The New Encyclopædia Britannica

Volume 14

MACROPÆDIA

Knowledge in Depth

FOUNDED 1768
15 TH EDITION

Encyclopædia Britannica, Inc.
Robert P. Gwinn, Chairman, Board of Directors
Peter B. Norton, President
Philip W. Goetz, Editor in Chief

Chicago
Auckland/Geneva/London/Madrid/Manila/Paris
Rome/Seoul/Sydney/Tokyo/Toronto

THE UNIVERSITY OF CHICAGO

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

The *Encyclopædia Britannica* is published with the editorial advice of the faculties of the University of Chicago.

Additional advice is given by committees of members drawn from the faculties of the Australian National University, the universities of British Columbia (Can.), Cambridge (Eng.), Copenhagen (Den.), Edinburgh (Scot.), Florence (Italy), Leiden (Neth.), London (Eng.), Marburg (Ger.), Montreal (Can.), Oxford (Eng.), the Ruhr (Ger.), Sussex (Eng.), Toronto (Can.), Victoria (Can.), and Waterloo (Can.); the Complutensian University of Madrid (Spain); the Max Planck Institute for Biophysical Chemistry (Ger.); the New University of Lisbon (Port.); the School of Higher Studies in Social Sciences (Fr.); Simon Fraser University (Can.); and York University (Can.).

First Edition	1768–1771
Second Edition	1777–1784
Third Edition	1788–1797
Supplement	1801
Fourth Edition	1801–1809
Fifth Edition	1815
Sixth Edition	1820–1823
Supplement	1815–1824
Seventh Edition	1830–1842
Eighth Edition	1852–1860
Ninth Edition	1875–1889
Tenth Edition	1902–1903

Printed in U.S.A.

Library of Congress Catalog Card Number: 89-81675
International Standard Book Number: 0-85229-529-4

CONTENTS

The Arctic

The Arctic is the area of the far north that is characterized by distinctively polar conditions of climate, plant and animal life, and other physical features. The term is derived from the Greek *arktos* ("bear"), referring to the northern constellation of the Bear. It has sometimes been used to designate the area within the Arctic Circle—a mathematical line that is drawn at latitude 66°30′ N, marking the southern limit of the zone in which there is at least one annual period of 24 hours during which the Sun does not set and one during which it does not rise. This line, however, is without value as a geographical boundary, since it is not keyed to the nature of the terrain.

While no dividing line is completely definitive, a generally useful guide is the irregular line marking the northernmost limit of the stands of trees. The regions north of the tree line include Greenland, Spitsbergen (Svalbard), and other polar islands; the northern parts of the mainlands of Siberia, Alaska, and Canada; the coasts of Labrador; the north of Iceland; also, a strip of the Arctic coast of Europe. The last-named area, however, is classified as subarctic because of other factors.

Conditions typical of Arctic lands are extreme fluctuations between summer and winter temperatures; permanent snow and ice in the high country, and grasses, sedges, and low shrubs in the lowlands; and permanently frozen ground (permafrost), the surface layer of which is subject to summer thawing. Three-fifths of the Arctic terrain is outside the zones of permanent ice. The brevity of the Arctic summer is partly compensated by the long daily duration of summer sunshine.

International interest in the Arctic and subarctic regions has steadily increased during the 20th century, particularly since World War II. Three major factors are involved: the advantages of the North Pole route as a shortcut between important centres of population; the growing realization of economic potentialities such as mineral (especially petroleum) and forest resources and grazing areas; and the importance of the regions in the study of global meteorology.

The article is divided into the following sections:

Physical and human geography

THE LAND

Relief and geology. The Arctic lands have developed geologically around four nuclei of ancient crystalline rocks. The largest of these, the Canadian Shield, underlies all the Canadian Arctic except for part of the Queen Elizabeth Islands. It extends eastward beneath Baffin Bay to include most of Greenland. The Baltic Shield, centred on Finland, includes all northern Scandinavia (except the Norwegian coast) and the northwestern Soviet Union. The two other blocks are smaller. The Angara (or Anabar) Shield is exposed between the Khatanga and Lena rivers in north central Siberia and the Aldan block is exposed in eastern Siberia.

Between the shields, particularly around their margins, there have been long periods of sedimentation, and consequently the shields are partly buried. In some areas thick sediments were subsequently folded, thus producing mountains, many of which have since been destroyed by erosion. Two main mountain-building periods have been recognized in the Arctic. In Paleozoic times there developed a complex mountain system that includes both Caledonian and Hercynian elements. It extends from the Queen Elizabeth Islands through Peary Land and down the east coast of Greenland. Mountain building occurred during the same period in Svalbard, Novaya Zemlya, the northern Urals, the Taymyr Peninsula, and Severnaya Zemlya. There is considerable speculation as to how these mountains are linked beneath the sea. The second period of folding was in Mesozoic and Cenozoic times. These mountains survive in northeastern Siberia and Alaska. Horizontal or lightly warped sedimentary rocks cover part of the shield in northern Canada, where they are preserved in basins and troughs. Sedimentary rocks are even more extensive in northern Russia and in western and central Siberia; they range in age from Early Paleozoic to Quaternary.

In the Tertiary Period (65,000,000 to 2,500,000 years ago) two Arctic areas experienced igneous activity. One

was associated with mountain building around the North Pacific, and active volcanoes are still found in Kamchatka, the Aleutian Islands, and Alaska. The other area of igneous activity extended across the North Atlantic and included the whole of Iceland, Jan Mayen Island, and east Greenland south of Scoresby Sound; it was probably connected to west Greenland and Disko Bay and to east Baffin Island. Volcanism continues in Iceland and on Jan Mayen, and hot springs are found in Greenland.

Little is known about the climate of the northern lands during Late Tertiary time; it is possible that the tree line was at least 1,000 miles (1,600 kilometres) farther north than at present, although even at this time Greenland may have been covered by ice.

Roughly 1,000,000 years ago the climate of polar regions began to deteriorate, and the Pleistocene ice sheets began to form. They are known to have expanded several times, but detailed information for the Arctic is available only for the last glaciation. In North America the main ice sheet of the final glaciation developed on Baffin Island and swept south and west across Canada, amalgamating with smaller glaciers to form the Laurentide ice sheet, covering much of North America between the Atlantic Ocean and the Rocky Mountains and between the Arctic Ocean and the Ohio and Missouri river valleys. A smaller ice cap formed in the western cordillera. The northern margin of the ice lay along the Brooks Range (excluding the Yukon Basin) and across the southern islands of the Canadian Archipelago. To the north the Queen Elizabeth Islands supported small, probably thin, ice caps.

Formation of Pleistocene ice sheets

The Atlantic Arctic islands were covered with ice except, as in Iceland, where isolated mountain peaks projected through the ice. In Europe the Scandinavian Ice Sheet covered most of northern Europe between Severnaya Zemlya in the Soviet Union and the British Isles. Northeastern Siberia escaped heavy glaciation, although, as in northern Canada, the ice sheet was more extensive in an earlier glaciation.

When the ice melted special landforms were left which, although not restricted to the Arctic, are often best developed there and in the absence of vegetation are clearly visible. In the areas of crystalline rocks, including large parts of the northern Canadian Shield and Finland, the ice left disarranged drainage and innumerable lakes. In the lowlands on weaker rocks, deep glacial deposits produced a smoother landscape, often broken by low ridges and hills of glacial material (drumlins). In the uplands the characteristic glacial landforms are the U-shaped valleys. Near the coasts these have been submerged partly to produce fjords, which are well developed in southern Alaska, along the east coast of Canada, around Greenland, in east and west Iceland, along the coast of Norway, and on many of the Arctic islands.

When the Pleistocene ice sheets melted the land remained depressed, and the sea flooded the Arctic lowlands. Subsequent emergence of the land has elevated beaches and marine deposits to considerable heights. The highest of these features are now found 500 to 900 feet (150 to 270 metres) above sea level in many parts of the western and central Canadian Arctic and somewhat lower in the east. Comparable elevations are found in the Eurasian Arctic. Comparable emergence is indicated on Svalbard, Greenland, the northern Urals, and on the Franz Josef Archipelago, where it may be more than 1,500 feet. In many of the emerged lowlands, such as those south and west of Hudson Bay, the raised beaches are the most conspicuous features in the landscape, forming hundreds of low, dry, gravel ridges in the otherwise ill-drained plains. Emergence is still continuing, and in northern Canada and northern Sweden uplift of two to three feet a century is found. Occasionally submergence is indicated, particularly at the mouths of such large rivers as the Mackenzie.

Although the detail of the terrain in many parts of the Arctic is directly attributable to the Pleistocene glaciations, the major physiographic divisions reveal close correlation with geological structure. The two largest shield areas, the Canadian and the Baltic, have developed similar landscapes. The landscapes west of Hudson Bay, on southern Baffin Island, and in Karelia are low and rocky with countless lakes and disjointed drainage. Uplands, generally 1,000 to 2,000 feet above sea level and partially covered with glacial deposits, are more widely distributed. They form the interior of Quebec–Labrador and parts of the Northwest Territories in Canada, and the Lapland Plateau in northern Scandinavia. The eastern rim of the Canadian Shield in Canada from Labrador to Ellesmere Island has been deeply dissected by ice to produce fjords and leave mountain peaks more than 6,000 feet high. The shield in Greenland is also a dissected upland, reaching similar elevations for long distances around the east and west coast. Where sedimentary rocks mantle the shields, as in north central Siberia and north of Hudson Bay, plains, plateaus, and hills result, the plateaus often being deeply dissected by narrow valleys. In Canada there are mainly plains and plateaus, while around the Angara Shield, hills and even mountain ranges are dominant.

Far beyond the margins of the shields, extensive plains have developed on soft sedimentary rocks. In North America these form the Mackenzie lowlands, Banks and Prince Patrick Islands, and the Arctic plains section of northern Alaska; in northern Europe they form the Severnaya Dvina and Pechora Plains. In Siberia the Ob Delta, its northeastern extension to the Laptev Sea, the north Siberian plains, the west Siberian lowlands, and farther east the Lena-Kolyma plains (including the New Siberian Islands) have developed on sedimentary rocks. Although there are differences in degree, these terrains are essentially flat, occasionally broken by low rock scarps, and covered with numerous shallow lakes. The plains are crossed by large rivers that have laid down deep alluvial deposits.

The strongly folded rocks associated with the two periods of mountain building in the Arctic form separate physiographic regions. The original mountains of the older, Paleozoic folding were long ago destroyed by erosion, but the rocks have been recently elevated and renewed erosion, often by ice, has produced a landscape of plateaus, hills, and mountains very similar to the higher parts of the shields. In Ellesmere Island the mountains are nearly 10,000 feet high. In Peary Land and Vestspitsbergen maximum heights are about 6,000 feet, while in east Svalbard and on Novaya Zemlya and Severnaya Zemlya the plateaus rarely exceed 2,000 feet.

The younger groups of fold mountains of northeast Siberia and Alaska are generally higher. Peaks of 10,000 feet are found in the Cherski Range, 15,000 feet in Kamchatka, and even higher in southern Alaska. Characteristic of this physiographic division are the wide intermontane basins often drained by large rivers such as the Yukon and Kolyma.

Throughout the Arctic, excluding a few maritime areas, the winter cold is so intense that the ground remains permanently frozen except for a shallow upper zone, called the active layer, which thaws during the summer. This permanently frozen ground (permafrost) covers nearly one-quarter of the Earth's surface. In northern Alaska and Canada scattered observations suggest that permafrost is 800 to 1,500 feet deep; it is generally deeper in northern Siberia. The deepest known permafrost is in northern Siberia, where it exceeds 2,000 feet. The depth of the permafrost depends on the site, climate, vegetation, and recent history of the area, particularly whether it was covered by sea or glacier ice. The very deep permafrost was probably formed in unglaciated areas during the Ice Ages. To the south in the subarctic, the permafrost thins and eventually becomes discontinuous, although it may locally still be 200 to 400 feet thick; along its southern boundary permafrost survives under peat and bogs. In areas of continuous permafrost the active layer may be many feet thick in sandy well-drained soils with little vegetation; it is usually less than six inches thick beneath peat.

Permafrost

Permafrost occurs in both bedrock and surface deposits. It has little effect on rock but in fine-grained, unconsolidated sediments, particularly silts, lenses of ice, called ground ice, grow by migration of water, and in extreme cases half the volume of arctic silts may be ice. Ground ice is often exposed in river banks and sea cliffs, where it may be 20 to 30 feet thick. In northern Siberia fossil ice has been reported up to 200 feet thick, although in these

cases it may be glacier or lake ice that has subsequently been buried under river deposits. If ground ice melts, due to a change in climate, pits are formed in the ground and quickly fill with water to form lakes and ponds. In their frozen state the silts have considerable strength, but if they thaw they change in volume, lose their strength, and may turn to mud. Variations in volume and bearing capacity of the ground due to changes in the permafrost constitute one of the major problems in Arctic construction.

Drainage and soils. Continuous permafrost totally inhibits underground drainage. Consequently shallow lakes are numerous over large areas of the Arctic, and everywhere in early summer there is a wet period before the saturated upper layers of the ground dry out. During the summer waterlogged active layers on slopes may flow downhill over the frozen ground, a phenomenon known as solifluction. It is ubiquitous in the Arctic but is particularly intense where the soils are fine-grained, as in the coastal plain of north Alaska, or where the precipitation is heavy, as on Bear Island (Bjørnøya) in the Norwegian Sea. The effect of solifluction is to grade slopes so that long, smooth profiles are common; slopes are normally covered with vegetation, but if the soil movement is too fast plants may not be able to survive. Under these conditions the surface material is often graded, with narrow strips of pebbles and boulders separated by broader strips of finer particles.

Formation of soil patterns

Many soils in northern areas show distinctive patterns produced by complex processes of freezing and thawing, which cause frost heaving and sorting of debris; although permafrost is not essential to these formations, it is usually present. There are many different types of pattern ground. In some, coarser material, pebbles, and boulders form polygonal nets with the finer materials concentrated in the centre. When sorting is widely spaced stone circles develop. Another kind of pattern, formed in sands and muds, is outlined by frost-crack fissures or strips of vegetation. Individual polygons vary from about one foot to more than 300 feet in diameter. Mounds due to frost heaving in the soil are also widespread. They grow rapidly, disrupting leveled fields in a few years.

Arctic soils are closely related to vegetation. Unlike soils farther south they rarely develop strong zonal characteristics. By far the commonest are the tundra soils, which are circumpolar in distribution. They are badly drained and strongly acid and have a variable, undecomposed organic layer over mineral horizons. Some of the drier heath and grassland tundras overlie Arctic brown soils, which have a dark-brown upper horizon with gray and yellowish-brown lower horizons. The active layer in the permafrost is normally deep in them.

Many exposed rock surfaces in the Arctic have been quickly broken up by frost action so that much rock is buried under a cover of angular shattered boulders. These mantles are known as *felsenmeer* (sea of rock) and are found principally on Actic uplands. Their continuity and depth varies with climate, vegetation, and rock type, but they may be as much as 12 feet deep. *Felsenmeer* are especially well developed on basalts and are consequently numerous on the Icelandic plateaus. They also develop quickly on sedimentary rocks and are widespread in the Canadian Arctic, all the way down to sea level.

Glaciation. Although the Arctic is commonly thought to be largely ice-covered, less than two-fifths of its land surface in fact supports permanent ice. The remainder is ice-free because of either relatively warm temperatures or scant snowfall. Glaciers are formed when the annual accumulation of snow, rime, and other forms of solid precipitation exceeds that removed by summer melting. The excess snow is converted slowly into glacier ice, the rate depending on the temperature and annual accumulation. In the Arctic, where most glaciers have temperatures far below freezing point, the snow changes into ice slowly. In northwest Greenland a hole 1,400 feet deep was made into the ice sheet without reaching glacier ice. The hole showed more than 800 annual snow layers, from which it was possible to determine precipitation changes for the last eight centuries.

The altitude at which ice accumulates over a period

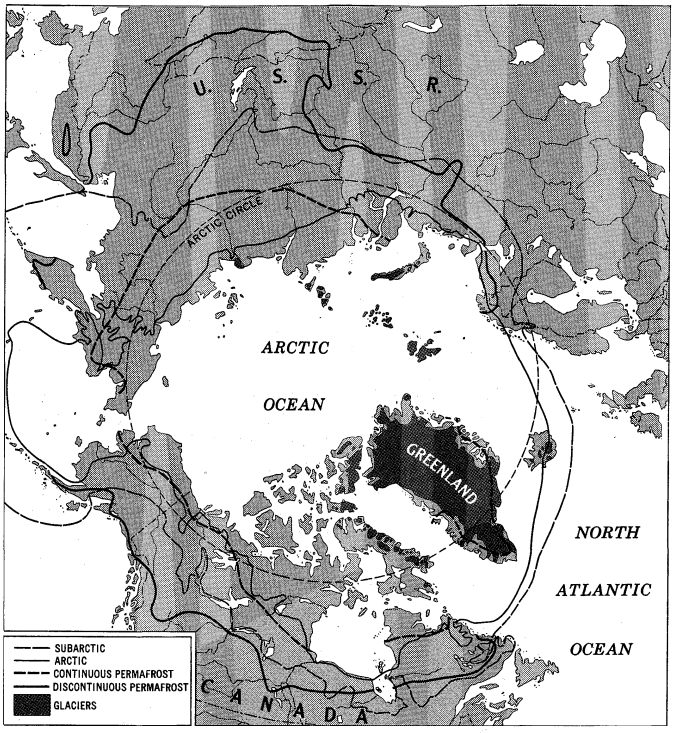


Figure 1: Division of subarctic and Arctic regions showing distribution of permafrost and glaciers.

of years is known as the glaciation limit and is roughly equivalent to the snow line. It is often very variable within short distances. On Baffin Island the glaciation limit is a little more than 2,000 feet above sea level in the extreme southeast, rising to more than 4,500 feet in the Penny Ice Cap 300 miles to the north and descending to about 2,000 feet in the north of the island. In Greenland the limit is at about 6,000 feet in the south and decreases irregularly to about 3,000 feet in the north. Nowhere in the Arctic regions is the glaciation limit at sea level; glaciers found close to the sea have all flowed down from higher levels. The domes of some arctic ice caps are below the glaciation limit but they continue to survive because of their low internal temperatures; the winter snowfall melts completely but refreezes in contact with the cold ice. This phenomenon, first observed on the Barnes Ice Cap of Baffin Island, is now known to be widespread in the high Arctic.

The glaciers of the north polar regions can be divided into two groups depending on the source of their snow; the larger group is around the Atlantic and the smaller around the Pacific Ocean. The largest ice mass, the Greenland inland ice, is part of the first group and is the largest glacier in the Northern Hemisphere. It extends about 1,570 miles from north to south, has a maximum width of some 600 miles, an average thickness of about 5,000 feet, and covers an area of more than 700,000 square miles (1,800,000 square kilometres), nearly 85 percent of Greenland. It is contained within an elongated saucer-shaped depression, the sides of which are mountains (called nunataks) that rise through the margins of the ice. In the centre the base of the ice is more than 1,000 feet below sea level. This discovery has led to the suggestion that Greenland is an archipelago rather than one large island. Although this might be so for a short time if the ice melted, the land would soon rise when the ice mass disappeared, forming a plateau at about 3,000 feet. The crest of the inland ice, which exceeds 10,000 feet above sea level midway between the east and west coasts in the northern part of the island, is displaced to the east in the south. In the centre the surface of the ice is undulating and is frequently covered with wind-drifted formations of snow called *sastrugi*. The ice cap slopes off to the sides, reaching the sea in a 240-mile front along Melville Bay in the west and along two smaller stretches in the northeast. Elsewhere huge outlet glaciers flow down to the sea. All combine to produce the vast numbers of icebergs that bar the coast of Greenland and

are carried south in the Labrador Current to the Atlantic shipping lanes. In summer the margins of the inland ice are covered with soft, sticky, granular snow. Three main ice-free areas are found in Greenland: in the southwest, where the inland ice is separated by 100 miles from Davis Strait; north of Scoresby Sound in the east; and in Peary Land in the north.

Location of ice caps

In Arctic Canada the permanent ice is restricted, with few exceptions, to the northeast, which is a result of the greater relief and precipitation around Baffin Bay and Davis Strait. The most southerly ice is found in the Torngat Mountains of northern Labrador, where there are small cirque glaciers. Immediately north of Hudson Strait on the plateau south of Frobisher Bay, there are two small ice caps. Larger ice caps and highland ice (through which mountains project) are present farther north along the east of Baffin Island and on Bylot Island; only the Barnes Ice Cap lies west of the coastal group. North of Lancaster Sound the ice is more extensive and large parts of Devon, Ellesmere, and Axel Heiberg islands are glacierized. In many ways these ice caps are small versions of the Greenland inland ice, with a central dome-shaped section and outlet glaciers pouring down through the mountains toward the sea. The ice cap on Meighen Island, the most westerly of the group, is an exception, as it is circular in shape and lies on low ground. Except for three small, unexplored glaciers on Melville Island, there are no glaciers in the Canadian western Arctic. Only a few of the Canadian glaciers reach the sea and form icebergs. In the Arctic Ocean off northern Ellesmere Island there is an area of floating shelf ice that may at one time have been joined by glaciers, but the glaciers no longer reach the sea. This shelf ice has been the principal source of the ice islands of the Arctic Ocean.

Other glaciers are found north and east of the Atlantic Ocean and its continuation in the Norwegian and Barents seas. Iceland has five major ice caps, the largest of which, Vatnajökull, covers more than 3,200 square miles. All have small outlet glaciers, although none reaches the sea. The ice caps owe their survival to extremely heavy precipitation. The western part of Vatnajökull buries a volcano, Grímsvötn, which erupts every six to 10 years; the heat of the eruption forms a vast subglacial lake which bursts in great floods over the margins of the glaciers.

North of Iceland, Jan Mayen Island supports a glacier on the volcano Mt. Beerenberg. The glaciers of Svalbard cover about 90 percent of the land. On the largest island, Vestspitsbergen, the plateaus are covered with highland ice from which outlet glaciers reach the sea; there are also numerous independent valley and cirque glaciers. North East Land, the second largest island, supports two ice caps on its plateaus. On the east side of the Atlantic Ocean precipitation is heavy over the Scandinavian highlands, but temperatures are also high, and the total area of ice is only about 2,000 square miles, a small part of which is in north Sweden and the remainder in Norway. To the northeast beyond the Barents Sea, precipitation is far less, but the summer is shorter and permanent ice is widespread.

Farthest north in this group are the islands of the Franz Josef Archipelago. Although at no point are they higher than 2,500 feet, probably more than 90 percent of their area is covered with ice; some of the smaller islands are completely buried by glaciers. The southern island of Novaya Zemlya supports a few small glaciers; on the northern island they become more numerous and the northern four-fifths of the island is ice-covered, with large outlet glaciers reaching the sea.

Weather disturbances penetrate into the Kara Sea beyond Novaya Zemlya and produce sufficient snow for ice to form on Severnaya Zemlya (North Land). There are four major and many minor islands in the group. Although they are low-lying, consisting primarily of plateaus less than 2,000 feet high, all the larger islands have ice caps that cover rather less than half the total area. Outlet glaciers reach the sea and are an occasional source of icebergs. Elsewhere the Russian northern areas are remarkably free of ice. Small cirque glaciers are found in the Ural Mountains and the mountains of northeastern Siberia. Not until the Pacific is approached do the mountains again have alpine glaciers, another indication of the importance of moisture for the development of glaciers.

The glaciers around the North Pacific are concentrated in Alaska. The glaciers of southern Alaska are alpine rather than Arctic and include some of the finest mountain glaciers in the world. All types of ice are present, from small valley glaciers to highland ice almost burying mountain ranges, with piedmont glaciers spreading out in the lowlands. The largest ice centres are around the Fairweather Range, the St. Elias Mountains, and the Chugach Mountains. Glaciers in these areas include the Hubbard, 90 miles long, intermontane glaciers such as the Seward, and piedmont glaciers such as the Malaspina. Smaller glaciers also occur inland on the Alaska Range and in the Brooks Range of northern Alaska; there is more ice farther east in the Romanzof Mountains, where one glacier, the Okailak, is 10 miles long. There are a few small glaciers in the Aleutian Range, and ice is found infrequently on the Aleutian Islands.

The overwhelming majority of Arctic glaciers, except for the Greenland inland ice, are in retreat, and many areas are more ice-free than ever before in the historical period. In Iceland, where glacier movements are well recorded, the ice appears to have been restricted from the 10th until about the 14th century. The ice then advanced, reaching a maximum about 1750. A second advance followed a minor retreat, culminating about 1850, and a major retreat set in about 1890. The recession was slow at first but by the 1930s it was generally rapid and has continued since.

Climate. The climates of northern lands are highly varied. It has been customary to divide them into two large groups corresponding to the climate of ice caps, in which no mean monthly temperature exceeds 32° F (0° C), and the tundra climates, with at least one month above 32° F but no month above 50° F (10° C). A more satisfactory division is to classify them as polar maritime climates, located principally around the Atlantic and Pacific oceans, in which winter temperatures are rarely extremely low and snowfall is high; and the polar continental climates, as in north Alaska, Canada, and Siberia, where winters are very cold and snowfall is generally light. Included in the polar continental climate type are the Arctic Islands (also called the Canadian Arctic Archipelago), which are influenced only slightly by the sea in winter because of thick ice. In addition to these two climates, there are smaller transitional zones, limited areas of "ice" climates and, adjacent to the Arctic, the subarctic climates.

Record low temperatures

Although the polar continental climates have low winter temperatures, they are not as cold as the subarctic. The lowest temperatures ever recorded in the Northern Hemisphere were measured in the subarctic of northeast Siberia, the coldest area being around Oymyakon. In North America a temperature of −81° F (−63° C) was observed at Snag, Yukon Territory. Temperatures below −80° F (−62° C) have been recorded on the Greenland Ice Cap. Colder temperatures have not been found in the tundra regions because generally windier conditions prevail.

Winter sets in toward the end of August in the far north and about a month later near the tree line. The temperature continues to drop rapidly until about December. January, February, and early March have uniform conditions with mean temperatures about −35° F (−37° C) in the central Siberian Arctic and −20° to −30° F (−29° to −34° C) in North America. The lowest extreme temperatures in the winter are between −50° and −65° F (−46 and −54° C). A better indication of low temperatures as they affect man is given by the windchill, a measurement of the cooling power of the atmosphere on human skin. It reaches a maximum north of Hudson Bay, where strong and persistent northwest winds, typical of the Canadian eastern Arctic, are combined with low air temperatures. This area is stormy in winter with moderately high snowfall (50 to 100 inches [1,300 to 2,500 millimetres]), rapidly changing temperatures, and even occasional rain. Elsewhere the winter continental climate is quiet, with long periods of clear sky and low snowfall. Visibility may be poor locally if there are open channels of water in the sea ice and is universally reduced when the wind blows drifting snow. The lowest snowfall is in the northern Canadian

islands and in north Greenland, where the total annual precipitation is frequently less than four inches of water.

Winter in the maritime Arctic (the Aleutians, coastal southwest Greenland, Iceland, and the European Arctic) is a period of storminess, high winds, heavy precipitation in the form of either snow or rain (the latter at sea level), and moderate temperatures. The coldest month is rarely below 20° F (−7° C) and extremely low temperatures are unknown.

Summer temperatures are more uniform across the whole of the Arctic. On the southern margin the monthly mean temperature reaches 50° F (10° C), and in continental positions short spells of hot weather with temperatures in the 80s, continuous sunshine, and calm weather are not uncommon; such weather often ends with thunderstorms. In the maritime climates, along the coasts, and on the northern islands when there is open water in the sea ice, the summer is relatively cool. In the south the temperatures are about 45° F (7° C), decreasing north to 40° F (4° C) or less; a maximum of 60° F (16° C) is hardly ever reached. Fog and low clouds are widespread, and at this time of the year these areas are the cloudiest in the world. In areas that experience continental winters, precipitation is heaviest during the summer months; light rain and snow showers are frequent, but the average fall is rarely high. The summer is everywhere a time of sudden changes. Calm, clear weather with sunshine and temperatures of about 50° F will be followed by sudden winds, often causing a temperature drop of 20° to 30° F and accompanied by cloud and fog. These changes are often local, and fine weather may continue in the next valley or at the head of a fjord.

The frost-free and growing periods are relatively short throughout the Arctic. For the most part there is no true frost-free period; frost and some snow have been recorded in every month of the year. At a few places near the tree lines, notably in the Canadian western Arctic, the frost-free period may be the same as the less favourable parts of the prairies.

The polar basin has a winter similar to that of northern Alaska and Siberia with long, clear, cold periods and occasional storms, accompanied by a brief rise in temperature. The summers are cloudy and foggy, winter snow melts off the pack ice, and the mean temperature rises to about 35° F (2° C). The climate of Svalbard and Novaya Zemlya is transitional between maritime and continental conditions with a cool rather than cold winter (mean temperature 0° F [−18° C]).

The only extensive ice climate in the Northern Hemisphere is in Greenland. In the south the ice cap has maritime characteristics with heavy precipitation, mainly snow from passing cyclone disturbances. In the north conditions are less severe and the snowfall consequently less. Although the air temperature may sometimes rise to 32° F, the mean temperature is much lower.

Changes in polar climates

The polar climates have changed noticeably during the 20th century, the climatic amelioration that produced the retreat of the glaciers having been widely felt. Sea ice around Iceland, Svalbard, and southwest Greenland has decreased in extent. The pack ice in the Arctic Ocean is thinner. Birds, animals, and especially fish have appeared farther north than before; in Greenland this led to a complete change in the economy, as its traditional dependence on seals has yielded to dependence on fishing, particularly cod, which are caught north of the 70th parallel. The cod, unlike the seal, does not provide fuel and clothing in addition to food; consequently the economy, once extraordinarily self-sufficient, is now based on cash and international trade. With the warming of the climate, sheep farming also has been introduced successfully in the southern part of Greenland.

The main changes in the climate result from increases in air and sea temperatures. The increases in air temperatures have been greatest during the North Atlantic Ocean winters around Svalbard but are found at all seasons and in all areas. The warming began at the beginning of the 20th century, reached a maximum in the 1930s, and declined briefly in many localities in the early 1940s, only to increase again temporarily. The primary cause is not known, although it results directly from increased penetration of southerly winds into the polar regions.

Plant life. Two main vegetation zones are found in the polar lands. In the south is the subarctic, formed by the northern subzones of the circumpolar boreal forest. To the north is the Arctic proper, where the vegetation is generally referred to as tundra, from the Finnish word for an open rolling plain; in North America the descriptive term Barren Grounds is frequently applied. The two zones are separated by the tree line. This is by definition the absolute northern limit of treelike species, although even beyond it the same species may be found in low shrubs and dwarfed forms. The tree line is composed of different species. In Alaska and northwestern Canada white spruce is dominant, while in Labrador–Quebec it is black spruce and occasionally larch. By contrast, in northern Europe and Siberia the tree line is formed by larch, pine, and fir. The tree line is related to summer warmth, which may be correlated closely with tree growth. Alexander Supan found good coincidence between the tree line and the 50° F July isotherm, a figure later modified by Otto Nordenskjöld to allow for spring temperatures.

In North America the tree line extends from the shores of Bering Strait along the Brooks Range of Alaska to the Mackenzie Delta and then curves southeastward across the Northwest Territories to Churchill and James Bay. East of Hudson Bay it crosses northern Quebec to Ungava Bay and then continues into Labrador. In western Scandinavia, the tree line is within a few miles of the sea; it curves east and crosses northern Siberia 50–150 miles south of the Arctic Ocean.

Arctic plants have to contend with a harsh environment including low temperatures, continuous daylight in summer, poor soil and permanently frozen ground, and in many areas strong, dry winds and blowing snow. The species that survive are few and are frequently dwarfed. Many plants grow in compact cushions for maximum protection from the climate. The growing season is so short that annuals are rare and perennials reproduce asexually by shoots or runners. Even so, Arctic plants have a rapid seasonal life cycle. Spring growth often begins when snow is on the ground and there are still heavy frosts; the flower and seed stages follow in a period as short as six weeks. The sudden blooming of flowers is spectacular, particularly along the southern edges of the tundra, and for a short time in July the Barren Grounds are covered with a mass of flowers. The species vary but typical are those in the western American Arctic, which include the blue-spiked lupine, wild crocus, mountain avens, arctic poppy, and saxifrage. By late August the cycle is complete, and the plants are awaiting winter.

At first sight many parts of the Arctic are rocky wastes without soil or vegetation. Closer inspection shows that some plant life is always present, and even on permanent ice there are often algae. The bare rock surfaces support thin brown, black, or gray crustaceous lichens that swell and become soft when wet; some of the larger black lichens are edible and are generally known as "rock tripe." In the past these lichens have been used for food by starving explorers. Higher plants grow in rock crevices and succeed in forming tussocks on patches of soil. Close to the southern edge of the Arctic, dwarf shrubs are found in protected sites on these rock deserts.

Tundra vegetation

Tundra areas have a continuous cover of vegetation, and many different tundra associations (plant communities) may be recognized. In the drier and better drained parts, heath tundra, made up of a carpet of lichens and mosses with isolated flowering plants, is dominant. In some areas, notably west of Hudson Bay, a similar environment results in tundra grassland. When there is more moisture, sedges and grasses become important and form tussock or hillock tundra; willow and dwarf birch may be found between the individual mounds. This type of tundra reaches its greatest development on the north Alaskan coastal plain.

In the warmer parts of the Arctic, woody dwarf shrubs, willow, birch, juniper, and, locally, alder are profuse. In the southern Arctic several of these shrubs modify the heath tundra, and low scrub woods may be quite extensive. On sheltered, south-facing slopes, tall thickets of willow, birch,

and alder develop, and under optimum conditions these bushlike "trees" may be more than 10 feet high. This type of vegetation is common in all circumpolar lands close to the tree line and is conspicuous in the inner fjords of southwest Greenland and in northern Iceland. The bushes may be used in the western Canadian Arctic for fuel or for mats, and in former times were used by Eskimos for arrow shafts. The wood is unsuitable for bows, spears, or boat building; for these purposes the Eskimos either had to travel to the tree line or search for driftwood, which was widely distributed along the Arctic coasts.

The tundra vegetation is the source of food for the northern grazing mammals but contains few foods of direct value to man. Berries are found throughout the southern Arctic. Most widely used by the native population has been the black crowberry (*Empetrum nigrum*), eaten either raw or mixed with animal oil. Europeans have found the cloudberry (*Rubus chamaemorus*), bilberry (*Vaccinium uliginosum*), and mountain cranberry (*V. vitisidaea minus*) more palatable. Mushrooms are widely distributed and can be used for a welcome change of diet.

South of the tree line is the subarctic forest-tundra. Its bare windswept ridges are covered with tundra associations, while in the sheltered valleys there are woodlands, which may become continuous near large rivers and, if the rivers flow north, may project many miles into the Barren Grounds. These areas, known as *galeria* forests, are found along the Coppermine River of Canada and frequently along the north Siberian rivers. The woods contain the same coniferous species as forms the tree line, together with several broad-leaved species, notably birch.

(J.B.Bd.)

Animal life. Animal life in the Arctic, compared with that of warmer parts, is poor in species but often rich in individual numbers. This is generally considered to be the result of at least two factors—the comparative novelty of polar glacial climates, allowing only a limited time for adaptation since their onset; and the much lesser variety of habitats available for colonization in the north as compared with the lower latitudes.

Fauna of the true Arctic Zone

The fauna considered in this section is from the true Arctic Zone only. On the land, this is the zone north of the tree line; in the sea, the area in which the upper water is of North Polar Sea origin, without admixture of Atlantic or Pacific water. This excludes most of the west Greenland waters and the waters of west and southern Iceland, the Faeroe Islands, and Norway; it also excludes the Labrador Sea and the waters of the Labrador coast south of Hudson Strait.

Animals of the land and fresh water. The typical and best known Arctic land mammals and birds are those highly successful forms, most of them circumpolar in distribution, that survived the Pleistocene glaciations probably both south and north of the ice sheets: south along the ice perimeter and north in ice-free refuges such as northern Alaska, the Bering Strait (then dry land) and northeastern Siberia, certain of the Arctic Islands, and probably northernmost Greenland. These include the polar bear (as much a marine as a terrestrial animal), caribou, Arctic wolf, Arctic fox, Arctic weasel, Arctic hare, the brown and collared lemmings, ptarmigan, gyrfalcon, and snowy owl. This fauna, together with the vegetation that feeds the lemming, ptarmigan, and caribou, forms a tight ecological system that is virtually self-sufficient. During the winter, and during periods of low lemming population, which occur every three to five years, the carnivores make some use of seashore life and (through the agency of the polar bear) of seal and fish. In extreme starvation conditions, there is a tendency for the snowy owls and gyrfalcons to go south in winter and for the foxes and wolves to become scavengers.

The caribou is a migrant, but only between the Arctic tundra and the conifer (subarctic) zone to the south, and there are far northern groups of caribou whose migrations are more restricted. The musk-ox is a special case. Now restricted to the North American Arctic (including north Greenland), it was formerly more widespread and is probably a "refugee" species, chased into the far north and on the defensive in the evolutionary sense.

Hibernation is not a phenomenon of animal behaviour in the Arctic because frost-free refuges are not available; all the nonmigrant, warm-blooded animals therefore must remain active all winter. Any incipient hibernation, shown for instance by the Arctic ground squirrel, proves abortive, as the animals will shiver themselves awake after only a few days.

Most of the birds of the Arctic Zone are migrants, coming from wintering grounds as far away as the southern United States, Central America, Brazil, or even the subantarctic zone. By migration the birds obtain the advantage of the long northern summer days and of the high productive capacity of plant foods in the short but intense growing season. There is increasing evidence that food is not a limiting factor on summer bird populations in the Arctic, except in the case of strictly predaceous species during years of scarcity of prey. Typical land and freshwater birds of the Arctic Zone are the redpoll, Lapland longspur, snowbird, wheatear, pipit, certain plovers and sandpipers, loons, rock ptarmigans, ducks, and geese.

There are no reptiles in the Arctic Zone, owing to the absence of frost-free winter refuges, but one amphibian, the wood frog, does penetrate just north of the tree line in Arctic Canada. It breeds in July and early August in ponds and small lakes and spends the rest of the year buried in the mud at the bottom. The mud does not freeze, and the frogs are able to breathe through the skin, which the reptiles cannot do.

Freshwater fishes in the Arctic are represented by a few species only: whitefish, lake trout and speckled trout, Arctic grayling, two species of stickleback, the Alaskan blackfish, and the Arctic char. In some regions the burbot, northern pike, and Atlantic salmon penetrate north of the tree line.

The invertebrate fauna of the Arctic land and fresh water consists largely of insects, including the chief scourges of the north, mosquitoes and black flies. Among the most northern navigators are certain species of spiders that winter even in northern Ellesmere Island. Crustacea are represented by the branchiopods, which form an important part of Arctic pond life, and by the copepods. There is, in addition, a very considerable number of smaller species belonging to many phyla. (For animals of the sea, see below *The Arctic Ocean.* For further details on physical and human geography, see below *The Arctic Islands; Greenland; Novaya Zemlya;* and *Svalbard.*)

(M.J.Du.)

THE ARCTIC OCEAN

The Arctic Ocean is by far the smallest of the Earth's five oceans, having only a little over one-sixth the area of the next largest, the Indian Ocean. Its area of 4,732,000 square miles (12,257,000 square kilometres), however, is five times larger than that of the largest sea, the Mediterranean. The deepest sounding obtained in Arctic waters is 18,050 feet (5,502 metres), but the average depth is only 3,240 feet.

Distinguished by several unique features, including a cover of perennial ice and almost complete encirclement by the landmasses of North America, Eurasia, and Greenland, the north polar region has been a subject of speculation since the earliest concepts of a spherical Earth. From astronomical observations the Greeks theorized that north of the Arctic Circle there must be a midnight sun at midsummer and continual darkness at midwinter. The enlightened view was that both the northern and southern polar regions were uninhabitable frozen wastes, whereas the more popular belief was that there was a halcyon land beyond the north wind where the Sun always shone and people called Hyperboreans led a peaceful life. Such speculations provided incentives for adventurous men to risk the hazards of severe climate and fear of the unknown to further geographic knowledge and national and personal prosperity.

The ocean floor. From the late 19th century, when the Norwegian explorer Fridtjof Nansen first discovered an ocean in the central Arctic, until the middle of the 20th century, it was believed that the Arctic Ocean was a single large basin. Explorations after 1950 revealed the true com-

A U.S. Coast Guard icebreaker and an oceangoing tugboat travelling through the Arctic Ocean on a mission during the brief Arctic summer.
By courtesy of the U.S. Coast Guard; photograph, Monkmeyer

plex nature of the ocean floor. Rather than being a single basin, the Arctic Ocean consists of two principal deep basins that are subdivided into four smaller basins by three transoceanic submarine ridges. The central of these ridges extends from the continental shelf off Ellesmere Island to the New Siberian Islands, a distance of 1,100 miles. This enormous submarine mountain range was discovered by Soviet scientists in 1948–49 and reported in 1954. It is named the Lomonosov Ridge after the scientist, poet, and grammarian Mikhail Vasilyevich Lomonosov.

The Lomonosov Ridge has an average relief of about 10,000 feet and divides the Arctic Ocean into two physiographically complex basins. These are referred to as the Eurasia Basin on the European side of the ridge and the Amerasia Basin on the American side. The Lomonosov Ridge varies in width from 40 to 120 miles, and its crest ranges in depth below sea level between 3,100 and 5,400 feet.

Eurasia Basin and Amerasia Basin

The Eurasia Basin is divided into two smaller basins by a trans-Arctic Ocean extension of the Mid-Atlantic Ridge. This Arctic segment of the global ridge system is called the Nansen Cordillera. Although it is a locus of active ocean floor spreading, its topography is less well developed than other oceanic ridges. The Fram Basin lies between the Nansen Cordillera and the Lomonosov Ridge at a depth of 14,070 feet. The geographic north pole is located over the floor of the Fram Basin near its juncture with the Lomonosov Ridge.

The smallest of the Arctic Ocean sub-basins, called the Nansen Basin, lies between the Nansen Cordillera and the Eurasian continental margin, and has a floor depth of 13,800 feet.

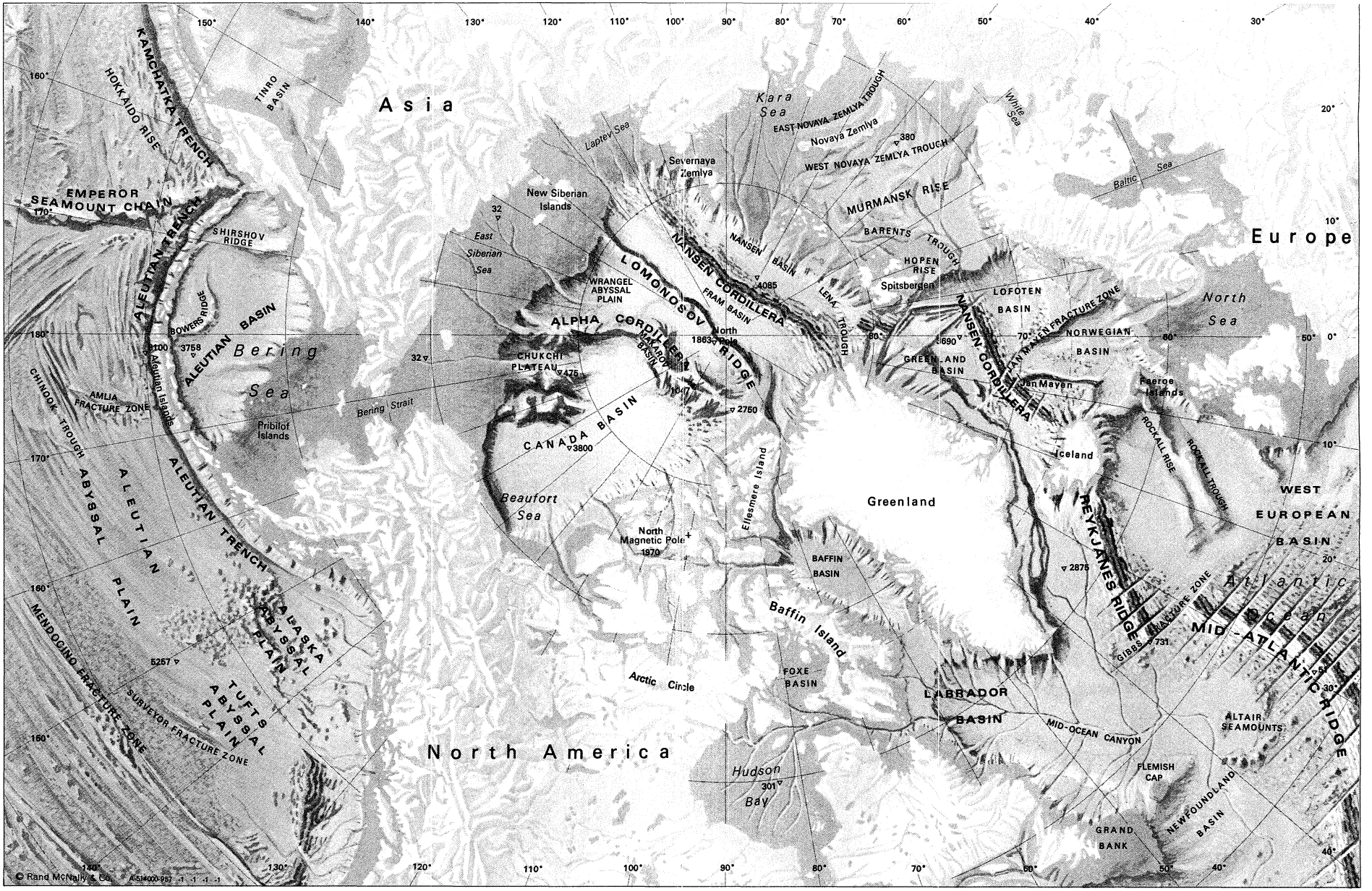
The Amerasia Basin is divided into two unequal basins by the Alpha Cordillera, a broad rugged submarine mountain chain that extends to within 4,600 feet of the ocean surface. The cordillera is believed to be a no longer tectonically active segment of the global rift system. The Makarov Basin lies between the Alpha Cordillera and the Lomonosov Ridge, and its floor is at a depth of 13,200 feet. The largest sub-basin of the Arctic Ocean is the Canada Basin, which extends approximately 700 miles from the Beaufort Sea Shelf to the Alpha Cordillera. The smooth basin floor slopes gently from east to west where it is interrupted by regions of sea knolls. The average depth of the Canada Basin is 12,500 feet.

The Arctic Ocean is unique in that nearly one-third of its total area is underlain by continental shelf, which is asymmetrically distributed around its circumference. North of Alaska and Greenland the shelf is 60 to 120 miles wide, which is the normal width of continental shelves. In contrast, the Siberian and Chukchi shelves off Eurasia range from 300 to 1,100 miles in width. The edge of the continental margin is dissected by numerous submarine valleys. The largest of these, the St. Anna Trough, is 110 miles wide and 300 miles long.

Oceanography. As an approximation, the Arctic Ocean may be regarded as an estuary of the Atlantic Ocean. The major circulation into and from the Arctic Basin is through its single, deep channel connection to the World Ocean: the passage that lies between the islands of Spitsbergen and Greenland. The connection of the Arctic Ocean with the Pacific Ocean is of minor importance. There are, however, factors in the Arctic Ocean that make its processes significantly different from those of the Atlantic. Most notable is the covering ice pack, which reduces the exchange of energy between ocean and atmosphere by about 100 times. In addition, sea ice greatly reduces the penetration of sunlight needed for the photosynthetic processes of marine life and impedes the mixing effect of the winds. A further significant distinguishing feature is the high ratio of freely connected shallow seas to deep basins. Whereas the continental shelf on the North American side of the Arctic Ocean is of a normal width (approximately 40 miles), the Eurasian sector is hundreds of kilometres broad, with peninsulas and islands dividing it into five main marginal seas: the Chukchi, East Siberian, Laptev, Kara, and Barents. These marginal seas occupy 36 percent of the area of the Arctic Ocean, yet they contain but 2 percent of its water volume. With the exception of the Mackenzie River of Canada and the Colville River of Alaska, all major rivers discharge into these marginal shallow seas. The combination of large marginal seas, with a high ratio of exposed surface to total volume, plus large summer inputs of fresh water, greatly influences surface-water conditions in the Arctic Ocean.

Main directions of flow

The balance of the Arctic Ocean's water budget (input and output) is still very poorly known. Although the largest volume of water transport is between Greenland and Spitsbergen, this mechanism is the least understood of all. The inflow of dense bottom water from the Atlantic occurs mainly on the eastern side of the passage and is estimated to be 78 percent of the total volume brought



ARCTIC OCEAN Depths in metres

Colours used are thought to be those of the various rocks and sediments on the sea floors. Differences in relief are shown by relief shading.

0 400 800 1200 km
0 400 800 1200 mi

into the Arctic Ocean. This is balanced by the surface outflow of colder but less saline waters in the form of the East Greenland Current. Very few data are available on the transport volume of this current, but, from indirect evidence, it is estimated to account for 83 percent of the total discharge from the Arctic Basin. Of this figure, 2 percent is in the form of icebergs. Although the volume of water transported as ice is a small fraction of the total, it is of considerable importance to oceanic heat balance and poses a major problem to North Atlantic shipping.

The flow of water out of the Arctic Ocean through the Canadian Arctic Archipelago is about 17 percent of the total discharge, and, compared to loss through the Greenland Sea, the transport of ice is negligible. There is essentially no northward flow of water through the shallow channels separating the islands of the Archipelago.

The volume of water entering the Arctic Ocean through the Bering Strait is better known than that from the Atlantic and is approximately 20 percent of the total. The amount of fresh water entering the Arctic Ocean is about 2 percent of the total input. Precipitation is believed to be about 10 times greater than loss by evaporation, although both figures can be only roughly estimated.

Through all of these various routes and mechanisms, the exchange rate of the Arctic Ocean is estimated to be approximately 210,000,000 cubic feet (5,900,000 cubic metres) per second.

All waters of the Arctic Ocean are cold. Variations in density are thus mainly determined by changes in salinity. On the basis of density, Arctic waters have a two-layer system: a thin and less dense surface layer is separated by a strong density gradient, referred to as a pycnocline, from the main body of water, which is of quite uniform density. This pycnocline restricts convective motion and the vertical transfer of heat and salt, and hence the surface layer acts as a cap over the larger masses of warmer water below.

Major water masses

Despite this overall similarity in gross oceanographic structure, the waters of the Arctic Ocean can be classified into three major masses and one lesser mass.

(1) The water extending from the surface to a depth of about 650 feet is the most variable and heterogeneous of all that in the Arctic. This is because of the latent heat of freezing and thawing; salt-brine addition from the process of ice freezing; freshwater addition by rivers, ice melting, and precipitation; and great variations in insolation (rate of delivery of solar energy) and energy flux as a result of sea ice cover. Water temperature may vary over a range of 7° F (4° C) and salinity from 28 to 34 grams of salt per kilogram of seawater (28 to 34‰).

(2) Warmer Atlantic water everywhere underlies Arctic surface water from a depth of about 650 to 3,000 feet. As it cools it becomes so dense that it slips below the surface layer on entering the Arctic Basin. The temperature of this water is about 34° to 37° F (1° to 3° C) as it enters the basin but is gradually cooled so that by the time it spreads to the Beaufort Sea, it has a maximum temperature of 32.9° to 33.1° F (0.5° to 0.6° C). The salinity of the Atlantic layer varies between 34.5 and 35‰.

(3) Bottom water extends beneath the Atlantic layer to the ocean floor. This is colder than the Atlantic water (below 32° F or 0° C) but has the same salinity.

(4) An inflow of Pacific water can be observed in the Amerasia Basin but not in the Eurasia Basin. This warmer and fresher water mixes with colder and more saline water in the Chukchi Sea, where its density enables it to flow as a wedge between the Arctic and Atlantic waters. The Pacific water, by the time it reaches the Canada Basin, has a temperature range of 31.1° to 30.8° F (−0.5° to −0.7° C) and salinities between 31.5 and 33‰.

Surface-water movement

Arctic waters are driven by the wind and by density differences. The net effect of tides is unknown but could have some modifying effect on gross circulation. The motion of surface waters is best known from observations of ice drift. The most striking feature of the surface circulation pattern is the large clockwise gyre (circular motion) that covers almost the entire Amerasia Basin. Fletcher's Ice Island (T–3) made two orbits in this gyre over a 20-year period, which is some indication of the current speed. The northern extremity of the gyre bifurcates and jets out of the Greenland–Spitsbergen passage as the East Greenland Current, attaining speeds of six to 16 inches per second. Circulation of the shallow Eurasian Shelf seas seems to be a complex series of anticlockwise gyres, complicated by islands and other topographic relief.

Circulation of the deeper Atlantic water is less well known. On entering the Eurasia Basin, the plunging Greenland Sea water appears to flow eastward along the edge of the continental margin until it fans out and enters the Amerasia Basin along a broad front over the crest of the Lomonosov Ridge. There seems to be a general anticlockwise circulation in the Eurasia Basin and a smaller clockwise gyre in the Beaufort Sea. Speeds are slow—probably less than two inches per second.

The circulation of the bottom water is unknown but can be inferred to be similar to the Atlantic layer. Measured values of dissolved oxygen show that the bottom water is well ventilated, dissolved oxygen everywhere exceeding 70 percent of saturation.

Sea ice. The Arctic Ocean's cover of sea ice suppresses wind stress and wind mixing, reflects a large proportion of incoming solar radiation, imposes an upper limit on the surface temperature, and impedes evaporation. Wind and water stresses keep the ice pack in almost continuous motion, causing the formation of cracks (leads), open ponds (polynya), and pressure ridges. Along pressure ridges the pack ice may be locally stacked high and project downward into the ocean several tens of metres.

Besides its deterrence to the exchange of energy between the ocean and the atmosphere, the formation of sea ice generates vast quantities of cold waters that help drive the circulation of the world ocean system.

Sea ice rarely forms in the open ocean below a latitude of 60° N but does occur in more southerly enclosed bays, rivers, and seas. Between about 60° N and 75° N the occurrence of sea ice is seasonal, and there is usually a period of the year when the water is ice-free. Above a latitude of 75° N there is a more or less permanent ice cover. Even there, however, as much as 10 percent of the area consists of open water due to the continual opening of leads and polynyas.

In the process of freezing, the salt in seawater is expelled as brine. The degree to which this rejection takes place increases as the rate of freezing decreases. Typically, newly formed sea ice has a salinity of 4 to 6‰. Even after freezing the process of purification continues but at a much slower rate. By the time ice is one year old it is sufficiently salt-free to be melted for drinking. This year-old, or older, salt-free sea ice is referred to as polar pack. It can be distinguished by its smoother, rounded surface and pale blue colour. Younger ice is more jagged and grayer in colour. Because the hardness and strength of ice increases as the salts are expelled, polar pack is a special threat to shipping.

Longevity of the ice pack

Important evidence concerning the long-term stability of the Arctic ice pack has been found by studying the sedimentary record of the ocean floor. This research has recently shown that there is strong evidence of an uninterrupted existence of sea ice on the Arctic Ocean during at least the past 3,000,000 years. Such evidence strongly refutes those theories of the cause of ice ages that postulate rapid variability of the Arctic ice cover.

Plant and animal life. *Plankton.* The chain of marine food supply is founded on floating microscopic plants, the phytoplankton. The annual productivity of these plants in the Arctic Basin is less than 10 percent of that of other ocean areas. The factor limiting this primary production is the ice pack itself, which reduces the penetration of solar energy needed for photosynthesis. Recent studies have shown, however, that Pacific water entering the Amerasia Basin is rich in detritus and nutrients. This is presumably true of Atlantic water coming through the Greenland Sea.

The next link in the food chain are zooplankton, small animals that feed on phytoplankton. Available information on the zooplankton under the permanent ice pack is even more fragmentary than for phytoplankton. But many of the Arctic zooplankton (and fishes) grow larger than their near relatives in warmer seas. This is believed

to be due to their delayed sexual maturity in cold water rather than to their speed of growth.

Fish. Free swimmers, or nekton, depend upon zooplankton as their food base. In truly Arctic waters fish play a minor role in human economy compared to marine mammals. The reverse is true in the bordering seas, in which there is an admixture of warmer Atlantic or Pacific waters and in which fisheries abound. The most important fish is the Arctic char (*Salvelinus alpinus*), a member of the salmon family. It spends most of its life in fresh water and comes into the sea for a few weeks in the summer. Only the polar cod (*Boreogadus*) has been taken from the central Arctic Basin, although bottom photographs have shown at least one other species of fish and shallow sonic scattering, apparently from fish schools, has been noted during summer months.

Mammals. The large sea mammals are the most conspicuous polar wild life. All are important to the hunting natives, and some of them are of commercial interest. These include the suborder Pinnipedia, including the eared or fur seals, the hair seals, and the walrus. Fur seals do not occur as close to the pole as the other groups, reaching only the southern edge of the Arctic. The Pribilof Islands in the Bering Sea are one of the main breeding areas. Hair seals include the harbour seal, the jar or ring seal, the bearded seal, the harp seal, and the hooded seal. The first three of these are widely distributed in the Arctic throughout the year and are of importance to the Eskimo. The latter two are migratory, and have been exploited commercially. The walrus is divided into two species, Atlantic and Pacific, and inhabits shallow waters, feeding on clams and other bottom fauna, which it can root up with its ivory tusks.

A wide variety of whales frequents Arctic waters, either for the entire year or during the summer months. They include sperm, killer, white, bottlenose, finback, humpback, blue, and narwhals.

Polar bears range throughout the Arctic, including the central ocean basin, throughout the year. They are frequently followed by the scavenging Arctic fox.

(N.A.O.)

The seabirds in the true Arctic Zone are represented by the auk family (murres, guillemots, auklets, and little auk), the sea duck (eider, scoter, old squaw), the gulls and terns (especially the glaucous and glaucous-winged gulls, many of the herring gull group of species, Sabine's gull, and the common and Arctic terns), the jaegers (parasitic, pomarine, and long-tailed), and the waders (sandpipers, etc.). One of the petrel group, the fulmar, breeds on certain Arctic cliffs. These are all migrant species that fly south in the fall. The Arctic tern makes a remarkable migration to subantarctic waters. (M.J.Du.)

THE PEOPLE

Ethnic groups. *Western Arctic.* There are no tribes of Eskimos, the word tribe being used in the sense of social groupings with internal political organization; nor may the term be used in the sense of populations conscious of their own tribal identities as opposed to other tribal groups. The Eskimos call themselves Inuit or Innuit ("the people") and are careful to distinguish themselves from non-Eskimos, and particularly from American Indians, who are their nearest neighbours; but this is a general identification and involves no elements of social structure or organization. The only Eskimo social units with validity from the Eskimo point of view are the immediate family, the household (often made up of two families), and the local residence group, which is identified only by a designation for the location, or place, with the suffix "miut," meaning the people of that place. Thus, the Birnirkmiut are those people who dwell at Birnirk (a location in the bend of a lagoon). The designators for larger groupings that have become standard in Eskimo studies are conventions introduced by field workers and scholars (see Figure 2 for the location of these conventionally named groups). Some names, such as Iglulik, Netsilingmiut, and Caribou Eskimos, refer to a particular subsistence characteristic (*iglu,* or "igloo"; *net,* or "seal"; caribou). Others have territorial referents, such as the North Alaskan Eskimo, Bering Strait Eskimo, and South Alaskan Eskimo. Whether in the first or the second category these conventionalized groupings have become established because they reflect various ecological adaptations within the general Eskimo way of life. The major groupings are as follows:

Social organization among Eskimos

Major Eskimo groupings

Siberian Eskimo (Yuit), including St. Lawrence Island Eskimo, hunters of walrus, whales, and seals, using the whaleboat (umiak) for whaling and voyages.

Aleut, fishers and hunters of seals, sea otters (for furs), and other sea mammals, using kayaks with two cockpits.

South Alaskan Eskimo, fishers (chiefly of salmon) and coastal dwellers in a forested area, using kayaks.

West Alaskan Eskimo (Nunivak Island and adjacent coast, Norton Sound), hunters of seals, walrus, and seabirds and fishers of salmon and other fish.

Bering Strait Eskimo (including Diomede Island and Cape Prince of Wales), hunters of walrus, whales, and seals, using whaleboats (umiak).

North Alaskan Eskimo (Point Hope, Point Barrow, and the North Alaskan Interior Eskimo), primarily whalers but also hunters of walrus, seals, waterfowl, and caribou.

Mackenzie Eskimo, hunters of whales and seals and fishers of salmon.

Copper Eskimo (including Victoria Island) and Netsilingmiut Eskimo (including Boothia Peninsula), hunters of seals (in winter, at breathing holes, and, in summer, from kayaks in open water) and of caribou.

Iglulik Eskimo, hunters of seals, walrus, and caribou.

Baffin Land Eskimo, hunters of seals and sea fowl.

Caribou Eskimo (Barren Grounds, Chesterfield Inlet, Back River), primarily hunters of caribou, but also of musk ox and, rarely, sea mammals, using no boats.

Southhampton Island Eskimo (extinct since 1904), hunters of caribou and seals, using no boats.

Labrador Eskimo, hunters of seals (from the ice edge and in open water from the kayak) and of caribou, and dwellers in or near subarctic forests.

West Greenland Eskimo, hunters of seals, using kayaks.

East Greenland Eskimo (Angmagsalingmiut), hunters of seals, using kayaks.

Polar Eskimo (northwest Greenland), hunters of seals, walrus, and seabirds, using no boats.

There is general agreement among specialists in studies of the indigenous peoples of North America that all of the Eskimo area, including the Aleutians, makes up one culture area in contrast to the various American Indian culture areas south of the Arctic. There also is considerable agreement among specialists that the Eskimo, or Western Arctic, culture area is divided into two major subcategories: a central and eastern subculture area, which extends across the northern Canadian archipelago and includes Greenland and the coast of Labrador, and a western subculture area, which coincides closely with the Alaskan regions, including the Aleuts and also the Siberian Eskimo (Yuit) on the northeast tip of Siberia. Various traits have different forms in the two subregions. Dog teams, for instance, are hitched differently. In the western subregion the dogs are attached in tandem along a central trace (a mode of hitching also used in the Eurasian Arctic). In the central and eastern subregion, dogs in a team are attached to separate traces of varying lengths and run in a fan formation. Kayaks, harpoon heads, house design, and other things show similar contrasts. There are a series of traits, not geared to mere survival but to more sophisticated living, that are found in the western area but not in the central and eastern region. Among these are labrets (ornaments worn in the perforated lip), dance masks, coiled basketry, pottery, grave monuments, mourning feasts, property distributions, war parties, and kin groups larger than the family. Collectively, these make the western cultures more elaborate or complex than those of the central and eastern region. One interpretation is that the western elaboration represents traits inherited from ancestral Asian cultures and that many such traits were sloughed off as the Eskimos migrated eastward and had to cope with a harsher environment that left them little opportunity for traits and practices that could have been considered superfluous.

Major Eskimo subculture areas

Some studies further divide the central and eastern areas

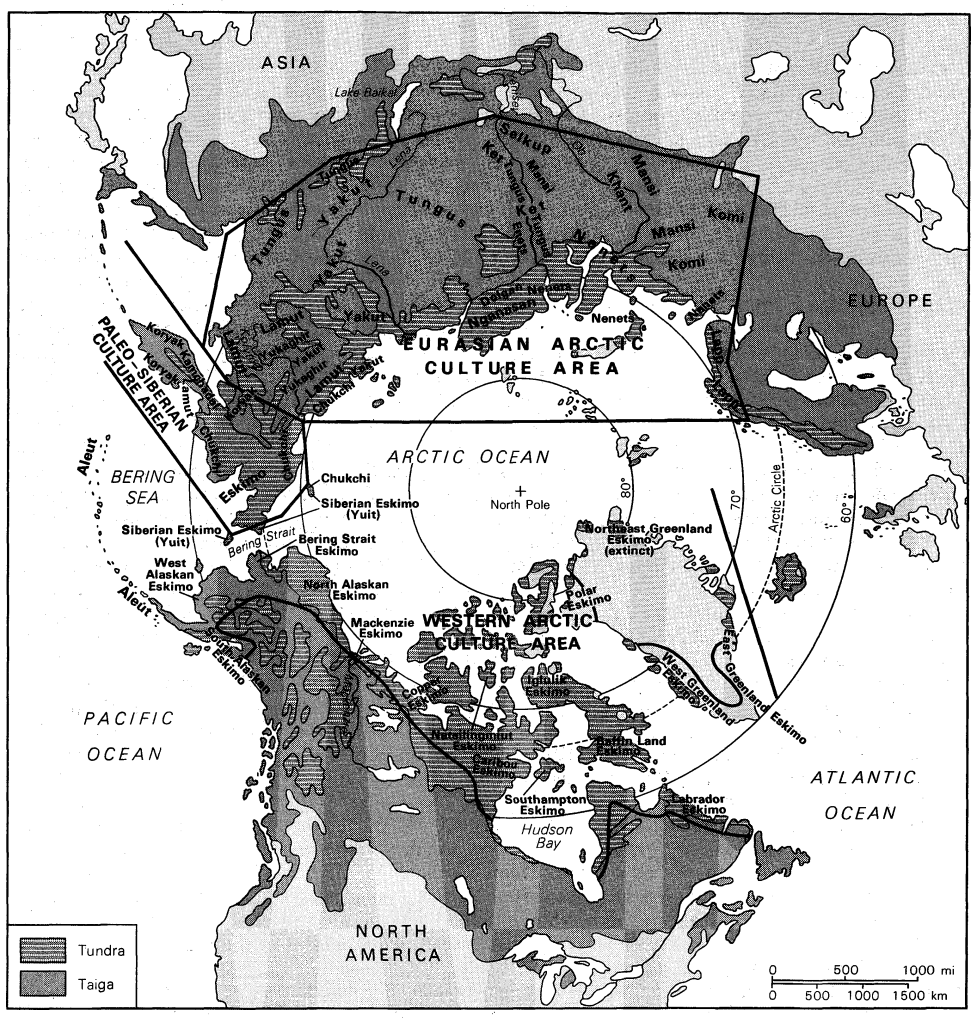


Figure 2: Distribution of Arctic peoples.

into separate categories. The people in the central area are largely nomads who dwell in snowhouses; the people in the east, particularly in Greenland, live in semisubterranean houses in fixed villages of considerable size. It is the nomad of the central groups (especially Iglulik, Netsilingmiut, Copper, and Caribou Eskimo) who provides the popular stereotype of Eskimo life; but actually most Eskimos, in both the east and west, are not nomads living in snow winter houses, but instead they are sedentary village dwellers living in subterranean sod-walled houses. Village living is made possible for the Eskimos by their efficient modes of transportation—dog teams in winter and boats made of skin (kayak and umiak) in summer. Only hunters with good transportation can range far enough to gain sufficient food for sedentary village living. (W.K.C.)

Eastern, or Eurasian, Arctic. Although many hunters and fishers are found in Arctic Eurasia, it is the keeping of domestic reindeer that defines the culture area and gives it unity. The herds and correlated traits, in harmony with the required mobility, set the Eurasian Arctic apart from the Western Arctic, despite many common circumpolar traits. The reindeer complex also differentiates Arctic from non-Arctic cultures in Asia; but some boundaries are blurred, particularly in southern Siberia, by transitions between the herding of reindeer and the herding of other domestic animals, such as cattle, or by transitions to agriculture within single ethnic groups.

The peoples of the Eurasian Arctic more genuinely divide into different ethnic groups (in contrast to the Eskimo groupings); each group has its own language (though many are linguistically related), and each considers itself as having its own name and tradition. The major Eurasian Arctic groups number 19 and go by various names (the alternative names appear below in parentheses):

Major Eurasian Arctic ethnic groups

Lapp (Saamian, Lopari), reindeer herders and fishers, speaking a West Finnic language (of the Finno-Ugric family, a subdivision of Ural-Altaic).

Komi (Zyryan), various groups of reindeer herders, cattle herders, hunters, and agriculturalists, speaking a Permian language (of the Finno-Ugric family).

Mansi (Vogul), reindeer herders, hunters, and fishers, speaking an Ugrian language (of the Finno-Ugric family).

Khant (Ostyak), reindeer herders, hunters, and fishers, speaking an Ugrian language (of the Finno-Ugric family).

Selkup (Ostyak-Samoyed), reindeer herders, hunters, and fishers, speaking Selkup (of the Samoyed family, a subdivision of Ural-Altaic).

Enets (Yenisey-Samoyed), reindeer herders, hunters, and fishers, speaking Entsy (of the Samoyed family).

Nganasan (Tavgi Samoyed), reindeer herders, hunters, and fishers, speaking Nganasan (of the Samoyed family).

Nenets (Samoyed, Yurak), reindeer herders, hunters, and fishers, speaking Nentsy (of the Samoyed family).

Yakut (northern Yakut: Sakha, Jeko), chiefly cattle and horse herders but also some reindeer herders, hunters, and agriculturalists, all speaking a Turkic language (of the Altaic family, a subdivision of Ural-Altaic).

Dolgan, reindeer herders, hunters, and fishers, ethnically a Tungus group but speaking the language of the Yakut.

Lamut (Even, Oven), reindeer herders, hunters, and fishers, speaking a Manchu-Tungus language (a subdivision of Ural-Altaic).

Tungus (Evenk), reindeer herders, hunters, and fishers, speaking a Manchu-Tungus language (a subdivision of Ural-Altaic).

Yukaghir (Odul), a Paleo-siberian group, formerly primitive hunters but now largely reindeer herders, speaking Yukaghir (an unaffiliated language).

Koryak (Nymylan), a Paleo-siberian group of reindeer herders and hunters and of sea hunters and fishers, speaking Koryak (of the Chukchi-Koryak-Kamchadal family).

Chukchi (Chukchee, Luorawetlan), a Paleo-siberian group of reindeer herders and hunters and of sea hunters and fishers, speaking Chukchi (of the Chukchi-Koryak-Kamchadal family).

Kamchadal (Itelmen), a Paleo-siberian group of sea hunters and fishers, speaking Kamchadal (of the Chukchi-Koryak-Kamchadal family).

Eskimo (Yuit), Asiatic Eskimo sea hunters and fishers, speaking Eskimoan.

Aleut (Unangan), Asiatic Aleut sea hunters and fishers, speaking Aleut.

Ket (Yeniseian, Yenisey-Ostyak), western Paleo-siberian hunters, speaking Ket (an unaffiliated language).

Lapp, Komi, and Samoyed

Of northern Finno-Ugrian-speaking peoples west of the Urals, the Lapp and the Komi are the only recognizable entities remaining in the late 20th century. The Lapp comprise small minorities in northern Norway, Sweden, Finland, and on the Kola Peninsula in the Soviet Union; their principal livelihoods are fishing and reindeer breeding. The Komi, by contrast, are concentrated in the northern Soviet Union between the Pechora and Vychegda rivers (southeast of the White Sea); most are agriculturalists, but in the northern part of their area, some are hunters and reindeer breeders. The Komi Autonomous Soviet Socialist Republic has been in existence since 1936.

East of the Komi and on both sides of the Urals lies the traditional area of the Samoyed peoples. Their languages and the languages of the Finno-Ugrians make up the Uralian language stock. The majority of the Samoyed are located along the coast and lower reaches of rivers between the White Sea and the Taymyr peninsula, where they have adapted to a tundra ecology based principally on reindeer breeding, hunting, and fishing. These groups are now known in the Soviet Union as the Nenets, Nganasan, and Enets. A fourth group of Samoyed, the Selkup, is settled well to the south of the others and along the upper reaches of the Ob River. Their nearest non-Russian neighbours are two Finno-Ugrian-speaking peoples who moved to the east of the Urals: the Ostyak, now called the Khant, and the Vogul, now called the Mansi. Earlier, the Selkup, Khant, and Mansi were hunters and fishers only; later, they adopted reindeer breeding (Nenets influence), fur trapping (tsarist influence), and fur farming (Soviet influence).

The Siberian home of most of the Tungus people, who speak what is called a Manchu-Tungus language, is in the vast area between the Yenisey River and Lena River, to the east and south of the Samoyed peoples. These Tungus are now known in the Soviet Union as Evenk. Like the Nenets, they are traditionally reindeer breeders, hunters, and fishers principally; the culture of the Evenk, however, is adapted to the taiga and not to the tundra. There are also the Dolgan, a small Tungus group of 5,100 persons, living north of the Evenk and in the southern part of Taymyr in ecologic conditions approaching those of the Samoyed groups. The Dolgan are now Yakut-speakers. One group of Siberian Tungus lives well to the east of the Yenisey–Lena area; they are the Lamut, now called Even, whose culture and language are held to be closely similar to the Evenk. The dispersed Tungus, or Evenk, population continues to the east and south beyond the Lena River to the Sea of Okhotsk and the Amur River; it is particularly in these areas that Evenk culture becomes blended with Yakut.

The Yakut are distributed on both sides of the middle basin of the Lena and on to the Kolyma River in Northeastern Siberia. The Yakut speak a Turkic language that combines with the Manchu-Tungus language of the Tungus to form the so-called Altaic linguistic family. Only a minority of the Yakut are reindeer breeders, most of them having kept cattle and horses; agriculture and hunting also are traditional practices. The Yakut may have moved northward into their present locations not long before the Russians arrived in that part of Siberia in the 17th century; local groups of other ethnic populations (besides Dolgan and Evenk, also Yukaghir and even Russian settlers) were assimilated by the Yakut, who have always had an important role as traders. Since 1922 there has been a Yakut Autonomous Soviet Socialist Republic.

Paleo-siberians

Far Eastern Siberia is the ancestral home of some of the different peoples known collectively as Paleo-siberians. They do not form a linguistic group of their own, and, unlike all other Arctic peoples of the Eastern Hemisphere, they do not have linguistic affinities with peoples to the south. They appear to have been in Siberia much longer than the other peoples, who are sometimes therefore designated collectively as Neosiberians. The Yukaghir, who inhabit the area immediately to the northeast of the Yakut, are Paleo-siberians who used to practice a primitive hunting economy over wide areas of tundra. Privation, disease, and assimilation (particularly to the Even, Yakut, and Russians) have taken a severe toll on their populations, however, and there are only about 800 persons who call themselves Yukaghir. Many of them are reindeer breeders, and about a third of them regard Russian as their native tongue. East of the Yukaghir are the Koryak and the Chukchi, both of whom are traditionally adapted to the tundra (reindeer breeding and hunting) as well as to the sea (hunting and fishing). Both offered relatively severe opposition to the Russians, the Chukchi not becoming Russian subjects until as late as 1789. Only the Chukchi expanded their territory and numbers in historical times, however, and this was at the expense of the Koryak as well as of the Yukaghir.

Along the Bering Sea are Asiatic Eskimo interspersed with Chukchi settlements. There are also Kamchadal, now called Itelmen, who live along the western coast of the Kamchatka Peninsula; the economies of both these peoples are marine. The languages of the Koryak, Chukchi, and Itelmen are related. Finally, there are Aleut on Commander Island off the Kamchatka coast, and far to the west there are the Ket, who are hunters, trappers, and fur farmers living along the middle reaches of the Yenisey. Both of these peoples are included among the Paleo-siberians; the language of the Ket is the sole survivor of a linguistic group earlier present in the area.

(W.K.C./R.P.B.P.)

Physical types. Two of the major varieties of mankind are represented among the Arctic peoples—Mongoloid (Asian) and Caucasoid (European). The idea that mankind varies in continuous distributions of physical characteristics, which should be plotted in terms of frequency occurrences (clines) of these traits rather than broken into discrete categories (races and subraces), is demonstrated in a general way among the Arctic populations. There is more or less a cline from Mongoloid to Caucasoid extending from Arctic America and eastern Siberia across northern Asia and into northern Europe, with the more generally Mongoloid nearer to the American and Siberian end and the more generally Caucasoid nearer to the European end. Comparative studies of the ethnic populations reveal, however, that the distribution of traits is not strictly continuous; some groups not fitting the pattern intrude here and there, and the cline extremes are not truly at the geographical ends of the distribution.

Biology of major groups. In regard to physical type, the Eskimos are much more closely related to northern Asians than they are to American Indians, as shown by blood types as well as by morphological features (principally facial features and body proportions). Sources variously refer to an "Eskimo" subrace of Mongoloids or to an "Arctic" subrace. When "Arctic Mongoloid" is used as a category, the Chukchi and some other groups of northeastern Siberia are included; however, field workers report that the Chukchi and their neighbours, the Siberian Eskimos, differ enough so that most individuals can be identified without reference to language or other cultural traits; and early sources reported that skulls unearthed from ancient tundra burials showed considerable differences, with Eskimo crania being higher and Eskimo facial breadths greater than those for the Chukchi.

The Tungus and Yakuts

West of the Arctic Mongoloids, the Tungus and Yakuts show some of the highest frequencies of extreme Mongoloid traits found anywhere. They are among the most Mongoloid of all peoples, with low-bridged noses and

forward-projecting cheekbones (malars) that are covered with pads of fatty tissue, resulting in extreme facial flattening. Eye forms show a high frequency of epicanthic (or "Mongoloid") eye folds. Hair is straight, black, and large in cross section; it is abundant and long on the head but sparse in beards and body hair. Skin colour is yellow or yellow brown, and eye colour is consistently brown. In stature the Tungus and Yakuts are short, with the shortness accentuated in lower limbs. Legs below the knee (tibia and fibula) are short relative to the legs above the knee (femur); and elbow to hand lengths (radius and ulna) are short relative to upper arms (humerus). Resulting total limb lengths are short relative to trunk lengths. The type is labelled "classic" Mongoloid or the "extreme" Mongoloid. Broad heads, relative to head lengths (brachycephalic), are also stated to be characteristic of extreme Mongoloids in most discussions of racial types; but this ratio depends on a combination of factors, and changes through time occur within ethnic populations. Many Tungus are longheaded (dolichocephalic), and in recent times there has been a trend in all northeastern Asian Mongoloids toward an intermediate head form (mesocephalic).

The Tungus and Yakuts have been frequently cited as extreme Mongoloids who are placed geographically between the Caucasoids of the Urals and the Caucasoid-like Ainu of the northern Japanese islands; moreover, the Tungus and Yakuts are equally intrusive between the less Mongoloid Arctic peoples who are their immediate neighbours, particularly the Yukaghirs and the Kets. The physical characteristics of these various groups and their distributions are consistent with the hypothesis that the extreme Mongoloids derived from a people who originally had more generalized Mongoloid characteristics (and were more like Caucasoids) and who, isolated in Siberia during a period of extreme cold in the Ice Age, developed the distinctive facial and body characteristics of the extreme Mongoloid as adaptations to the glacial environment. Prior to the development of the extreme Mongoloid, some people of the generalized Mongoloid type had migrated across the Bering Strait area to become ancestors of the American Indians, while other generalized Mongoloids remained in scattered populations elsewhere in Asia. After the ice receded, the physical characteristics of the extreme Mongoloids, though adaptive to very cold conditions, were no handicap in warmer environments and were continued and spread by migrations; but the centre of highest frequency for extreme Mongoloid traits remained near the point of origin. Seen relative to this hypothesis, the Tungus and Yakuts are representative of the extreme Mongoloid type, and the Yukaghirs and Kets are near the generalized Mongoloid in physical type. In appearance the Kets are suggestive of some American Indians. Old photographs of the Yukaghirs show Caucasoid-like facial features, but this is not inconsistent with the idea of a generalized Mongoloid type.

West of the Yenisey River, the Khants and Mansi (Voguls) are basically Caucasoids (as are the Komi and Lapps farther to the west), although the frequency of blood type B is somewhat higher (a Mongoloid trait), and individuals with epicanthic eye folds are more common than in European populations. The Khants, Mansi, and Komi are usually classified as belonging to a "Ural" subrace of Caucasoids. (Soviet sources refer to a "Europoid" type.) The Ural type is characterized by straight, brown or light-brown hair, fine in cross section; light skin colour; mostly light or mixed eyes; moderately thick beards; comparatively short and moderately wide faces; concave or upturned noses; thin lips; medium ratio of head breadth to head length (mesocephaly); and short stature (with a mean of about five feet, three inches [160 centimetres] for adult males).

Compared to their neighbours, the Khants and Mansi, the Samoyedic speakers—Nenets, Nganasans, Enets, and Selkups—have a higher frequency of Mongoloid features, including particularly the epicanthic eye fold, and they are often identified as Mongoloid outliers in Europe and western Siberia. Compared with the extreme Mongoloids, however, they are clearly closer to Caucasoids in many features. They lack pronounced malars and have higher nasal profiles, and their facial proportions are more comparable to Europeans than to Asian types. Clearly, they are intermediate in the continuum found among Arctic peoples.

Caucasoid features of Lapps

The Lapps of the Scandinavian countries and northwestern Russia are now recognized as Caucasoids. Formerly, they were not always so classified, probably mainly due to the marked contrast with their Indo-European-speaking Scandinavian neighbours, who show a high proportion of tall Nordic types. Whereas the Scandinavians, particularly the people of Sweden, are the tallest of Caucasoids and among the tallest groups of mankind, the Lapps are among the shortest, with a mean stature of only five feet, three inches for males and four feet, 10 inches for adult females. Some sources formerly classified the Lapps as Mongoloids, but they lack most sorting criteria for Mongoloids, having only a moderate incidence of B-type blood (an Asian trait), shovel-shaped incisors (also Asian), and other dental characteristics to suggest this affiliation. Some authorities have even assigned the Lapps to their own individual major racial category, thinking of them as a remnant of some archaic Caucasoid-like people. In most traits, however, they are close to the Ural type, and systematic assessment places them among the Caucasoids. Their Mongoloid traits are terminal effects of the Arctic Mongoloid-Caucasoid cline.

In most physical characteristics, the Eskimos are close to the extreme Mongoloid type. They also share the general characteristics of short stature, thick torso, and small extremities found among all the Arctic peoples. Uniquely Eskimoid features include very broad faces, thin noses, and high cranial vaults with sloping sides ("gabled" parietals). Eastern Eskimos and the crania from old Eskimo archaeological sites are longheaded (dolichocephalic), and for this reason many summaries characterize the Eskimos as the longheaded subtype among Mongoloids. Not all Eskimos, however, are dolichocephalic. Group means vary from dolichocephaly through mesocephaly to the lower range of brachycephaly. Anthropometric studies also report variations in stature from group to group (from about five feet, two inches, to five feet, six inches, for adult males).

General adaptive features. Considered as a whole, the Arctic peoples share certain physical features regardless of the predominance of Mongoloid or Caucasoid traits. The most evident of their shared traits is their small size—or, more correctly, short stature, since such people as the Chukchi are described as being massive in build, though only two or three inches above five feet in mean stature for adult males. Other shared characteristics include large girths relative to body lengths, short limbs relative to torso lengths, short lower legs and lower arms relative to upper limb segments, and small hands and feet. All of these characteristics are in harmony with general biological principles regarding adaptation to cold environments.

Cold-adaptive body mechanisms

Experimental studies, particularly since World War II, have identified at least two specific cold-adaptive body mechanisms (largely vascular) that have become hereditary among the Arctic populations. One of these results in increased blood flow in exposed body parts, such as the hands, and heightened basal metabolism to keep blood temperatures within tolerable limits (found among Eskimos and other Mongoloids in the Arctic). The second form of cold-adaptation involves shunting of blood from one artery to a paired artery and warming of venous blood leaving the body extremities before it returns to the heart. Thus, the extremities can be reduced in temperature without affecting more critical body parts such as the brain. (This second cold-adaptive characteristic has been found among the Lapps.) Even before these cold-adaptive features were identified by experimental physiology, they were acknowledged by ethnographers in statements asserting that the Chukchi men could work for considerable periods with bare hands when temperatures were −30° F (−35° C), or that the Chukchi would strip to the waist and complain of the heat if the temperature rose much above 40° F (4° C).

ARCTIC CULTURES

Two major manners of earning a living are found among the cultures of the Arctic—that of the hunter and that of the reindeer herder. In the Western, or New World,

Arctic, only hunters are found and all are Eskimos (including the closely related Aleuts of the Aleutian Islands). Hunting cultures also occur in the Eastern, or Eurasian, Arctic; but more typical are reindeer herders distributed in a great arc across northern Eurasia, from the Chukchi of Northeastern Siberia to the Lapps of northern Europe.

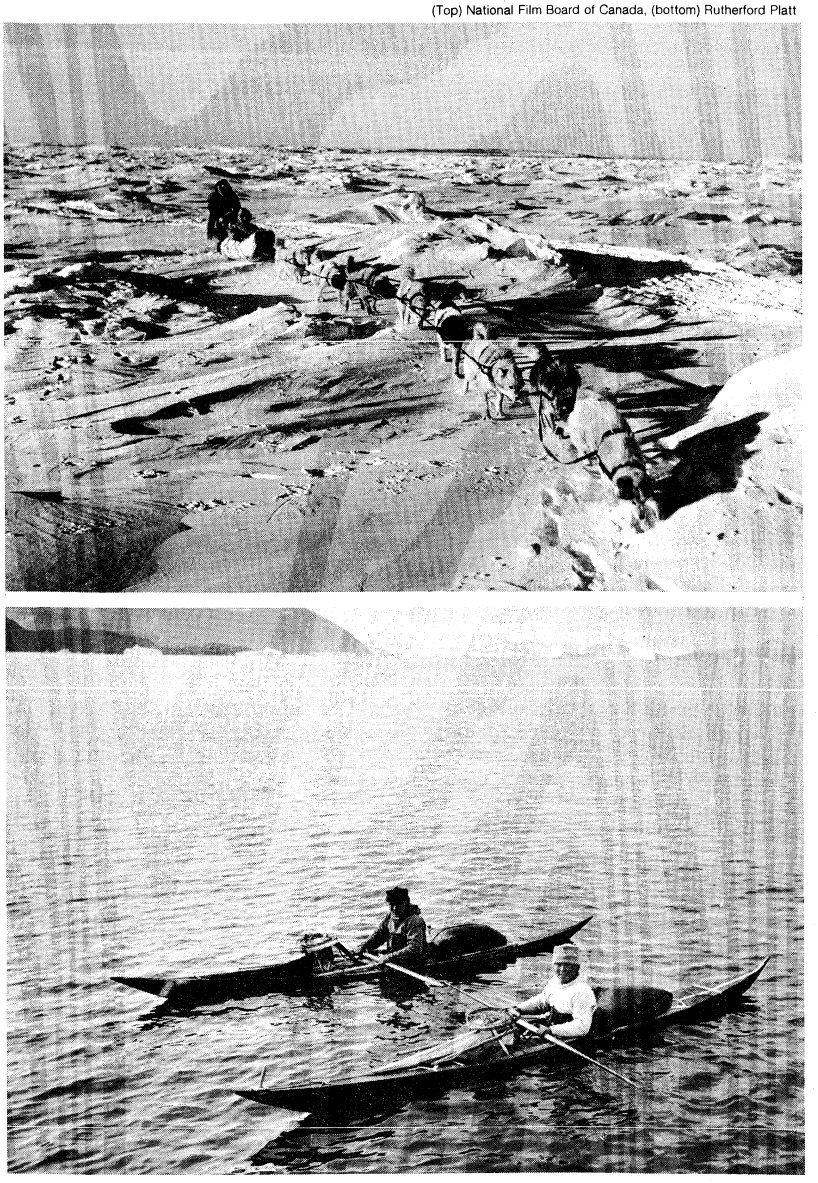
When defined as a culture area, the Arctic includes some regions that are subarctic in climate. Physical geographers often place the southern boundary of the Arctic either at the Arctic Circle or at a line marking the regions where the warmest month has a mean temperature that is not above 50° F (18° C); in Eurasia this coincides closely with the northern limit of trees and represents a real division in vegetation, climate, and habitat. But this geographical boundary is not a cultural boundary, since most of the Eastern Arctic cultures are found in both the treeless desert-like tundra of the Arctic and the northern reaches of the swampy, coniferous forests, or taiga, of the subarctic. Some groups live in the taiga and some on the tundra, and others are seasonal nomads (transhumants), wintering their reindeer in the taiga and moving them in the spring and summer to graze on the tundra along the Arctic Ocean. Considered as a culture area, the Arctic is determined by relative uniformity of practice and custom, not solely by geography.

In many ways the Eskimos are the people most adapted to the Arctic, the people most specialized for Arctic living. Most Eskimos are coastal dwellers, specialized in hunting sea mammals, particularly the several species of seals, and in some places walrus and whales. Fish and birds also are important. Although there are some inland Eskimos in northern Alaska and Canada who depend especially on hunting the caribou, a wild relative of the domesticated reindeer of Eurasia, an orientation toward the sea characterizes the patterning of most Eskimo cultures. And although some Eskimos live in the forested areas of the subarctic (in Labrador and southern Alaska), the characteristic Eskimo settlement is along a coast of the Arctic Ocean or along some northern strait with tundra at the back of the village. Most Eskimo villages are between latitudes 60° and 72° N. The most northerly of the Eskimos—and most northerly of all the world's peoples—are the so-called Polar Eskimos of northwestern Greenland, who are found as far north as 79°.

Ecology and cultural adaptations. The most demanding environmental conditions for Arctic peoples are the cold and high winds, the seasonal variations in the hours of daylight, the limited numbers of natural species, both flora and fauna, and the restricted range of raw materials for making houses and tools. Contrary to widely held impres-

Character of the Arctic environment

(Top) National Film Board of Canada, (bottom) Rutherford Platt



Modes of Eskimo transportation.
(Top) Copper Eskimos drive a dog sledge near Cambridge Bay, Victoria Island, Canadian Arctic Archipelago, Northwest Territories. (Bottom) West Greenland Eskimos in kayaks on Baffin Bay, off Greenland.

sions, the people who live on the tundra are not buried in the snow. Precipitation is very light in much of the Arctic (often less than four inches of water a year). The Arctic would be recognized as a desert except that the long, cold winters and short summers reduce evaporation and prevent runoff of groundwater or ice; much of the land is permanently frozen soil, or permafrost. Expressed in temperatures, the cold is more intense in the boreal forests of the interior taiga; but on the tundra high winds persist throughout the year, and in winter the resulting windchill makes exposure critical and the needs for protective housing, clothing, and equipment more demanding. (See above *Climate.*)

Arctic habitations, clothing, and equipment. For those peoples who still follow their traditional way of life, survival in such conditions requires highly functional housing and clothing. The provisions to meet these requirements are particularly revealing. The construction materials have to be those readily available in the immediate vicinity.

Types of housing

Some Eskimos use snow blocks to build the domed house widely identified as an igloo; but most Eskimos (before the introduction of prefabricated structures in the Arctic) lived in earth- or sod-walled structures built on a frame of driftwood, sometimes with large whale bones as structural members. Many of the design features to offset cold and wind are common to both the snowhouse and the semi-subterranean sod house: both include an entrance shelter or storm shed, frequently long and narrow, to block cold winds, and a raised bench ("sleeping bench") to keep the occupants up out of the colder levels of the house. The sleeping bench is the centre for many activities other than sleeping during the long winter nights, activities such as eating, making tools, and mending equipment. In summer, the Eskimos move out onto the tundra, living in skin tents.

For the nomadic or seminomadic reindeer herder of the Eurasian Arctic, the winter habitation is more apt to be a tent, in many cases a double-skin tent (subterranean houses are possible only among some taiga pastoralists who can winter their herds in a sheltered area and keep them in one place for several weeks or more). The Eurasian tent, which is used both summer and winter, is conical (somewhat resembling the tepee of the North American Plains Indians) and is covered with reindeer skins on the tundra and with bark in the subarctic forests. The winter tent is sometimes banked with sod and often, if not always, with snow. Some are fitted with raised benches and with storm sheds. The main tent surface may be doubled with insulating materials between the layers; in other cases, as among the Chukchi and Koryaks, there may be separate boxlike sleeping shelters within the main canopy, each for several occupants. Toward the western end of the Eurasian area, approaching northern Europe, the winter house is more akin to European prototypes, such as the Scandinavian log house, usually with a flat roof that is earth-covered; some Lapp herders of Scandinavia, however, spend winters in conical tents banked with sod and snow. The Arctic Eurasians who do not depend on reindeer but are seamammal hunters and fisher folk, and thus largely sedentary, live in semi-subterranean winter houses.

Arctic clothing

Clothing is no less critical (and perhaps more critical) than housing for survival in the Arctic. Eskimo clothing, similar in general to clothing in the Eurasian Arctic cultures as well, is so suited to the environmental conditions that until recently none made elsewhere could equal it, and explorers and others from the outside world adopted it after a short period in an Arctic winter. The general aspects of Eskimo clothing are well known—garments fashioned of the skins of fur-bearing animals with emphasis on a pullover jacket with an attached hood (parka) and bootlike footgear with watertight seams. The parka and trousers (worn by both sexes) are double-skin garments—two layers of skins placed back to back with the fur surface of the outer layer on the outside and the fur or hair of the inner layer inside, against the wearer's skin. Caribou skins with the hairs at just the right stage of development are selected. Not only does the hair of the inner layer provide warmth, but air trapped between the short, fine hairs of the inner caribou skins provides additional insulation. Other functional aspects of clothing design provide for ventilation and perspiration control. Even at temperatures well below zero, a man exerting considerable effort, as in running after a dog sledge, often perspires. If he is wearing conventional undergarments, these inner garments become damp with perspiration; and when the activity rate is decreased and the body cools, the perspiration-dampened clothing freezes. In the fully developed Eskimo garments, drawstrings are provided at the end of each sleeve and at the opening for the face. A belt closes the lower edge of the parka. If the wearer perspires, he loosens these bindings, permitting the outside air to percolate through the hairs of the inner fur layer. The cold, dry, ambient air, which is warmed by passage over the person's body, becomes extremely dry and quickly absorbs the moisture. With the perspiration stopped, the wearer pulls the bindings tight again to prevent overcooling. Other items of Arctic clothing include various types of footgear for differing weather, outer rain garments to keep the exterior fur layer dry, and mittens, each item with design features providing similar functional fit and solving particular environmental problems.

Technology and equipment

It has become customary to state that Arctic dwellers were able to occupy their difficult environment only through application of much ingenuity. Other references credit the Eskimos, for instance, with the invention of the toggle-headed harpoon (one whose spurred head detaches from the shaft after piercing a struggling animal); the semi-subterranean house; the composite bow (one of wood, bone, and sinew glued and lashed together); the throwing stick, or atlatl (a rodlike device for hurling darts or spears); and other items of material culture. In terms of skill in manufacturing and eagerness to adopt new items, Arctic peoples were indeed ingenious; but this does not mean that they invented all of the items that they used. With the exception of the domed snowhouse, a trait exclusive to the Eskimos, even the most typical items in the material culture of Arctic peoples were widely shared with other peoples in more southerly areas.

A list of material items used in both the Eurasian and Western Arctic would include the semi-subterranean house with entranceway and sleeping bench; tailored clothing, including the parka; the toggling head and other harpoon parts; blunt arrowheads for birds and small game; the atlatl; barbed darts for taking birds; the semicircular slate knife (ulu); snow goggles with slits as eye openings; sledge forms; dog teams for traction; skin boats both large and small (kayak); the bow drill; the composite bow; ice picks for harpoon shafts; the tambourine-type drum used in healing rites; and so on. The semi-subterranean house is shared with American Indians south of the Eskimo area and distributed across the Eurasian Arctic. The hooded parka was used as far west as northern Europe and can be identified in art objects far back in time (as in the Siberian Neolithic Period). The toggling harpoon head was used in hunting sea mammals by some prehistoric and later cultures of northern Europe. (W.K.C.)

Traditional culture patterns. *Western Arctic.* The cultural and social life of the peoples of the North American Arctic can in large measure be interpreted in terms of an adaptation to Arctic hunting. Theirs was a highly successful material adaptation, marked by an ability to exploit an essentially hostile environment. A price had to be paid, however, for the stress on technology and hunting. Because the Eskimo lacked agriculture and, for reasons unknown, were uninclined to exploit whatever wild vegetation was available, their populations and settlement patterns were limited by the available game. In general, they stressed cooperation among close kindred in the pursuit and use of game—this necessarily at the expense of any tribal definition or solidarity. They had to be pragmatic, relying on the physical strength, technical ability, and survival skills developed in their own small groups.

Importance of hunting

It is not wholly accurate to describe the Eskimo culture as nomadic; although a group might range widely through a given territory in seasonal search for food, a group knew its own territorial range and generally kept to it. Every cliff, rock, bay, and discernible hummock of tundra land was named, and it was to this familiar terrain that

the individual and family were generally bound. On the other hand, this attachment to a traditional area was not immutable. People could on occasion move and establish relations or affirm kinship ties with other groups.

The family group

The primary unit of affiliation lay in the family, not only in the nuclear family of spouses and children but in the extended family in which both the maternal and the paternal lines figured. The system made no provision for an elaborate genealogical accounting; indeed, one's relatives were those who stood close to one in both reckoning and residence—brothers and sisters, parents and their siblings, the children of such, one's own children and grandchildren. One might try to keep alive a sense of relationship with a person who had moved to a different group for marital or other reasons, but usually such ties eroded over time. In line with the eminently practical sense of the Eskimo, one's relatives were those with whom one could interact directly, from whom one could freely ask support and have it demanded in turn. This feature in turn gave rise to the sense of collective responsibility, the sense of loyalty and obligation to one's resident relatives, which so vitally underlay Eskimo societal and legal organization.

Alliances among kin were nevertheless sought beyond domain of the immediate family. Local groups of Eskimo and Aleut developed mechanisms for the formation of cooperative alliances, those designed to effect success in collective hunting, as well as in the exchange of goods and services. The individual could thus make demands on affinal kindred, persons who were related to him through marriage.

Marital patterns

The basis of marriage lay in a fairly sharp division of labour between the sexes; the men hunted, made tools and household goods, and built the homes; the women cooked the food, dressed the skins, and made the clothing. Each partner owned his own personal goods, such as clothing and tools; shelters and food might be held in common, though men had personal rights to the kayak and the umiak (an open boat) and to sleds, sledges, and dog teams. Indeed, the whole rationale for marriage among the Eskimo lay in the recognition of such sexual division of labour and property.

In general, Eskimo marriage was not arranged but open to freedom of individual choice. A young man could settle down with an unmarried woman, either within his own family grouping or hers; children might be born to the couple; but, until there was public recognition of the marital status, legal marriage did not in fact exist. It was not unusual for women to cohabit with several men in succession or even, as in the central regions, with several at the same time, giving rise to a kind of polyandry. Actually, the birth of a child, increasing as it did the cooperative circle of kindred, tended to give some permanence to the marital union. At that, however, a woman whose fertility had been demonstrated was much in demand as a wife, especially if she exhibited the cardinal virtue of industriousness. Divorce under such circumstances was relatively simple, meaning no more than that one or both partners broke the residential unit to find a new spouse. In cases in which there were children, however, even divorce did not entirely break the relationship, the child having relations with both his paternal and maternal kindred and they to each other through him.

Although polyandry could occur, it was almost never of the fraternal type; that is, the husbands could not be brothers. Similarly, two or more wives could reside with a skilled hunter in polygynous unions, but the wives were rarely sisters—the point being that sisters should marry divergently so as to extend the cooperative bonds of kinship. A man, it is true, might be responsible for his wife's unmarried sister and engage in sexual relations with her, but she was not his wife and could leave almost at will. An interesting feature among the Eskimo was the stress on competition for women; a man might easily lose a wife to an abductor, particularly if the union lacked strong support from his kindred.

An important mechanism for ensuring alliances was wife exchange—that is, the arrangement whereby a husband allowed another, unrelated man to have temporary sexual relations with his wife, with the understanding that the favour would be reciprocated. Partly this was an expression of hospitality, but the sharing of wives between two men could initiate and foster a lasting partnership and friendship. Lacking either a tribal or national sense, the various Eskimo groups were obliged to devise means whereby kinship right and privilege could in some measure be extended to persons unrelated by blood or marriage. The problem, as viewed by the Eskimo, was that anyone not related by blood or marriage was a potential enemy.

Legal mechanisms

Involved in these issues of kinship and alliance was the concept of blood feud. Because there were no chiefs or headmen, no formal arbiters of disputes, and clearly no sense of broad territorial community as such, there had to be means whereby any group could be held responsible for wrongdoing by one of its members. The means was the blood feud, whereby not only the malefactor in a group but also his kin could be pursued by an outside group that had been wronged. One result of this practice was to deter wrongdoing in the first place; group opinion tended to restrain any individual from putting the whole group into jeopardy. Another more fearful result was that a vengeful killing might touch off a series of counterkillings—bloody retaliations back and forth between the groups, a true blood feud. Fear of the blood feud, it may be said, probably played an important role in keeping populations territorially apart. Finally, the blood feud, with its implicit element of collective responsibility, also served to enforce family solidarity.

There were of course other means of social control. Public opinion and public ridicule and shaming were effective in maintaining order. There was also, among the central and eastern Eskimo, a kind of surrogate blood feud called the song duel, in which, at a public gathering, two rivals would voice their quarrel by insulting each other in extemporaneous song. The loser, bested in the singing, was often forced to leave the community. Similarly, men might wrestle to settle disputes.

Among the Eskimo there was a large measure of freedom of individual choice in economic pursuits; that is, men were free to seek whatever way they could find to eke out a living in the harsh environment. This freedom, however, tended to be socially divisive; it could mean the breakaway of certain individualistic members of a group. It was not unknown, for example, for a man to leave his wife and children to settle in another place in order to take advantage of economic opportunities there. While away, he might marry again, produce a new family, and ultimately lose contact with his original group. A wife had no recourse in such a case but to fall back on the resources of her own kindred; she too might marry again. Nor did the authority of a parent extend to the demand that a child follow a particular activity. Everyone had the primary goal of seeking food, but each person was free to realize the goal in whatever way suited him best. A result was the beginning of some economic and labour specialization. One person might specialize in a certain kind of hunting or trapping; another might excel in organizing a collective hunt; still others might make weapons, boats, or other items that could be traded for food. The sole demand was that each individual produce.

Trade

Trade and barter figured prominently in the native cultures. Crafted items could always be exchanged informally within a community, but beyond this there were the very important trading patterns that developed between groups—as, for example, in northern Alaska, where the inland caribou hunters traded skins and furs to the maritime residents for pokes of seal oil, ivory, whalebone, and other products of the sea. In the west, these trading arrangements took on the form of elaborate ritualistic feasts called "asking feasts" or "messenger feasts," which, while not basically religious, involved a strong element of ceremony. In these ceremonies, one man of demonstrated skill and leadership publicly feasted his counterpart from another community, and in the process there would be a large exchange of surplus goods, all attended by humour and banter.

Food resources varied across the vast expanse of the Arctic regions. Maritime Eskimo hunted seals of various species and populations. The Aleut sought principally the

sea otter but also hunted seals and whales. All along the north Alaskan coasts, west and south of Point Barrow to the Bering Strait, the natives specialized in the tracking of whales. Other groups hunted the walrus (both Pacific and Atlantic subspecies), polar bears, and various Arctic birds. Many Eskimo groups, in northern interior Alaska and Canada, hunted the caribou, which were especially prized not only for food but also for their skin and fur. Sea fishing was not very important, except perhaps among the western Eskimo; freshwater-stream fishing was important almost everywhere.

Against land animals the Eskimo used the bow and arrow, along with a great variety of traps, and against sea mammals they used the retrieving harpoon. Some groups of Eskimo also used rawhide nets for capturing seals, as well as hooks and tridents, weirs, and traps for gathering fish. From the skin-covered, one-man kayak they harpooned the seals in the sea and speared the swimming caribou in the lakes and rivers. This vessel was not practical, however, for whaling; for this, Eskimo employed, in summer, the umiak, a larger boat, also skin-covered, which resembled the European whaleboat and which served otherwise to transport the family and its possessions. Winter transport was provided by dogsled.

Group hunting

The hunting of such large and often elusive game as whales, walrus, and caribou usually required the efforts of a large expedition, often larger than the labour resources a single kinship group could supply. In such cases the mechanisms of alliance between different groups would come into play. The means could be informal; a man of proven skill and ability as a leader of hunts could randomly attract a following of both relatives and nonrelatives for a specific enterprise. Or the means could be more formal (as it was most characteristically among western Eskimo); a would-be leader might bribe men to join his hunting crew, and the relation between leader and crew member might be further cemented by temporary wife exchange. Whether formal or informal, however, the leadership lasted for only the length of the hunting season or the particular enterprise.

Individual hunting alternated with group hunting. An individual hunter, for instance, might snare seals under the ice during the dark winter days, join a crew for caribou or sea hunting in the spring, move with his family in summer to a fishing site, engage in trade with others, and then return to a permanent encampment in the winter months to begin the cycle again.

It seems true of the Eskimo, as of many other nonliterate peoples, that there was little interest in either the problem of origins or a life hereafter. There was a vague concept of reincarnation but not a well-defined eschatology. Indeed, the Eskimo were less concerned with the dead or less afraid of the dead than many of the world's peoples; a corpse was frequently left for the wolves. It was rather the animal domain that excited interest and about which myth evolved. If there was any ultimate philosophical notion, it lay in a concept of the continuity of life flowing from animal to man and back again into the animal domain. All animals were endowed with human characteristics and, indeed, were considered morally and intellectually superior to men. The beasts and birds allowed themselves to be taken by the hunter either because they felt pity for helpless man or because man, through ritual and magic, could attain some degree of temporary control over them. In any event, they had to be placated, respectfully treated, and accorded the proper ritual handling. To offend animals was to invite sickness, accident, and death. It was hardly sufficient to hunt without making adequate preparation in the way of ritual. Each hunter owned magical songs that were sung over weapons, boats, amulets, any devices relating to hunting. These songs were viewed as property and could be traded or sold.

Endowing animals with spiritual force

There were countless regulations and taboos, those relating directly to food being the most numerous and complicated. They differed from region to region, but everywhere had the same purpose, to avoid antagonizing the spirits of animals and nature and thus preserve harmony between man and the environment. Sometimes the taboos were collective; all members of a group might be forbidden the flesh of wolves or killer whales, for example. Or the taboos might be personal; an individual's song, talisman, or name might be operative, for instance, only if he avoided the flesh from the front flippers of the ribbon seal or from the back fat of the female caribou. Personal prohibitions were particularly intense in the western regions.

Faced with the hard practical realities of living, the Eskimo were not much given to speculating on the nature of the cosmos or devising explanatory myths. In the eastern regions, there was a supernatural woman of the sea, Sedna, who was looked upon as a kind of deity because she was presumed to control the supply of seals, but she was not worshipped; the Eskimo sought to control or curb her by use of magic or incantation. In the West, there was an anthropomorphic raven who was presumed to have created the world; but it was more the subject of trickster tales than a true divine figure. In general, the purpose of Eskimo myths and folktales was to define or illustrate relations between man and environment, not to provide a corpus of actual belief.

Communal festivals generally centred on the hunt. The whaling cult in northern Alaska, for example, brought men together well in advance of the hunt itself to clean and refurbish the gear, to clean out the meat cellars, and to sing the ritual songs associated with whaling. Later, in the hunt, the boat owner and crew leader greeted the whale, giving it fresh water to drink, a common treatment of sea mammals, and generally acting in a priestly role. Such ceremonial festivals (others were conducted for the seal and the caribou) not only brought groups of people together in meaningful social ways but also affirmed the continuity of life and its source, the continual renewal of the world.

Shamanism and illness

Illness resulted from the violation of taboo or ritual prohibition or because of the malevolence of an enemy; death, in any case, was not considered natural. The Aleut performed autopsies to determine the ritualistic cause of death and so to locate the nature of the infraction or the person of a magician. In this context arose the shaman. The term is drawn from the Tungusic languages of Central Asia and refers to a person who enters into various ecstatic or spiritually possessed states and, in such states, is capable of many feats of magic and can cure or send diseases, control weather, and find lost persons and articles. Not a priest, the shaman was the principal focus of day-to-day religious activity among the Eskimo. In this sense he differed from a hunt leader who, in a priestly role, might make an offering of thanks to game for allowing itself to be taken.

Illness resulted when the soul, a kind of spiritual double of the individual's personality, departed from the body and went wandering; it would ordinarily leave because it was offended by the actions of the body (which may have violated a taboo or a regulation regarding animals). The shaman's task was thus to find the soul and entice it back into the patient's body. Alternatively, illness could be caused by the actions of a hostile shaman, who would steal the soul. This ability not only to cure but also to cause disease did not make shamans wholly popular. Feared by the general population, shamans ran a risk of being killed by an offended patient or the patient's kinsmen. Finally, another cause of disease was the penetration of the body by some foreign object, either real or imagined; the shaman, in a trancelike state, located the object and operated to remove it, singing as he did so. (A broken bone, a toothache, or a headache, it should be noted, might not qualify as illness and could be treated by anyone with the necessary practical skills.)

In spite of their severe environment and their severe belief about it, the Eskimo remain a generally cheerful people, given to much joking and entertainment. Their cheerful temperament undoubtedly contributed to their ability to endure the great natural perils without complaint or fear and to find happiness in a land that others have thought uninhabitable. (R.F.S.)

Eastern, or Eurasian, Arctic. From earliest times to the present day, the way of life of many of the Eastern Arctic peoples has centred upon the reindeer; indeed, they have had to adjust their lives to the pecularities of the species.

Important in this regard are the seasonal migrations of reindeer between tundra and taiga; their movement in herds; their seasonal preferences for certain grasses, mosses, and fungi; their dependence upon salt; and their relative helplessness—and panic—when attacked by mosquitoes and other flying insects in the summer and such predators as the wolf and wolverine in the winter. The situation has always given rise to a number of similarities in all types of reindeer economy, among the more important of which are the seasonal movements of camps and the seasonal differences in their size and composition.

Reindeer hunting, breeding, and herding

There are, however, distinctive ecological and economic differences among the Arctic peoples of the Old World, relating to each of the three principal ways in which they utilize reindeer. The traditional Paleo-siberian Yukaghir, for example, were simply reindeer hunters. Although they "monitored" the seasonal movements of the reindeer, they made no effort to herd the animals or to use them for transportation (the Yukaghir adopted both these practices later, however, after contact with neighbouring peoples, particularly the Even). They used decoy reindeer in the hunt and travelled on foot or ski, by dogsled (though this was never general), and by canoe or raft. Reindeer were by far their most important source of food and of skins for clothing and tent covering, so that their lack of control over the condition and movement of the herds left the Yukaghir susceptible to unpredictable periods of privation. Fishing, the hunting of other species, and berry gathering were also essential to the economy.

Another basic way of using reindeer is represented by the many Tungus groups that partly controlled reindeer breeding as an adjunct to a basic economy of hunting and fishing in the taiga. Geldings were ridden, used as pack animals, and sometimes harnessed to sleds; females were milked, making possible various dairy products; dead reindeer were an important source of skins, bone, antler, and sinew for clothing, shelter, and simple utensils. These Tungus secured their meat, however, not from their domesticated reindeer but from wild reindeer, Asiatic red deer, and elk that they hunted. The domesticated herd was always small (herds of up to 25 reindeer per household were usual); the animals usually pastured in the vicinity of the camp, securing protection from wolves there in the winter and finding relief from the summer mosquitoes in the smudge fires provided by the camp for that purpose. These hunting and herding Tungus, it may be noted, also used dogs—as camp sentinels to warn against predators or strangers and as "pointers" in hunting small game such as sable or squirrel.

The third basic method of using reindeer was controlled breeding and herding. This was the principal or sole livelihood of several other Tungus groups (many Nenets and Enets and some Nganasan), of the tundra Chukchi and Koryak in Northeastern Siberia, of the northern Samoyed in Western Siberia, and of many Lapp. All these groups depended upon the herded reindeer for food, and many used dogs to shepherd the herds, which sometimes were large in number. Prestige and social status, indeed, were measured by the size of a household's herd; and competition for prestige sometimes had damaging ecological consequences: pastures would be overgrazed or aggressively expanded, or epidemics would decimate large, closely packed herds. As a result, impoverished reindeer owners would be driven to find alternative livelihoods in villages or elsewhere; some would revert to reindeer nomadism. (State-imposed regulations in both the Soviet Union and Scandinavia are ameliorating such problems.) As a supplement to herding, most of these peoples pursued fishing in the summer, and they bartered extensively with urban, village, and agricultural populations.

The peoples of the Eastern Arctic have been studied and written about far less than the Eskimo of the Western Arctic. In tsarist Russia and the Soviet Union, in particular, there has always been some tendency to play down any attention to national differences. Most investigations have thus depended on inferential sources. They have tried, that is, to reconstruct the social relations of certain Eastern Arctic peoples by piecing together information derived from myths, legends, and other folklore that have been published. They have also tried to infer the social systems of Arctic peoples by reference to the social structures of their Turkic and Mongol neighbours and kinsmen farther south, who have been more thoroughly studied—the assumption being that the systems are probably similar.

Tribes and clans

Among all the Eastern Arctic peoples, there appear to have been large territorial groups that may be called tribes—*i.e.*, clusters of communities that could act together for such common purposes as warfare and trade. One tribe could make alliances and establish systems of exchange with other tribes. Below the level of the tribe, at least among the Neo-siberians, there were probably clans, groups of people having a stricter sense of kinship. Normally, clans would be exogamous; that is, marriage within the clan would be forbidden. Further, a person's lineage within the clan would be important in determining his residence, role, and inheritance.

From the 17th century onward, however, clans—and even tribes—had to leave their traditional territories and disperse in smaller units and "mixed" kin units to search for new territories. This happened to the Tungus when they were faced with the exhaustion of reserves of fur that had to be supplied to the Russians and when they were faced with the continual encroachment of Russian settlement. The movement of Tungus groups created pressures on the territories of other neighbouring peoples; the northern movement of the Yakut had similar effects in Northeastern Siberia. The competitive expansion of grazing lands could also disrupt traditional social relations, as it apparently did among the northern Samoyed. Various demographic and economic changes, in short, could have serious social consequences. Finally, there was the political influence. The tsarist Russian government tried to consolidate the tribe–clan system and use it to administer all northern peoples, but the administrative use of tribe and clan, being a foreign idea, tended paradoxically to sever the already weakening association between kinship and territorial organization. Among most groups, therefore, the village or camp emerged as the relevant territorial and social unit in the 19th and early 20th century.

Dominance of village and camp

Village or camp enterprises of a seasonal nature generally involved each of the three principal livelihoods: reindeer breeding, hunting, and fishing. When hunting and fishing was pursued collectively, the game or catch was distributed collectively. When production was not collective, responsibilities for distribution were limited, and they were minimal or absent when furbearing animals were caught for trade with foreigners. Villages and camps (and subdivisions made within these units) claimed rights to the produce from hunting and fishing grounds usually on an annual or lifetime basis. Among reindeer breeders, herding was undertaken collectively, and rights to pasture were claimed collectively, also. The animals, however, were owned individually. The size and composition of herding camps could be adjusted at each season of the year.

In all ethnic groups, close interpersonal bonds of friendship arose between non-kinsmen within the village or camp. Kinship nevertheless remained in many ways the most important principle of interpersonal and group relations. Many privileges and obligations of kinship practiced by the Eskimo were also recognized by Eastern Arctic peoples. Among the more onerous was the prosecution of the blood feud, a retaliatory killing and counterkilling between different kin groups that might continue for generations. Blood feuds seem to have been widespread even at the end of the 19th century, indicating that clanship was still a recognizable principle but without any territorial meaning, for villages and camps were then composed of persons from different clans.

The incidence of cross-cousin marriage, in which certain categories of first cousins were obliged to marry, remained notable among many of the eastern Arctic peoples. Practiced in conjunction with the general rule of patrilocal residence, this marriage preference had earlier supported the territorial groupings of persons in clans; the preference remained even after the development of villages and camps in which cousins remained anyway, without the compulsion of marriage. The matching of several siblings of one family with those of another and a close bond

between husbands of sisters—sometimes denoted as closer than that between brothers—were prevalent among the northern Eurasians. So, too, was the injunction that a bridegroom provide several years' service to his father-in-law before being allowed to set up his own marital household. Common among the reindeer breeders only was the establishment of marriages through contributions of wealth (including reindeer) from both the parties (bride price from the groom, dowry from the bride). Having two or more wives was a privilege of personal rank and occurred most frequently among preeminent hunters or warriors, shamans and smiths, and well-to-do reindeer owners. These marriage practices (except those prohibited by law) persist among some contemporary Lapp groups in Scandinavia. Among the Lapp, building up a reindeer herd was probably the most influential factor governing marital patterns.

In all these societies there have been strains of social and economic inequality. At the very least, the need for authority in organizing work teams promoted village or camp elders, local headmen, lead hunters, and so on; among some peoples—the Chukchi are notable—an ethic of success and of male reputation licensed the presence of "strong men" and aggressive behaviour. Intertribal warfare, fairly general though sporadic throughout northern Eurasia, led to the use of captives as slaves. Among the reindeer breeders, differential accumulation of herd capital placed some herdsmen in employer–employee relationships.

Animism and shamanism

Traditionally, there were broad similarities among the spiritual cultures of Eastern and Western Arctic peoples. Like the Eskimo culture, Eurasian Arctic culture emphasized the practice of animism—*i.e.*, a belief that everything in nature, every beast and bird, stone and earth, wind and snow, is alive with soul or spirit—and all believed in shamanism. Siberian shamans were frequently women; male and female transvestism in shamans was common.

The shaman's power was often associated with the properties of iron, and there was a corresponding association of the profession of shaman with that of smith. The role of the smith had particular prestige among the Tungus and Yakut.

Christianity reached these peoples at different historical periods; some Lapp were converted as early as the 16th century, whereas some Chukchi probably were never converted, even before the Russian Revolution discouraged religious faith. Christianity, when it was adopted, usually came in the wake of trade with the Western world and with the West's imposition of taxes. In any event, animism and shamanism rarely died completely but became blended with Christian beliefs, rituals, and liturgy. Christian motifs, for instance, appeared in the drawings on Lapp shamans' drums; Christian crosses and crucifixes were worn together with amulets and other talismans of animistic belief. (R.P.B.P.)

Contemporary trends. *Western Arctic.* The establishment of trading centres and, more recently, industrialization and mining have influenced Arctic living. In the period 1900–50 such developments tended to diminish the diffuse residential pattern of the Eskimo, concentrating the Eskimo more and more in towns and other more concentrated population centres. Generally Eskimo and Aleut are moving into the competitive situation of modern life and have demonstrated their ability to adapt well to it.

Eskimo involvement in the money economy

Since early in the 20th century, paternalistic administrations have attempted to devise ways by which the native peoples could develop their Arctic resources to secure much-needed cash. Such attempts go back to the promotion of fur trapping in the 1910s. Some native groups in Canada still depend on furs as a means of trade and a primary source of cash. Clearly, money, in the Western sense, is a present necessity. The purchase of guns, ammunition, whaling bombs, outboard motors, yard goods for clothing, and innumerable other industrial items are now viewed as a necessity. The snowmobile, with the costly gasoline that it requires, has begun to replace the dog team. Adoption of the snowmobile accelerates the change to a moneyed economy and may cause profound changes in Arctic ecology as well. The utilization of the oil resources in Alaska and Canada, along with the natural-gas potential, provides new sources of employment (it has already happened in Alaska) and draws the Eskimo more directly into modern Western culture.

Eastern, or Eurasian, Arctic. The abolition of serfdom in 1861 brought millions of Russian peasants to Siberia in a comparatively short time; Soviet settlement, development, and industrialization during the 20th century has affected the demographic structure of Northern Siberia even more. Although it must be recognized that no non-Soviet observer has been permitted to study any of the northern peoples since the 1930s or even earlier, some historical trends and events are clearly evident. First, it can be noted that attempts to deal administratively with northern peoples began as early as 1922. Then, between 1924 and 1935, a Committee for Assistance to the Peoples of the Northern Regions (the Committee of the North) sought determinedly to bring the native peoples into the Soviet administrative structure and to establish among them "integral cooperatives," collective farms, and educational, medical, technical, and social programs. In 1925 Leningrad State University accepted the first students from among the northern peoples, and in 1930 the Institute of Northern Peoples was established at the university with 195 pupils in attendance from 19 ethnic groups. The institute closed in 1941, but its work was continued after World War II at several regional institutions. Among its principal objectives has been the training of native persons as teachers and administrators.

Collectivization and russification

The editors of the monumental Soviet work *The Peoples of Siberia* (1964; originally published in Russian, 1956) stated that, "By the present time, the village soviet has become the principal local unit of Soviet administration among the peoples of the North, while the collective farm now is universally the fundamental economic unit." The progress toward this state of affairs, however, was beset by difficulties and controversies. The Committee for Assistance to the Peoples of the Northern Regions, for instance, ruled, in 1926, in favour of "clan-based native soviets," but this arrangement proved unworkable because the clan elders were not too willing to support, explain, and enforce Soviet policies. Accordingly, the native soviets were redesigned between 1929 and 1932 as territorial units at the bottom rung of the standard Soviet administrative structure; above the soviet was established the region (*rayon*) and then the district (*okrug*). The northern *okruga* were named after the dominant minority people of the region; by the end of 1932 there were nine such *okruga* of northern peoples, 82 *rayony,* and 462 soviets. Native membership in the executive committees of these *okruga* and *rayony* has varied from 33 to 97 percent.

The progress of collectivization among the northern peoples was recorded as 12 percent of households in 1931, 36 percent in 1934, 75 percent in 1940, and 97 percent in 1948. The program met its chief resistance from reindeer breeders. The problem at heart is a conflict between ideology and ecology or economics. Official Soviet ideology holds that nomadism is an unworthy way of life in a Socialist society, but this is not easily reconciled with the ecological fact that reindeer pastoralism can be conducted successfully only on a nomadic basis, for reindeer must move seasonally. These difficulties have adversely affected not only production in the Soviet reindeer industry but also recruitment to it. Lately, however, ethnographers have tried to persuade administrators and other officials that "production nomadism" is necessary and is compatible with an otherwise stable or sedentary way of life. Probably of greater significance for the future is the issue of whether northern peoples should be encouraged to join the Siberian industrial force or, on the contrary (as Soviet ethnographers seem to urge), be encouraged to remain in their traditional pursuits and to retain their traditional skills.

In education, a number of native grammars were published and used in schools in the early 1930s. The change from a Roman-based to a Cyrillic-based alphabet in 1937, however, presaged greater Russian influence. This seems to have continued, and Russian is now the lingua franca of much of the north, although the Komi and Yakut languages still appear to retain a respected status.

One of the most notable aspects of ethnic relations in

the Soviet north is its complexity. Even at the level of the village soviet, there can be polyethnic membership, and this mixture of representatives from various ethnic groups increases at the levels of the *rayon* and the *okrug;* there are polyethnic collectives such as the Nenets–Enets–Nganasan–Dolgan collective in the Taymyr and the Yukaghir–Lamut–Chukchi one in the northeast; Russians also sometimes belong to northern peoples' soviets and collectives. Some ethnic groups—the Yakut and the Komi—are deemed to have reached the level of "Socialist nation"; others, having national *okruga* and written languages, are perhaps progressing in that direction. But the majority are unlikely ever to achieve nationhood. It seems clear that they will either remain as minorities or disappear. Russification is the dominant trend.

The Lapp of Norway, Sweden, and Finland today form small minorities even in the northern regions of these countries, but they are receiving greater recognition than ever before. The explanation may lie more in the changing attitudes of Scandinavians, who are becoming more tolerant of diversity, than it does in any activism on the part of the Lapp themselves. In any event, the Lapp of Sweden, for instance, have their own ombudsman, whose work is particularly concerned with the legal position of Lapp groups faced by public or private encroachments on their reindeer pastures. There are Lapp language primers, and school readers are now available in each of the three countries (though by no means are all schools with Lappish-speaking children using them); and there are Lappish high schools, Lapp-language newspapers and radio programs, and Lapp-language courses in several Scandinavian universities.

Combined ethnic and national trends among the Lapp

The reindeer-breeding industry is, in general, in good condition. In Norway it is predominantly a Lappish livelihood, and in Sweden it is exclusively so (in Finland, however, the Finns predominate in most reindeer breeding localities). The nomadic way of life associated with reindeer pastoralism is becoming seriously modified, but most of the modifications either are initiated by individual households or camps (and hence vary locally) or emanate from the regional and national associations that have been formed by the reindeer owners themselves. Direct legislation by the state is less important in this process, though the indirect influence of the state—as the principal source of capital loans, subsidized housing, and infrastructure amenities—should not be minimized. Cooperatives for the purchase and marketing of reindeer products are established in many areas.

Lappish individuals and communities in the latter half of the 20th century, without exception, have a heightened sense of national citizenship, and all persons except the very old and perhaps the very young are conversant in the language of the country of their citizenship. Indeed, individuals who leave reindeer breeding are likely to assume a non-Lappish identity: in Sweden such persons usually join the northern industrial labour force or move to Stockholm. In Norway, where the majority of Lapp reside, most families of Lappish descent have never been involved in reindeer herding; most of them have been and are fishermen. Among these people, Lappish culture, in all its respects, is fast being replaced by Norwegian culture.

(R.F.S./R.P.B.P.)

History

ARCTIC PEOPLES

Eurasian prehistory. The polar regions were not populated during much of the Pleistocene Epoch (from 2,500,000 to 10,000 years ago), when most of man's prehistory was unfolding. Man began as a tropical animal, and the Arctic was the last area (except for the Antarctic, only now being explored) to become a habitation area. Areas just south of the Arctic, which, in Pleistocene times, had climates similar to the Arctic, do show significant early developments. At archaeological sites on the Upper Yenisey River and at Lake Baikal and at other locations across the steppes into European Russia, for instance, there have been found the world's oldest known houses—semi-subterranean dwellings with central fire pits, probably ancestral to the earth dwellings of the Eskimos and other Arctic peoples. (For such a functional trait, the possibility of independent invention in different times and places must be considered.) Associated with these finds were flint points and scrapers, fine blade tools, and bone, antler, and ivory tools.

Lateness of Arctic settlement

The Lake Baikal region also supplies evidence for later developments (this time supplemented by some discoveries within the Arctic itself). Around Lake Baikal are to be found a number of stages of development. The earliest of these stages, beginning between 5000 and 4000 BC, is marked by a hunting culture using the bow and arrow. In the second stage, after 4000 BC, pottery and ground-stone tools are present, and in the third, beginning in the 3rd millennium BC, reinforced bows (wooden with bone plates), finely worked implements of bone and ground stone, and effigies of animals (especially fish) are suggestive of artifacts found in later periods all across the Arctic. These materials and additions from later periods in Siberia are reflected in old cultures around the Bering Strait and in the American Arctic.

The Arctic sites contain no evidences of cultivated plants and no evidence of domestic animals other than the dog, so the use of "Neolithic," or New Stone Age, to describe the sites is based on the presence of other Neolithic traits such as ground stone implements and pottery. Even reindeer seem to have been hunted in the wild only, until later times when they were substituted for other domestic animals by groups moving northward, who had kept horses and cattle in warmer regions.

The origin and movements of the Eurasian Arctic peoples are not clear and are subject to scholarly conjecture and controversy. In general, however, rather recent movements from south to north are probable. The Chukchi, Koryaks, and Kamchadals are believed to have entered Northeastern Siberia about 2,000 years ago, finding the ancestors of the Eskimos already living along the coasts. Because the Yakuts and Tungus were still expanding into the northlands after the Russians arrived, it was formerly held that they had migrated there only recently; but the archaeological evidence renders this questionable. Some sources refer to movements of the Yakuts from south to north and of the Tungus from north to south; however, the most tenable hypothesis at present suggests that both the Yakuts and the Tungus developed from groups associated with the Baikal region (or from groups further to the south and east), with the Yakuts expanding north and west over a long period, arriving in the middle Lena by the 17th century and assimilating earlier Arctic residents along the way, while the Tungus expanded both north and south.

Prehistorical migrations in the Eurasian Arctic

Among those assimilated and replaced were many Yukaghirs, who are thought by Soviet prehistorians to have been in the area from about 4000 to 3000 BC and formerly to have extended westward to the Yenisey. The Kets also are held by some scholars to be remnants of a cluster of groups who were early Arctic dwellers, although some Ket traits suggest recent south-to-north movements along the Yenisey. For the Samoyedic speakers (Nenets, Nganasans, Enets, and Selkups), different categories of data provide conflicting conclusions. Linguistic distributions and other language studies suggest that they once lived west of the Urals in the Pechora Basin (now occupied by the Komi and Russians). Their physical type, however, suggests Asian origins. A tenable hypothesis held by some scholars is that a Caucasoid people speaking a Samoyedic language migrated eastward through the Urals about the beginning of the Christian era and intermixed with Mongoloids who had moved westward into Northwestern Siberia. Beginning about AD 1000, the Khants and Mansi, under pressure from the Komi, followed the same migration route, completing their movement across the Urals about 1400. The Komi, themselves under pressure from the Slavs, had moved northward and arrived in their present area about 800 years ago.

More study by linguists, physical anthropologists, ethnologists, and archaeologists has been devoted to the origins of the Lapps than to any other Arctic people (with the possible exception of the Eskimos)—but with no certain results. Some scholars conclude that the Lapps originated

in central Europe, where they were adapted to an Alpine habitat. Others consider them to be an ancient Arctic people who adapted to polar conditions during the Pleistocene. So far, archaeological studies of their shifting camps have shed little light on the question. The most developed hypothesis maintains that the modern Lapps are descendants of a people of Caucasoid physical type who were Stone Age hunters living south and east of their present area and outside the present Arctic but in an area of similar environmental conditions. When the Pleistocene ice retreated, they moved north and west and were the earliest Finnic-speaking people in northern Europe. They adopted reindeer herding according to patterns of animal domestication used for cattle by Europeans further south, augmented with ideas of reindeer herding diffused across the Urals from Asia.

Western prehistory. Two general problems are dominant in archaeological studies in the American Arctic. These are, first, the origin of the Eskimo and, second, the migrations of the ancestral American Indians as they crossed the Bering Strait region from Asia and moved through Arctic America to spread out in other parts of the continent.

Origins of the Eskimo

Two major theories have dominated discussions of Eskimo origins: (1) that the Eskimos evolved from a northern group of American Indians who began to emphasize their winter living patterns and became specialized to tundra dwelling and sea-mammal hunting, or (2) that the Eskimos adopted their way of life in Asia, later migrating to fill a northern void in the New World at a time much later than when most of the ancestors of the American Indians left Asia. In the latter hypothesis, the Asian Eskimos (Yuit) are seen as a remnant along the migration route. In the former hypothesis, the Siberian Eskimos are regarded as a late backwash in counterdirection to the major migrations. Archaeological discoveries, almost from the beginning, have seemed to support the Asian origin, for the earliest sites clearly in the Eskimo tradition have been found around the Bering Strait (Okvik-Old Bering Sea) and in Siberia (Uellen).

There is little evidence of man in the American Arctic in the period between 25,000 and 10,000 years ago, and there are only a few scattered artifacts for the period between about 8000 and 5000 BC, even though the ancestors of the American Indians passed through the Alaskan area in the earliest years of these periods. For the period from 5000 to 3000 BC, however, in the Seward Peninsula and in the Brooks Range of Alaska, there are indications of a land-hunting and, to some degree, sea-mammal-hunting group that may have been related to some of the forest cultures to the south. Then, between 3000 and 2000 BC,

Arctic Small Tool Tradition

the so-called Arctic Small Tool Tradition developed in northwestern Alaska. It was based on the hunting of caribou and other tundra animals, along with some hunting of sea mammals. This culture included elements from the northern movement of the American Plains hunters to northwestern Canada but was primarily derived from northeastern Siberia. It gradually spread eastward in the Canadian tundra to the northwest and northeast side of Hudson Bay, into extreme northeastern Canada, and to western and northern Greenland. This eastern spread was accomplished by 1000 BC.

The Arctic Small Tool Tradition is considered by most prehistorians to be too distinctive to be called Eskimo. The oldest known Eskimo culture is that in the so-called Aleutian and Southern Tradition, dating from as early as 2000 BC (though the earliest archaeological discovery, stone implements found at Anangula Island in the Aleutians, dates back about 8,000 years). Prehistoric south Alaskan and Aleutian culture was basically Eskimo in character but specifically different in many respects from that of western and northern Alaska. It was somewhat simpler technologically; its most important achievements were the production of things made of stone (lamps, chipped knives, and points), bone (harpoon heads, fishhooks, and wedges for splitting driftwood), and ivory (lip ornaments, figurines, and needles). Although the Aleutian and south Alaskan region had the greatest population density of any part of the Eskimo territory, its culture was remarkably stable, in contrast to the Bering Strait region, where culture changes were frequent and pronounced.

In western and northern Alaska is to be found what is probably the next most ancient Eskimo cultural sequence—that of the Old Whaling, Choris, Norton, and other cultures, each known from only one or a few sites north of Kotzebue Sound near the Bering Strait, the earliest dating from about 1800 BC. Included in this sequence is the Ipiutak culture, which, discovered in 1939, is a major enigma in Eskimo prehistory. Ipiutak, known from the type site at Point Hope and a few other western locations, contains many traits of the Eskimo tradition, including toggle-headed harpoons and other items for sea hunting (although harpoon heads are outnumbered by arrowheads for hunting caribou); but the culture lacks many typical Eskimo items and stands by itself in the large size of the village and the regular arrangement of houses along streets. The Ipiutak site contains more than 600 shallow semi-subterranean dwellings with central fireplaces. The slight accumulations of debris argue for a short period of occupation (estimated to be about 100 years). Dating also is somewhat uncertain. Some radiocarbon dates for Ipiutak are as late as AD 1000, whereas some artifacts suggest overlap with the earlier phases of the Northern Maritime Tradition. Most summaries conclude that Ipiutak dates from about AD 300 to 400 and drew upon both the Northern Maritime Tradition (for seal-hunting techniques) and the northeastern Asian cultures (for certain implements and ceremonies).

Dorset culture

In the eastern parts of the American Arctic—including the Hudson Bay region, Labrador, Newfoundland, and Greenland—there is another very early culture, known as Dorset, dating from about 1000 BC and lasting to about AD 1000 or 1200 (when Eskimos in the Northern Maritime Tradition overran the area). The relationship between Dorset and other Eskimo traditions is somewhat enigmatic. Dorset sites contain semi-subterranean houses and some Eskimo traits such as the toggle-headed harpoons, lamps, and ivory carvings; but many of the artifacts were somewhat different in form, smaller in size, and more delicate; moreover, these sites lack dogsleds, bone arrowheads, lip ornaments, and other typical Eskimo traits. Many prehistorians think that the Dorset culture developed from the Small Tool Tradition and marks the migration of people who began moving from the western end of the American Arctic about 3000 BC, bringing with them implements in the Arctic Small Tool Tradition, and who by 1000 BC had reached the eastern extent of the Eskimo area and developed a largely independent culture. Some materials in later Dorset sites, however, resemble similar items in the Northern Maritime Tradition. Most summaries of Eskimo prehistory conclude that the Dorset Eskimos were overrun by the eastward migration of the Thule Eskimos, who were carriers of the Northern Maritime Tradition. Some scholars believe that the Eskimos of Southampton Island, extinct in 1904, and somewhat different from neighbouring Eskimos, represent a late survival of the Dorset peoples.

Northern Maritime Tradition

The best known early Eskimo cultural sequence is that of the Northern Maritime Tradition, which began in the area from the Bering Sea to Point Barrow about 500 BC. In this tradition, the most typical traits of Eskimo culture, including the semi-subterranean house and the toggle-headed harpoon (with all other parts of the seal-hunting complex), are present from the start. Although some less typical traits were added as the tradition developed and the emphasis in some areas changed (as from chipped-stone to ground-stone implements), the phases in the Northern Maritime Tradition are mainly demarcated only by stylistic changes, as in the design of harpoon parts, arrowheads, needle cases, and art objects.

In art the general trend is from more elaborate to less elaborate. Tools and other implements also become plainer and more utilitarian in later periods, although more efficient. The Birnirk and Thule cultures are the most widely distributed examples of the Northern Maritime Tradition, with Birnirk sites reported from the mouth of the Kolyma River in Siberia to the Mackenzie Delta in Canada, and Thule even more widely distributed, reach-

ing from the Bering Strait to Greenland. The spread of Thule, after it developed out of Birnirk some time before AD 1000, has been traced eastward from Alaska, arriving in Greenland about 1200, where it was influenced by the medieval Norse settlements in the subsequent century. Later there was a resurge of Thule back toward the west, reaching all the way to Bering Strait and accounting for the relative uniformity in language and other aspects of Eskimo culture in early contact times.

At many locations throughout the area of the Northern Maritime Tradition, archaeological materials have been studied that trace the development of Thule into historically identified groups of Eskimos. Although there are many specific problems yet to be resolved, in general the Northern Maritime Tradition has the feeling of history.

Modern populations. Although many nations participated in the exploration of the Arctic, only a few have contributed settlers, and all are Indo-European speakers. Russians predominate in Eurasia (and numerically throughout the Arctic), although they were preceded in Arctic Europe by Scandinavians. Currently, Norway, Sweden, and Finland all have settlements north of the Arctic Circle. In Arctic Alaska and Canada, English-speaking peoples are found in small numbers; and in Greenland, Danish settlers have intermixed with the Eskimos, though Eskimo speech has been retained.

Russian settlement of Siberia

Russians began crossing the Urals in the 16th century and continued in flood tide after the abolition of serfdom in the Russian Empire in 1861. Settlement of the northlands was further accelerated by the building of the Trans-Siberian Railroad in the late 19th and early 20th centuries and was continued as state policy by the Soviet government after the Revolution. In the late 20th century there were more than 5,000,000 people in the parts of the Soviet Union north of latitude 60° N. A condition of note for the modern populations of the Arctic Soviet Union is that around three-quarters live in urban concentrations, whereas only around half of the people of the Soviet Union as a whole are urban. Concentrated living in the north shows that the modern populations are mainly involved in extractive industries, such as mining, petroleum, and lumbering (in the subarctic). The Russian settlers, who outnumber the indigenous peoples, are the inhabitants of the Arctic cities; but some of the native peoples are also becoming urban dwellers, particularly the Komi.

American Arctic settlements

In the American Arctic there are no comparable populations. Canada has more than 1,000,000 people in the Arctic and subarctic, almost all of them in the subarctic. In Alaska the population is mostly concentrated in the subarctic, although one major city, Fairbanks, is close to the Arctic Circle. (In this comparison, however, it should also be noted that the "Soviet North" includes much of the subarctic, and most of the Russian settlers live outside the Arctic proper.) A process of population concentration has, in fact, occurred in the American Arctic, even though it has not resulted in cities. In Alaska and Canada, the Eskimos of the north have abandoned many small settlements and concentrated in fewer towns—a response to opportunities for employment and other appeals of living in larger communities. Only in Greenland, with its rapidly growing population, has there been an increase in the number of communities. (W.K.C.)

EXPLORATION

The earliest references to Arctic exploration are shrouded in superstitious beliefs concerning the uninhabitable "frigid zone," and also in an obscurity resulting both from inaccurate ideas of the shape of the Earth and from primitive navigation techniques, which make it very difficult to interpret early maps and accounts of voyages. Probably the first to approach the Arctic regions was a Greek, Pytheas, who in the 4th century BC made an astonishing voyage from the Mediterranean, around Britain, to a place he called Thule, variously identified as the Shetlands, Iceland, and Norway. The accounts of this remarkable explorer were for centuries discredited, but the idea of his Thule, shrouded in fog and believed to be the end of the Earth, caught the imagination of many.

Iceland is known to have been visited by Irish monks in the 8th and 9th centuries, but it was the Vikings from Norway who settled the island, late in the 9th century. In the course of the next four centuries, these hardy seamen established trade routes to the White Sea; visited Greenland (*c.* 982) and founded two settlements on the southwest coast (which disappeared, for unknown reasons, before the 16th century); reached the coast of North America; and probably also reached Spitsbergen and Novaya Zemlya. Unfortunately, however, they left scant records of their voyages, and many of the places they visited had to be rediscovered by others.

Viking settlement

The Northeast Passage. After a long period of inactivity following the decline of the Norsemen, leadership in Arctic exploration was assumed in the early 16th century by the Dutch and the English. The motive was trade with the Far East. The known sea routes around the southern tips of Africa and South America had been claimed as a monopoly by Portugal and Spain respectively and besides were long and arduous; the overland routes were even worse. There remained, however, the northern latitudes, and the attempts of English and Dutch merchants to find a Northeast and Northwest Passage gave a strong stimulus to exploration of the Arctic.

In 1553 the English sent three ships to the northeast under the command of Sir Hugh Willoughby, with Richard Chancellor as chief pilot. Willoughby, with two ships, wintered in a harbour on the Kola Peninsula, where he and all his men perished. Chancellor, who in the "Edward Bonaventure" had become separated from the others in a gale, reached what is now Archangel and made an overland journey to Moscow (1,500 miles in all) before returning home to England. It is interesting to note that these waters were already well known to Russian seamen, who used the route around North Cape to western Europe as early as 1496, but this was not generally known at the time.

After Chancellor's voyage the Muscovy Company was formed, and there grew up a very lucrative trade with Russia, the success of which rather distracted the minds of the English from the Northeast Passage. Nevertheless, in 1556 Stephen Borough sailed in the "Searchthrift" to try to discover the Ob River. He was stopped by ice and fog at the entrance to the Kara Sea, which he described in such discouraging terms that it was not until 1580 that another English expedition, under Arthur Pet and Charles Jackman, attempted its passage. They too failed to penetrate it, and England lost interest in searching for the Northeast Passage.

In the meantime, however, the Dutch had taken up the search, largely because of the efforts of Olivier Brunel, who in 1565 established a trading post at Archangel, to the consternation of the English traders. In the course of an eventful career, Brunel made an overland journey to the Ob and in 1584 tried to reach it by sea, but like Pet and Jackman he got no farther than Yugorski Strait. He was followed by Willem Barents, an outstanding seaman and navigator, who in 1594 discovered Novaya Zemlya and sailed to its northern tip; his two companions, Cornelis Nai and Brant Tetgales, penetrated a little way through Yugorski Strait into the Kara Sea. In 1596, with Jan Cornelisz Rijp and Jacob van Heemskerck, he was more successful. Heading due north from Norway instead of following the coast around, Barents discovered Bear Island (Bjørnøya) and Spitsbergen, which he mistook for Greenland. Rijp then went home with one ship, but Barents and Heemskerck in the other headed east and rounded the north of Novaya Zemlya. They were forced to winter in Ice Haven on the northeast coast and thus became the first Europeans to winter successfully in the Arctic. They built a house of driftwood and passed the season with remarkable fortitude and success, although they were in almost every way unprepared for it. In the spring, the ship being hopelessly damaged, they escaped across the open Barents Sea in two small boats. Barents died on the journey. In 1609 Henry Hudson, the Englishman, sailed in the "Half Moon" to the Barents Sea in the service of the Dutch East India Company, but his crew, afraid of having to winter like Barents, mutinied and forced him to sail west, where he explored the coast of North America north of Virginia and ascended the Hudson River.

Dutch voyages

While the English and the Dutch were thus engaged, Siberia was gradually being annexed to Russia by bands of Cossacks spreading slowly eastward from the Urals. Conquering and trading as they moved, the Cossacks reached the Pacific Ocean by the middle of the 17th century and traversed the northern coastline in small, flat-bottomed boats. One of these hardy wanderers, Semen Dezhnev, is believed to have sailed through the Bering Strait in 1648. Like the Norsemen, however, the Cossack settlers made few records of their activities, and only rumours of the existence of a strait separating Asia from America reached the ears of Peter I the Great.

In order to find and above all to map the sea route to the east, Peter conceived a series of expeditions that culminated in the Great Northern Expedition (1733–42) and, under his successors, mapped the Arctic coastline of Siberia as far east as the Kolyma River. The work of the expeditions was planned by Vitus Bering, a Dane with long service in the Russian Navy, and directed by the admiralty college. In 1728 Bering, who himself took charge of activities in the Pacific, sailed through Bering Strait from the south, but because of bad weather failed to recognize its significance. In 1741 he crossed the North Pacific and landed on the coast of Alaska. Other important figures among the many who made up the Great Northern Expedition are Lieut. Semyon I. Chelyuskin, who in 1741 reached the cape named after him, the most northern point of the Siberian mainland; and the brothers Laptev, Dmitry and Khariton, who charted the coast from the Taymyr Peninsula to the Kolyma River.

The question of the existence of a strait separating the continents remained uncertain until 1778 when Capt. James Cook, the great British explorer, sailed through Bering Strait and reached North Cape (now Cape Schmidt) on the Siberian coast. Unlike Bering, he was able to see both sides of the strait and so establish fairly conclusively that the continents were separate.

Attempts to complete the mapping of the northern coast were made by a Russian merchant named Nikita P. Shalaurov in 1760–62 and by Joseph Billings, an Englishman in the service of the Russians, in 1787; it was eventually finished by Baron Ferdinand von Wrangel, a Russian of Swedish parentage, in a series of sledge journeys from 1820 to 1824.

Wrangel's companion, P.F. Anjou, worked in the New Siberian Islands. These strange islands of silt and ice were discovered in 1770 by a merchant named Ivan Lyakhov, who saw caribou coming south to the mainland and followed their tracks north. The islands produced a fortune in mammoth ivory and were further explored and mapped in the years that followed, notably by an official survey party under an exile, Matvey M. Hedenström.

The main features of the Northeast Passage were now investigated and mapped; it remained to be successfully navigated. One attempt to do so, by an Austrian expedition under Julius Payer and Karl Weyprecht in 1872–74,

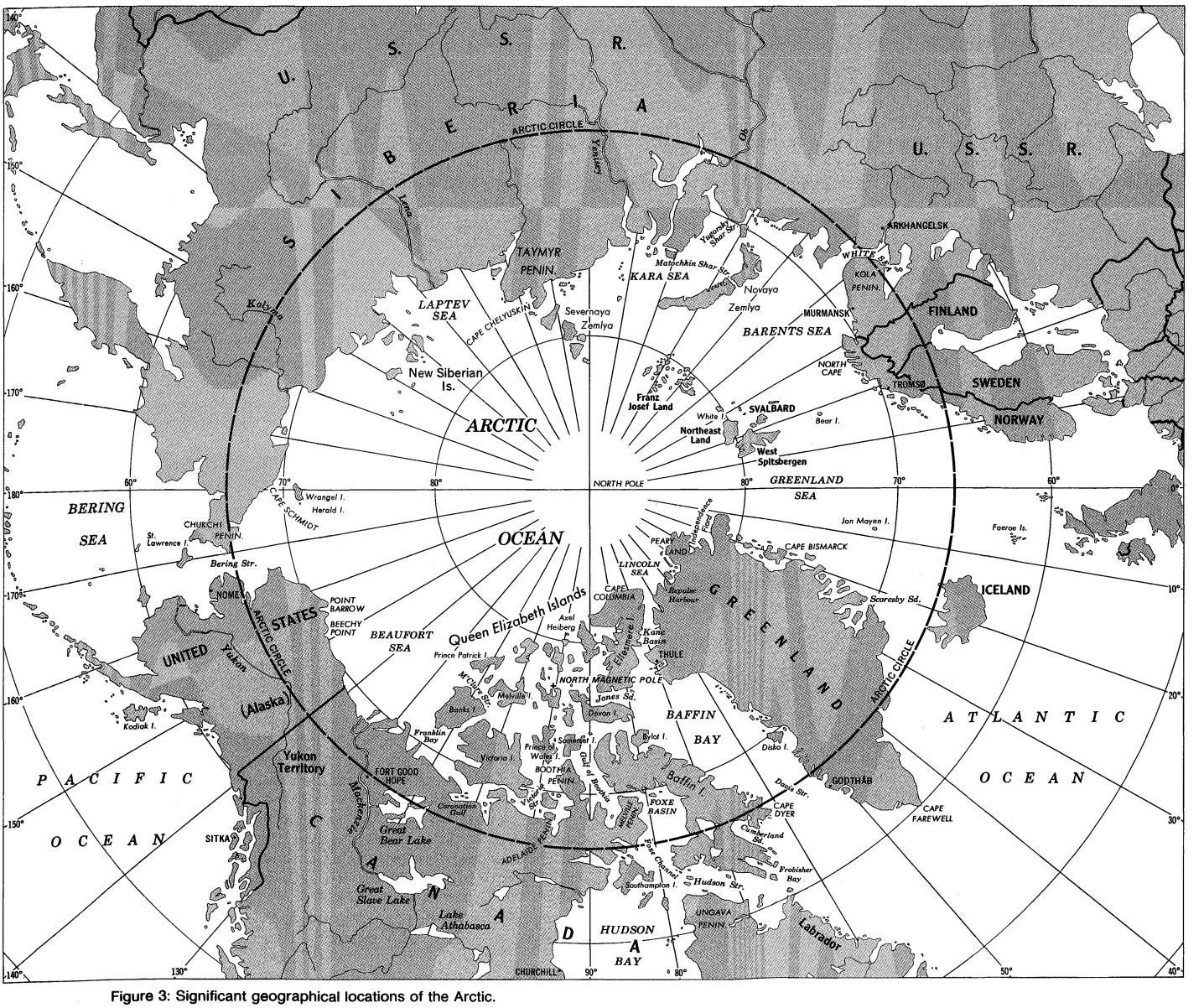


Figure 3: Significant geographical locations of the Arctic.

led to the accidental discovery of Franz Josef Land. While trying to sail around the north end of Novaya Zemlya, their ship, the "Tegetthoff," became caught in the ice and drifted helplessly north and west for more than a year, to be deposited by the ice on the shore of a new land. In the spring of 1874 Payer made a long sledge journey, reaching the northernmost point of the archipelago. The ship had to be abandoned, and the party sailed in small boats to Novaya Zemlya, where they were picked up by a Russian fishing vessel.

A.E.Nordenskiöld

The first successful navigation of the passage was accomplished in 1878–79 by one of the greatest of Arctic explorers, the Swedish baron A.E. Nordenskiöld, who saw in it not so much a route to the Far East as an avenue of trade between Siberia and the rest of the world. After two reconnaissance trips to the Yenisey in 1875 and 1876, Nordenskiöld set out in 1878 in the "Vega" accompanied by three cargo ships, two bound for the Yenisey and one for the Lena. Aided by a good ship and a year of rather light ice conditions, he managed to reach the Chukchi Peninsula before being forced to winter only 120 miles from the Bering Sea. The passage was next navigated from east to west by Comdr. B.A. Vilkitski of the Russian Navy, who was engaged in hydrographic work. In 1913, while trying to pass Cape Chelyuskin, he was forced north and discovered Severnaya Zemlya, first called Nicholas II Land. He had to turn back and winter on the Pacific coast but succeeded in completing the passage in 1914–15. Two west–east attempts in 1912 by G.L. Brusilov and V.A. Rusanov ended disastrously with the almost complete loss of both expeditions. In 1932 the icebreaker "Sibiryakov" was the first to make the passage in one season, sailing from west to east. During the late 20th century the passage has been kept open by Soviet icebreakers, such as the nuclear-powered "Sibir," which opened a southern route across the northern Siberian coast in 1978.

The Northwest Passage. The search for the Northwest Passage may be said to have begun with the European discovery of America, for the voyages of Cartier and his successors to the St. Lawrence and the Cabots and Côrte-Reals to Newfoundland and Labrador were all undertaken with the aim of finding the passage. The first such voyage to enter the Arctic, however, was that of Sir Martin Frobisher in 1576. At this time the map was a blank between southern Labrador and Greenland. To complicate matters further, there existed a spurious map, published in 1558 from the voyage of Niccolò Zeno in 1380, that showed an entirely imaginary country called Friesland lying between Iceland and Greenland. Because of the supposed existence of this land, confusion in mapping and identifying newly discovered lands persisted for 200 years.

Frobisher set out with the "Gabriel" and "Michael" and a tiny pinnace of 10 tons. He made his North American landfall on the southeast coast of Baffin Island, after weathering terrible storms in which the pinnace was lost with all four of its crew and the "Michael" deserted and went home. In the "Gabriel" Frobisher sailed about 60 miles up the long inlet named after him, which he took to be a strait, and brought home among his trophies a rock sample that was identified wrongly as containing gold. The Northwest Passage was forgotten, and in the next two years Frobisher made two further voyages for the sole purpose of establishing a gold mine. The last voyage was an astonishing enterprise involving 15 ships and had the aim of founding a settlement of 100 men. Included in the cargo was a prefabricated house. The ships, however, were scattered by storms, at least one was sunk, and Frobisher, unable to set up his colony, loaded the remaining ships with ore and returned home only to find that his cargo was worthless.

Next to seek the passage was John Davis, one of the finest of the early seamen and something of a scientist as well. In three voyages, 1585–87, Davis rediscovered Greenland (lost to Europeans since the decline of the Norse settlements); he visited the southeast coast and sailed up the west coast to beyond Disko Island (72° N). He also traced the coasts of Baffin Island and Labrador from Cape Dyer south. Davis showed great imagination in his dealing with the Greenland Eskimos. He took musicians with him and had his sailors dance to the music, thus immediately establishing cordial relations with these sociable people.

In 1602 George Weymouth sailed a short way into Hudson Strait, and in 1610 Henry Hudson made his last voyage, sailing the "Discovery" into Hudson Bay and down to James Bay, where he was forced to winter. In the spring there was a mutiny aboard; Hudson and about eight of his crew, including his young son, were set adrift in a small boat to die, while the mutineers sailed the ship home. Retribution, however, overtook the ringleaders in Hudson Strait, where they were killed by Eskimos in one of the rather rare instances when these peaceful and good-natured people have attacked explorers. The remnant that reached England in September was imprisoned; nothing was ever heard from the deserted men.

Discovery of Hudson Bay

The early exploration of Hudson Bay led to a long series of voyages seeking a western outlet to this large inland sea, and at the end of 20 years its shores were fairly well known. Sir Thomas Button in 1612–13 (with Robert Bylot, a Hudson survivor, as pilot) was the first to reach the west coast of the bay, wintering near the site of York Factory and discovering Roes Welcome Sound; William Baffin, again with Bylot, sailed halfway up the northeast coast of Southampton Island in 1615; Jens Munk, a Dane, wintered at the mouth of the Churchill River in 1619–20, where nearly all his men died of scurvy, only Munk and two others surviving to sail one of the two ships home; and in 1631 Luke Foxe sailed into Foxe Channel.

In the meantime Baffin, the outstanding navigator of his day, had explored Baffin Bay (1616), but the significance of this exploration was not recognized for 200 years. With Bylot as master of his ship (Hudson's old "Discovery"), Baffin sailed up the west coast of Greenland to the head of Baffin Bay (78° N) and down the west side of the bay, discovering the three sounds that lead out of it, Smith, Jones, and Lancaster. Unfortunately, he reported that all three were merely bays and that there was no passage out of Baffin Bay. Even more unfortunately, his map was never published, and in time the very existence of "Baffin's Bay" came to be doubted. Search was continued in the Hudson Bay area up to the end of the 18th century, but at a greatly reduced pace. By that time traders were opening up the interior and hope of a passage by this route was fading.

The 19th century brought an entirely new era. Up until that time the search had been undertaken mainly by merchants; now it was the turn of governments. The end of the Napoleonic Wars had left the British Navy relatively unemployed, and the British government, spurred by the enthusiasm of Sir John Barrow, secretary to the admiralty, was persuaded to equip a whole series of large naval expeditions for the discovery of the Northwest Passage. The first of them, under Capt. John Ross in 1818, retraced almost exactly Baffin's journey of two centuries earlier, and repeated his error of mistaking the sounds for bays. Second in command to Ross was Lieut. W.E. (later Sir William Edward) Parry. He was not convinced that no sound existed, and in 1819–20, in HMS "Hecla" and "Griper," he made a voyage through Lancaster Sound to Melville Island in the western part of the archipelago, where he wintered. Blocked by ice in M'Clure Strait, he next (1821–23) tried the route through Foxe Channel, spending two winters in Foxe Basin. Again he was stopped by ice in the narrow Fury and Hecla Strait (named after the two ships he used on this expedition). A number of rather unsuccessful ventures followed. Parry on a third voyage (1824–25) explored Prince Regent Inlet; Capt. G.F. Lyon and Capt. George Back made unsuccessful attempts to reach Repulse Bay; and Capt. John Ross, on a privately financed venture in 1829–33, sailed down Prince Regent Inlet into the Gulf of Boothia, passing by one of the keys to the Northwest Passage, the narrow Bellot Strait, which washes the northernmost tip of the North American continent. The latter expedition added greatly to the extent of mapped territory, mostly through the work of Ross's nephew, J.C. Ross, who established the position of the North Magnetic Pole in southwest Boothia Peninsula. After three winters trapped in the ice, Ross had to abandon his ship, the "Victory," and retreat by sledge and boat,

spending a fourth winter on the way before being picked up by a whaler in Lancaster Sound.

In the meantime the British were also attacking the problem from the west by both sea and land. In 1819–22 and 1825–27 two expeditions under Capt. John Franklin, working overland and by boat from wintering bases in the Mackenzie Basin, surveyed the coastline from Turnagain Point, about 200 miles east of the Coppermine River, to Cape Beechey, Alaska. There Franklin almost made contact with the survey of Lieut. F.W. Beechey, who in 1825–26 reached Point Barrow from the west. In 1833–35 Capt. George Back discovered the Back River and mapped it to its mouth in Chantrey Inlet, and in 1837–39 P.W. Dease and Thomas Simpson, Hudson's Bay Company employees, made three coastal journeys by boat, filling in the gap in the Alaska coastline left by Franklin and joining Franklin's survey to Back's at Chantrey Inlet. In 1847 another Hudson's Bay Company employee, John Rae, joined Parry's Fury and Hecla Strait to Ross's survey in the Gulf of Boothia. Rae was a most remarkable traveller, far ahead of his time in adopting Eskimo methods and living off the land.

Capt. John Franklin

Most of the continental coastline and a considerable amount of the Canadian Arctic Archipelago had now been charted, and still the Northwest Passage remained elusive. The British government was getting tired of the project but prepared to send out one last expedition. This was the famous and tragic last voyage of Sir John Franklin, who sailed into Lancaster Sound in 1845 in HMS "Erebus" and "Terror" and was never seen again. The loss of this expedition produced a reaction of profound shock and resulted in a 12-year search that contributed tremendously to geographical knowledge. At its peak in 1850, as many as 14 ships were in the area at the same time, and a further expedition was at work from the mainland. The story eventually pieced together was that Franklin had wintered at Beechey Island at the west end of Lancaster Sound, after sailing up Wellington Channel to 77° N, and in the spring of 1846 turned south down Peel Sound, hitherto unnavigated, to Victoria Strait, off the north tip of King William Island, where his ships eventually had to be abandoned. There were no survivors.

The first to become anxious was Sir John Richardson, a veteran of Franklin's earlier expeditions, who in 1847–49 conducted a search along the northwest mainland coast, accompanied by Rae. The first official search parties were sent out in 1848; Sir James C. Ross, with "Enterprise" and "Investigator," was to enter from the east, and Capt. Henry Kellett, with the "Herald" and "Plover," had orders to stand by in Bering Strait to meet Franklin on his way out. Ross wintered in Somerset Island and traced most of its coastline before returning in 1849 without news. In an atmosphere of mounting alarm, a more thorough search was made in 1850. Capt. Horatio Austin wintered with four ships, the "Resolute," "Assistance," "Intrepid," and "Pioneer," off the south coast of Cornwallis Island, from which base extensive sledge trips traced many miles of coastline. Two more ships, under Capt. William Penny, a whaler, were in the same area, as was also Sir John Ross, then 73 but still active. The first U.S. expedition to the Arctic, financed by Henry Grinnell and led by Lieut. E.J. de Haven, sailed in the "Advance" and "Rescue" to Wellington Channel. Franklin's winter quarters at Beechey Island were found by Austin's and Penny's expeditions, but no record had been left to point the way from there.

Search for Franklin party

At the same time, in 1850, Capt. Richard Collinson was to enter from the west and meet Austin in a pincer movement. His two ships became separated in the Pacific, however, and operated independently. Comdr. Robert McClure in the "Investigator" discovered Prince of Wales Strait, rounded Banks Island by the west, and entered Mercy Bay on the north coast, where the ship remained frozen in for two years and was finally abandoned. McClure and his men were rescued by another expedition and returned home in 1854 by the eastern route. He was thus the first to make the Northwest Passage, though in more than one ship and partly on foot. Collinson in the "Enterprise" spent three years in Victoria Island, reaching Victoria Strait. There he was within a short distance of the place where Franklin's ships had been abandoned, as was also Rae, travelling by boat two years earlier. Neither found any clues. In 1852 a private expedition financed by Lady Franklin and led by a whaling captain, William Kennedy, discoverd Bellot Strait, named after a French volunteer in the search.

Voyage through Northwest Passage

After this the search moved north, which was generally thought to be the most likely direction; only Lady Franklin held firmly, and correctly, to the view that her husband had gone south as directed by his orders. In 1852 Capt E. Inglefield in the "Isabel" sailed north up Smith Sound to 78°35′ N, and another large expedition, under Sir Edward Belcher and Henry Kellett, sailed into Lancaster Sound with Austin's four ships plus a supply vessel, the "North Star." Splitting into an eastern and a western party and spending two winters in the Arctic, this expedition mapped many miles of new coastline north of Lancaster Sound, rescued the survivors of McClure's expedition, and then without apparent justification abandoned all four ships in the ice and sailed home in the "North Star." One ship, the "Resolute," was later found drifting in good condition in Davis Strait by a whaler, sailed to New England, refitted, and returned to the British government by the United States.

In 1853 an American, E.K. Kane, sailed in the "Advance" to Kane Basin, wintering twice and searching northward to Kennedy Channel. In the same year Rae was sent by the Hudson's Bay Company to complete the charting of the mainland coast between Chantrey Inlet and Boothia. The company had given up the search as hopeless, but ironically it was this expedition that brought back the first real news, obtained by Rae from Eskimos in Pelly Bay and backed by identifiable relics. The British government considered the search closed, but Lady Franklin was not satisfied; she financed a final expedition in the "Fox" under Capt. F.L. McClintock, who had been on three previous expeditions and was largely responsible for developing the sledging techniques used during the search. He travelled around the coasts of King William Island, finding many bodies and relics of the expedition, and also the only record left by it, at Victory Point.

So the Northwest Passage was at last found to be a reality, and official recognition went to McClure as its discover, though Franklin too had proved its existence and should share the honour. An unsuccessful attempt to navigate it was made by Allen Young in the "Pandora" in 1875. In 1903 Roald Amundsen, the great Norwegian explorer, sailed down Peel Sound in his tiny yacht "Gjöa" and passed around the east side of King William Island, where he spent two winters taking magnetic and other scientific observations. After a third winter spent west of the Mackenzie, he passed through Bering Strait in 1906, the first to navigate the Northwest Passage. It was navigated again in 1940–42 and 1944 by Henry A. Larsen of the Royal Canadian Mounted Police in the "St. Roch," west–east by way of Bellot Strait and east–west in one season by Prince of Wales Strait. In 1954 the first passage by a deep-draught vessel was made by HMCS "Labrador," a Canadian naval icebreaker. In 1969 the "Manhattan," the largest and most powerful commercial ship ever built in the United States to that time, smashed through 650 miles of ice between Baffin Bay and Point Barrow, Alaska, to assess the commercial feasibility of the passage. The future of the Northwest Passage as a regular commercial route remained uncertain, however, in the late 20th century.

Navigation of Northwest Passage

Whale fisheries and the fur trade. Many advances in geographical knowledge were due, directly or indirectly, to the whale fisheries that flourished in the Arctic for three centuries. On his first voyage in 1607, in the service of the Muscovy Company, Henry Hudson sailed up the east coast of Greenland in search of a direct polar route to Cathay. Reaching 73°30′ N, he turned east along the edge of the pack ice to Spitsbergen and on his way home discovered Jan Mayen Island. He was impressed by the number of whales he saw, and, as a result of his recommendation, a thriving whaling industry prospered under English, Dutch, Danish, French, and other companies, which continued until the whales were gone. In the 18th century the whaling spread to East Greenland waters, and

in the 19th century to Baffin Bay, Hudson Bay, and, after 1850, the Beaufort Sea, where U.S. whalers played a prominent part. By 1920 the industry was dead.

Much of the geographical knowledge accumulated by the whalers was never recorded and died with them; some, especially in the early days, was deliberately suppressed so as to keep it from competitors, but a great deal did find its way onto the maps. The coasts of Spitsbergen were first mapped by Dutch and English whalers, and the Dutchman Cornelis Giles discovered an island east of Spitsbergen that was long known as Giles (Gillis) Land, now White Island. Later details were added by Norwegian sealers. The considerable part played by whaling captains in the Franklin search has already been noted; in addition, the names of many whalers are perpetuated on the maps of Baffin Island and Hudson Bay. A whaler, William Adams, established the insularity of Bylot Island, and another, George Comer, made the first complete map of Southampton Island. Going farther west, Wrangel Island was discovered by Thomas Long, a U.S. whaler.

By far the most famous of the whalers were the William Scoresbys, father and son. Scoresby Sr., a farmer's son, was a first-rate navigator, inventor of the crow's nest and other aids to ice navigation, and the first to suggest the use of sledges to reach the pole. His son, who inherited his father's talents and added to them a scientific education, wrote two important books on the Arctic. In 1806 the Scoresbys reached 81°12′ N, north of Spitsbergen, a record northing at the time, and in 1822 the younger Scoresby made a detailed map of the east coast of Greenland from 75° to 69° N.

Just as whaling led to improved knowledge of the coastlines, the fur trade helped to open the interiors of Arctic lands. The Cossacks who settled Siberia were mainly engaged in this profitable trade, and the interior of Spitsbergen was first frequented by Russian hunters, who in the 18th century hunted the furbearers almost to extinction. In North America the influence of the fur traders was also great.

The formation of the Hudson's Bay Company was a direct result of the 17th-century voyages into Hudson Bay in search of the Northwest Passage. Following an exploratory voyage by Capt. Z. Gillam in 1668, the company was incorporated in 1670 and a trading post established at the foot of James Bay. Soon posts were set up on the west side of the bay at York Factory and Churchill, and these served as bases for further exploration. It was a hundred years, however, before the Hudson's Bay Company made any real attempt to explore the hinterland; Samuel Hearne was sent to look for a source of copper reported by Indians who traded at the coast. Hearne set out from Churchill with a band of Indians in 1770–71 and with them made a remarkable journey to the mouth of the Coppermine River, returning by way of Great Slave Lake. In 1789 Alexander Mackenzie of the rival North-West Company of Montreal travelled by canoe from Lake Athabasca down the Mackenzie River to the sea, establishing a second known point on the Arctic coast. By the time the two companies merged in 1821, there were trading posts on Great Slave Lake and down the Mackenzie to Ft. Good Hope; it was the existence of these posts that made possible the overland expeditions of Franklin and his successors, among whom were many Hudson's Bay Company men.

The Hudson's Bay Company

Pushing westward from the Mackenzie through the mountains and into Alaska, the Hudson's Bay Company met Russian traders working from the west coast. The Russians had established a colony in Alaska toward the end of the 18th century and carried on a vigorous trade at Kodiak, Sitka, and other settlements. In 1831 Baron von Wrangel, governor of the colony from 1829 to 1834, established a post on St. Michael's Island and had the lower Yukon explored. Competition and strife between the Russian and British traders ended with the purchase of Alaska by the United States in 1867 and the joint survey of the Alaska–Yukon boundary.

The North Pole. The North Pole did not become in itself a goal of exploration until fairly late; the few early expeditions that tried to reach the pole were looking for a polar route to the East rather than for the pole itself. After Hudson's first attempt in 1607, nearly 200 years elapsed before the British government sent another expedition, under Capt. C.J. Phipps in 1773, to try to sail across the North Pole. He got only as far as the north of Spitsbergen, and Capt. David Buchan in 1818 did no better. In 1827 Parry was the first to try the sledging technique suggested by Scoresby, breaking the latter's record by reaching 82°45′ N.

All these attempts had been in the area between Greenland and Spitsbergen, which actually is not the accessible route to the Arctic Ocean that it appeared to be, owing to the strong southerly drift of the ice. The Franklin search opened a new route, up the west coast of Greenland. In 1860 I.I. Hayes attempted to reach the pole by this route in the schooner "United States." Hayes was a firm believer in a theory, then prevalent, that the polar sea was ice-free and that it could be reached by breaking through the fringing belt of pack ice. Ironically, he met with unusually heavy ice conditions and got only as far as Etah on the coast of Smith Sound. In 1871 Charles Francis Hall, another American, with more luck and a better ship, reached 82°11′ N and charted both sides of the channel to its northern end at the entrance to the Lincoln Sea. Hall himself died during the winter and his ship, the "Polaris," was caught in the ice on the voyage south and drifted to Smith Sound, where it was almost wrecked. A party of 19, including an Eskimo mother with a two-month-old baby, became separated from the ship and drifted all winter on an ice floe, being picked up by a whaler in April 1873 off the coast of Labrador. In 1875–76 a British expedition under Capt. G.S. Nares in the "Alert" and "Discovery" reached the Lincoln Sea by ship, the "Alert" wintering near Cape Sheridan on the north coast of Ellesmere Island, the "Discovery" farther south at Lady Franklin Bay. Sledge parties in the spring traced the coasts of Ellesmere Island and Greenland to Yelverton Bay and Sherard Osborn Fjord respectively, and one, under Comdr. A.H. Markham, reached 83°20′ N over the pack ice, a new record northing.

In the meantime the Spitsbergen route was not neglected. In 1869–70 a German expedition under Karl Koldewey in the "Germania" sailed up the east coast of Greenland to 72°30′ N and traced it by sledge to Cape Bismarck. A second ship, the "Hansa," became separated and was crushed in the ice, and the crew drifted south on a floe around Cape Farewell, reaching the settlement of Frederiksdal in safety. A.E. Nordenskiöld made two journeys toward the pole from Spitsbergen, in 1868 by ship and in 1873 by reindeer sledge.

An entirely new approach was tried in 1879 by a U.S. expedition in the "Jeannette," led by Lieut. Comdr. G.W. De Long. At that time it was believed by many that Wrangel Island was a large land mass stretching far to the north, and De Long hoped to sail north as far as possible along its coast and then sledge to the pole. Accordingly, he sailed through Bering Strait, but his ship was caught in the ice near Herald Island and drifted west for 22 months, passing north of Wrangel Island and revealing its limited extent. The "Jeannette" sank near the New Siberian Islands and the crew reached the Lena Delta on foot, where many of them died, including De Long himself. A search expedition under Lieut. R.M. Berry surveyed Wrangel Island.

Wreckage from the "Jeannette" was found a year or two later on the southwest coast of Greenland, having apparently drifted right across the Arctic Ocean. Fridtjof Nansen conceived the daring idea that a ship might be made to do the same, thus providing a base for scientific investigation of the Arctic Ocean and incidentally a means of reaching the pole. In a new vessel, the "Fram," specially designed to rise under lateral pressure and so avoid being crushed, Nansen left Norway in 1893 with Otto Sverdrup and sailed into the Kara Sea. Near the place where the "Jeannette" sank they drove the "Fram" into the pack and began a drift that lasted almost three years and ended with the safe release of the vessel north of Spitsbergen in 1896; a formidable amount of scientific data was collected. Nansen himself left the "Fram" in 1895 with one companion, Hjalmar Johansen, in an attempt to reach the pole by

First voyage of the "Fram"

sledge, starting from 84° N in the longtitude of Franz Josef Land and setting a new record of 86°13′ N before having to turn back and winter in Franz Josef Land. In the spring, by a strange and lucky coincidence, he met Frederick Jackson, a British explorer, and returned home in his ship. Jackson was investigating Franz Josef Land as a possible steppingstone to the pole, but on hearing Nansen's account gave up the polar attempt. In his three-year stay (1894–97), however, he revolutionized the map of this complicated collection of islands and did a great deal of valuable work.

Up to that time the desire to reach the pole had been coupled with that of mapping unexplored territory and collecting scientific data; after the "Fram" expedition there was no longer any doubt that the central part of the polar basin was an ice-covered sea, and that any land still to be discovered would be peripheral. The race for the pole then degenerated into an international sporting event. Several expeditions, following in Jackson's footsteps, tried to reach the pole from Franz Josef Land. Three were American: Walter Wellman in 1898–99, the Baldwin–Ziegler expedition in 1901–02, and the Fiala–Ziegler expedition in 1903–06. An Italian expedition led by the Duke of the Abruzzi set a new record in 1900, when Capt. U. Cagni reached 86°34′ N.

Robert E. Peary started working toward his polar journeys in 1891–92 and 1893–95, when he made two long journeys across northwest Greenland, discovering the largely ice-free Peary Land. He intended to use north Greenland as his jumping-off place but later changed this to north Ellesmere Island, farther from the pole but more accessible. In 1898–1902 he laid a large supply cache in Lady Franklin Bay from bases in Smith Sound, sledged around the north coast of Greenland, and reached 84°17′ N from Cape Hecla, Ellesmere Island. In 1905, aided by the expert ice navigation of Capt. Bob Bartlett, he sailed in the "Roosevelt" to Cape Sheridan, near the "Alert's" old winter quarters, and from Cape Hecla set a new record of 87°6′ N. He also sledged around the north coast of Ellesmere Island, mapping the coast from where Nares left off. In 1908–09 he returned and from Cape Columbia in 1909 made a sledge journey to the pole. His technique was to start with a large party and set up depots at regular intervals, the support sledges turning back one by one as the depots were laid. With Peary on the final dash were his black dog-driver Matthew Henson and three Eskimos.

Just before Peary's return to the United States in September 1909, Frederick A. Cook, an American who had been with Peary in Greenland in 1891–92 and who had spent 1907–09 in the Arctic, announced that he had reached the pole the year before with two Eskimos, from the north point of Axel Heiberg Island. The matter aroused considerable controversy, and, to substantiate his claim, Cook submitted his journal to the University of Copenhagen, which considered it inadequate proof; he did not appear to challenge the decision. The question of whether Peary or Cook actually reached this theoretical point on the moving ice pack is hard to prove or disprove; but until there is fresh documentary evidence or improved knowledge that supports a different interpretation of the existing documents, Peary's claim seems the more valid one. In any event, it was accepted by the U.S. Congress and geographical institutions in many countries.

Early flights to the pole

The first attempt to fly to the pole was made as early as 1897 when a Swedish scientist, S.A. Andrée, left Spitsbergen in a balloon. It was not until 1930 that the fate of Andrée and his two companions was known, when their bodies and diaries were found on White Island. In 1909 Walter Wellman made an unsuccessful attempt by dirigible, and in 1925 Roald Amundsen, with two Dornier-Wal flying boats, reached 87°44′ N. The first to reach the pole was Richard E. Byrd, who with Floyd Bennett as pilot flew from Spitsbergen to the pole and back on May 9, 1926. Two days later Amundsen, with Lincoln Ellsworth and Umberto Nobile, set off from the same base in a semirigid airship and flew across the pole to Alaska.

Scientific exploration. An important secondary motive in much of the exploration so far discussed was pure scientific curiosity, the desire to add to the general store of knowledge of the world. With the passage of time and the disappearance of the old dreams of gain and glory, this became more and more the primary motive of exploration, though it was sometimes combined with a hope of developing such mineral or other resources as might be found.

In 1875 an important proposal for international cooperation in collecting scientific data was made by Karl Weyprecht, and the suggestion led to the establishment of the first International Polar Year, 1882–83, during which stations throughout the Arctic took observations and pooled the results. The countries participating were Norway, Sweden, Denmark, Finland, Russia, The Netherlands, Germany, Austria, the United States, and Great Britain. The 11 stations, reading eastward from Spitsbergen, were Isfjord (Ice Fjord), Spitsbergen; Bossekop, north Norway; Sodankylä, Finland; west coast of Novaya Zemlya; Sagastyr Island, Lena Delta; Point Barrow, Alaska; Great Slave Lake; Lady Franklin Bay, Ellesmere Island; Cumberland Sound, Baffin Island; Godthåb, Greenland; and Jan Mayen Island. In 1932–33 a similar pattern was followed by the second International Polar Year, but with more stations, and the technique was extended to cover the whole world in the International Geophysical Year of 1957–58.

Spitsbergen. Starting in 1827 a series of expeditions, most of them Swedish, surveyed Spitsbergen and studied its geology and natural history. Among those who carried out this work were B.M. Keilhau, Otto Torell, and A.E. Nordenskiöld. Sir Martin Conway crossed the interior of Vestspitsbergen in 1896–97, and in 1898 A.G. Nathorst explored the east coast and adjacent islands. Oceanographic and other work was done by the Dutch in the "Willem Barents" after 1878, by the Prince of Monaco and W.S. Bruce (1898–1914), and by the Russian admiral S.O. Makarov in the icebreaker "Yermak" (1899). At the turn of the century coal mining was begun in Isfjord, and this led to further survey activity by Norwegian government expeditions and others. A British expedition from Oxford University under George Binney in 1924 was the first scientific expedition to make extensive use of an aircraft.

The Russian Arctic. Between 1821 and 1824 F.P. Litke of the Russian Navy made four voyages to Novaya Zemlya, surveying the west coast and improving the mapping of Matochkin Strait and the White Sea coast, and in 1832–35 P.K. Pakhtusov surveyed much of the east coast of Novaya Zemlya. In 1880 Leigh Smith made the first of two voyages to Franz Josef Land and was the first to sail a ship there under its own power. On his second voyage his ship, the "Eira," was nipped by ice and sank. Smith built a hut on the shore and wintered, surveying the south coast and collecting scientific data. In the spring the party sailed to Novaya Zemlya in small boats. In 1886 and again in 1893 and 1900–02 Baron E. von Toll, a Russian explorer, worked in the New Siberian Islands. He perished in an attempt to find Sannikov Land, an island reported north of the New Siberian Islands, which, like many similar "lands" in the Arctic, probably does not exist. Some coordinated hydrographic work was done by the Russians in the Barents Sea from 1898 to 1908, in the Kara Sea from 1894 to 1904, and east of Cape Chelyuskin from 1910 to 1915.

In 1918 Amundsen set out in the "Maud" to emulate Nansen's drift in the "Fram" but with the hope of getting into a more northerly latitude by starting the drift nearer to Bering Strait. He took three seasons to sail east through the Northeast Passage, and it was not until 1922 that the "Maud" began its drift, under the scientific leadership of H.U. Sverdrup. In two years it was carried back to the New Siberian Islands, duplicating the path of the "Jeannette" rather than the "Fram," but useful scientific work was done throughout both phases of the expedition.

Increase in Russian exploration

After the Russian Revolution in 1917, the scale and scope of exploration increased greatly as part of the work of developing the northern sea route. Polar stations, of which five already existed in 1917, increased in number, providing meteorological, ice reconnaissance, and radio facilities. By 1932 there were 24 stations, by 1948 about 80, and by the 1970s more than 100. The use of icebreakers, and, later, aircraft, as platforms for scientific work

was developed. In 1929 and 1930 the icebreaker "Sedov" carried groups of scientists to Franz Josef Land and also to Severnaya Zemlya, the last major piece of unsurveyed territory in the Soviet Arctic; it was completely mapped under G.A. Ushakov between 1930 and 1932.

The one-season voyage of the "Sibiryakov" through the passage in 1932 accomplished much scientific work and was the first to use the route north of Severnaya Zemlya. It gave a further stimulus to development of the sea route, and icebreaker operations to study sea and ice became annual. Particularly worth noting are three cruises of the "Sadko," which went farther north than most; in 1935 and 1936 the last unexplored areas in the northern Kara Sea were examined and the little Ushakova Island discovered, and in 1937 the ship was caught in the ice with two others and forced to winter in the Laptev Sea, adding valuable winter observations to the usual summer ones.

Greenland. The history of modern Greenland may be said to begin with the voyage in 1721 of Hans Egede, a Danish-Norwegian missionary whose aim was to find and reestablish the Norse colonies. Travelling up and down the west coast, he found no survivors of the old colonists, but he stayed to found his own settlement at Godthåb and to begin the enlightened development of the country and its Eskimo people that has made Danish Greenland a model of colonial administration.

Greenland has received a great deal of study. The west and north coasts became fairly well known during the 19th century. The east coast was less easily explored because of severe ice conditions that make it hard to approach by ship. In 1806–13 Karl Ludwig Giesecke, a German mineralogist, used the native umiak to study the southeast coast, and so did Lieut. W.A. Graah in 1829–30. In 1823 captains D. Clavering and E. Sabine, following in the steps of William Scoresby the year before, carried the survey north to 76° N and took pendulum observations. In 1876 the Danish Committee for the Geographical and Geological Investigation of Greenland was formed, and since then a consistent program of research has been carried out. The gaps in the southeast coast were filled in by naval expeditions under L.A. Mourier (1879), G. Holm (1883 and 1885), and C.H. Ryder (1891–92), and the rest by Lieut. G.C. Amdrup in the "Antarctica" (1898–1900), the Duke of Orléans in the "Belgica" (1905), and Ludvig Mylius-Erichsen in the "Danmark" (1906–08). On the latter expedition long sledge journeys by J.P. Koch and Mylius-Erichsen traced the whole northeast corner of the island, but Mylius-Erichsen and two companions were lost. The last details were recorded in 1901–12 by Ejnar Mikkelsen, who traced the route followed by the dead men and found their records. A series of expeditions, known as the Thule Expeditions because they were based on the little trading settlement of Thule in northwest Greenland, did considerable work in north Greenland between 1912 and 1921 under Knud Rasmussen and Lauge Koch.

Expeditions to the Greenland ice cap

The Greenland ice cap presented a formidable barrier to travellers and at the same time a challenge to both adventurer and scientist. Early attempts to penetrate it from the west coast settlements were made by Edward Whymper in 1867, by Nordenskiöld in 1870 and 1883, by J.A.D. Jensen in 1878, and by Peary in 1886. Peary, the most successful, penetrated 100 miles from the coast. In 1888 the young Fridtjof Nansen adopted the bold plan of starting from the uninhabited east coast, thus leaving himself no retreat but at the same time avoiding the necessity of retracing his steps. With five companions, using snowshoes and skis, he crossed the ice cap from 64°23′ N on the east coast to Godthaab on the west. Since then the ice cap has been crossed many times, even in its widest parts. Peary was the first to cross the northern part, from Inglefield Gulf to Independence Fjord in 1892. Other crossings have been made by Knud Rasmussen and by A. de Quervain in 1912, by J.P. Koch in 1913, by Lauge Koch in 1921, and by others.

In 1930–31 three expeditions, simultaneous but independent, maintained stations on the ice cap throughout the winter, securing meteorological data vital to the study of world air circulation. They were the British Arctic Air Route Expedition led by H.G. Watkins, the German Greenland Expedition under Alfred Wegener, and the University of Michigan Expedition under W.H. Hobbs. After World War II this work was continued on a larger scale by the French explorer Paul Emile Victor (1947–53) and the British North Greenland Expedition under Comdr. C.J.W. Simpson (1952–54).

The North American Arctic. By the beginning of the first Polar Year in 1882, most of the coastlines of the North American Arctic were known except for the islands west of Ellesmere Island and the south and west coasts of Ellesmere. The U.S. Polar Year station at Lady Franklin Bay, in addition to its scientific program, explored a considerable amount of new terrain in Ellesmere Island and reached 83°24′ N on the north coast of Greenland, a record northing at the time. Led by Lieut. A.W. Greely, the expedition set up its station, Ft. Conger, in 1881. In 1883, as no supply vessel had arrived, Greely started south in five small boats, according to instruction, and reached Cape Sabine in Smith Sound. There a quite inadequate depot awaited him, together with a record to the effect that the supply ship "Proteus" had sunk in Kane Basin. After a terrible winter the survivors—seven from an expedition of 25—were rescued by Capt. W.S. Schley in the "Thetis."

Three earlier expeditions by Americans in search of Franklin's records are worth noting. C.F. Hall, having failed in a plan to reach King William Island by boat from Baffin Island, spent the years 1860–62 in Frobisher Bay, which only then, three centuries after its discovery, was proven not to be a strait; he found interesting relics of Frobisher's visits. From 1864 to 1869 he lived among the Eskimos at Repulse Bay and made an overland trip to the south coast of King William Island. In 1878 Lieut. F. Schwatka travelled overland from Hudson Bay and made the first summer search in the area, returning by a remarkable winter journey to Hudson Bay. Franklin's scientific records were never found, either then or since.

Exploration by the Geological Survey of Canada

Further exploration of the interior was carried out by the Geological Survey of Canada, notably by the journeys of J.B. Tyrrell, A.P. Low, and Robert Bell. Between 1884 and 1897 four Canadian government expeditions studied conditions in Hudson Strait and Hudson Bay with a view to establishing a sea route, and after 1903 a series of voyages into the archipelago by Low and Capt. J.E. Bernier visited many of the islands and did some survey and geological work. Two Germans, Franz Boas, the well-known anthropologist (1883–84), and Bernhardt Hantzsch (1909–11), contributed to the geography of Baffin Island.

In 1898–1902 a Norwegian scientific expedition in the "Fram" under Otto Sverdrup did a tremendous amount of work in south and west Ellesmere Island and north Devon Island and discovered three islands to the west—Axel Heiberg Island and the Ringnes Islands. Sverdrup's original intention had been to try to circumnavigate Greenland, but heavy ice in Kane Basin forced him to change his plans. The last gaps in the outline of Ellesmere Island were filled in by W.E. Ekblaw, geologist and botanist with the Crocker Land Expedition (1913–17) under D.B. MacMillan. Crocker Land, which Peary in 1906 conjectured to be north of Axel Heiberg Island, proved to be nonexistent; MacMillan failed to find it in a 200-mile journey over the ice.

The last large-scale expedition in the old tradition in the North American Arctic was the Canadian Arctic Expedition, 1913–18, led by Vilhjalmur Stefansson. It was divided into two parties, of which the southern one, under R.M. Anderson, did survey and scientific work on the north mainland coast from Alaska to Coronation Gulf, while the northern travelled extensively in the northwest, discovering the last remaining islands in that area. Stefansson, a magnificent hunter, successfully adopted Eskimo methods and was able to travel long distances by living off the land, avoiding the necessity of carrying large quantities of supplies. In 1921–24 the fifth Danish Thule Expedition under Rasmussen worked in Melville Peninsula and Baffin Island, and Rasmussen journeyed overland to King William Island and on to Alaska, studying the Eskimos. In the 1930s a number of British expeditions under Noel Humphreys, J.M. Wordie, and T.H. Manning worked in the Canadian Arctic; Manning completed the mapping of

the west coast of Baffin Island, the last major gap on the map of Canada. The last new land, however, was not added until 1948, when a Royal Canadian Air Force photo-survey aircraft found three islands in Foxe Basin, one of them, Prince Charles Island, about 3,500 square miles in area.

The exploration of Alaska after its purchase by the United States in 1867 proceeded slowly at first but later at a rapid speed. Coastal surveys by the Coast and Geodetic Survey were started immediately; among others, inland journeys were made by I. Petrof (1880); by Lieut. F. Schwatka (1883) and Lieut. H.T. Allen (1885), both of the U.S. cavalry; and by Lieut. G.M. Stoney, U.S. Navy (1883–85), one of whose party made the first overland journey to Point Barrow. After the Yukon gold strike of 1897, the Geological Survey began a large-scale systematic study of all Alaska that has continued along with work in other fields of natural science. Outstanding in the early days of the project were A.H. Brooks, chief geologist of the Geological Survey in Alaska, 1903–17, and E. de K. Leffingwell, who worked on the North Coastal Plain between 1906 and 1914.

The Arctic Ocean. It is a comment on the unimportance of the North Pole as an incentive to exploration that hardly any of the real exploration of the Arctic Ocean can be credited to the pole seekers. The great exception is Nansen, whose work in the "Fram" stood alone until the 1930s; but, although Nansen made a bid to reach the pole, his primary aim was rather to study the waters and bottom contours of the Arctic Ocean and the drift of the ice and to find out whether there were new lands still to be discovered in the centre of the polar basin. In accord with popular opinion, Nansen expected to find only shallow water in the North Polar Basin. Consequently the "Fram" was equipped to make soundings to only 6,000 feet. Soon after the drift began, depths in excess of this were encountered. Unravelling one of the ship's steelwire cables and joining the individual strands end-to-end permitted 11 successful soundings to be completed under most difficult conditions. These soundings gave depths ranging from 11,000 to 13,000 feet, showing that there was a deep basin under at least part of the North Polar Sea. These deep soundings mark the true discovery of the Arctic Ocean.

Exploration by aircraft

The advent of the airplane revolutionized exploration techniques and made possible investigation of the Arctic Ocean on a scale never before dreamed of. Following the polar flights of Byrd and Amundsen, G.H. (later Sir Hubert) Wilkins and C.B. Eielson made the first flight by airplane across the Arctic Ocean in 1928, from Point Barrow to Spitsbergen, and in 1937 two long-distance transpolar flights were made by Soviet flyers, V.P. Chkalov and M.M. Gromov. These flights were mainly intended to prove the capabilities of aircraft, but a third Soviet flight in the same year made a large but tragic contribution to exploration. A four-engined aircraft piloted by S.A. Levanevski disappeared in the Arctic Ocean and set in motion a large-scale, though unsuccessful, search which covered vast areas hitherto unexplored and added tremendously to flying experience.

It was also in 1937 that the U.S.S.R. set up the first floating scientific station, using four-engined aircraft based on Franz Josef Land to land a four-man party under I.D. Papanin at the North Pole in late May. The station, now known as North Pole 1, drifted south for nine months and was taken off its melting ice floe in the Greenland Sea. In the same year the icebreaker "Sedov" was caught in the ice in the Laptev Sea and began a 27-month drift that almost duplicated that of the "Fram" and yielded interesting comparative data. In 1941 an aircraft carrying a team of scientists made three landings on the ice at about 80° N, 175° E.

After World War II scientific work in the Arctic Ocean increased greatly, and today there remain no unexplored areas. After 1947 the United States carried out routine weather-reporting flights over the Arctic Ocean from Alaska and used icebreakers and aircraft to do oceanographic work in the Beaufort Sea. In 1952 a weather station was established on the ice island T3 and maintained for two years, being reoccupied briefly in 1955, and on a more permanent basis as an International Geophysical Year station in 1957. From that time there was continuous occupation of stations, usually two at any given time.

The Russians explored the Arctic Ocean on a large scale by means of floating stations and large-scale airborne expeditions which made many landings on the ice to take observations. Station North Pole 2 was established in 1950 north of Wrangel Island and was maintained for a year. After 1954 there was a continuous succession of stations, usually two at once and each occupied for one or two years or more, until they drifted into a region where they ceased to be of interest or joined the drift to the Greenland Sea. North Pole 15 was set up in April 1966.

Technological advances in power (nuclear engines), marine technology (especially designed icebreakers, scientific vessels, and commercial ships), and navigation (inertial and satellite guidance systems) have permitted expeditions such as the 1958 voyages of the USS "Nautilus"

By courtesy of the U.S. Department of the Navy

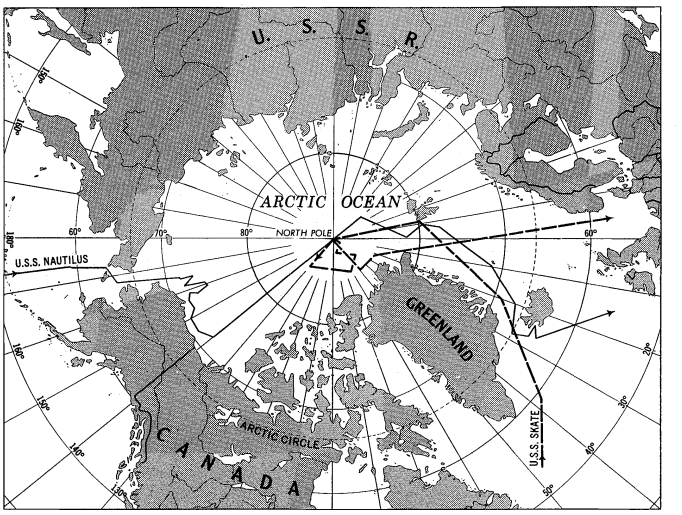


Figure 4: Arctic routes of the transpolar voyages of the U.S. submarines "Skate" and "Nautilus," August 1958.

and "Skate," which made successful polar cruises, in both winter and summer, and proved to be excellent platforms for oceanographic work; the 1969 and 1970 voyages of the "Manhattan," an ice-breaking oil tanker; and the 1978 voyage of the Soviet nuclear icebreaker "Sibir," opening a southern passage across the coast of northern Siberia. The discovery of oil in near- and offshore polar areas of North America provided the economic impetus for much geological exploration of the Beaufort Sea, the Canadian Arctic Islands, and other areas in the 1970s and 1980s.

(N.A.O.)

The discovery phase of Arctic exploration is over; there is no longer any possibility of finding new lands. Photo surveys have provided reasonably accurate maps, and improved aircraft and base facilities are making the once formidable "frigid zone" increasingly accessible; commercial airlines fly across the North Pole. Nevertheless, much detailed exploration of lands and seas remains to be done.

(M.Dr.)

The Arctic Islands

Location and general characteristics

The islands lying north of the Canadian mainland, forming one of the world's great archipelagos, extend about 1,500 miles east to west and 1,200 miles from the mainland to Cape Aldrich, the northernmost point of Ellesmere Island. Two large peninsulas, Boothia and Melville, extend northward from the mainland and will be discussed here as parts of the archipelago. The land area of this region exceeds 500,000 square miles, about one-seventh of the area of Canada.

The Arctic Islands, also called the Canadian Arctic Archipelago, lie far north of the temperature line, or isotherm, along which 50° F (10° C) is the mean July temperature; this line coincides closely with the northern limit of trees

Figure 5: The Arctic Archipelago of North America.

on the mainland and commonly is used to define the southern limit of the world's Arctic region. Like much of the adjoining mainland, the Arctic Islands are characterized by permanently frozen ground, or permafrost, except for a thaw to the depth of about one foot during the summer months.

Ice restricts sea access to most of the islands to the late summer and early winter. The few and widely scattered permanent settlements lie mainly on the southern islands. Although remote, the region has become accessible the year round through a network of airfields and air routes connecting centres of population to the south. For thousands of years the archipelago was peopled only by Eskimo, who evidently migrated to this land from Asia. Norsemen may have reached the easternmost islands in the 10th and 11th centuries, and from the 16th century numerous European expeditions carried out a prolonged search for a northwest passage through the islands to the Orient.

A few minerals and rocks have long been used by the aboriginal Eskimo—native copper and flint for knives and weapons and soapstone for lamps and cooking pots—but economically valuable minerals have played only a small part in the history of the Arctic Islands. Exploitation began with mining and shipping of "gold ore" (probably amphibolite with minor, bronze-coloured biotite) from Baffin Island in the 16th century by the English explorer Sir Martin Frobisher. Graphite, mica, and coal have been mined on the same island. In the late 20th century, increased mineral exploration has resulted in promising discoveries of iron and lead–zinc. Most spectacular, however, was the modern oil and gas exploration program that followed discovery of a major oilfield in northern Alaska in 1968. An oil-producing zone also has been found by drill, and several wells have been capped as gas wells.

The Canadian Arctic Islands and Boothia and Melville peninsulas form the District of Franklin, one of three administrative divisions of Canada's Northwest Territories. Government is by both elected and appointed councillors seated at the territorial capital, Yellowknife, on the mainland.

PHYSICAL GEOGRAPHY

Relief and geology. The surface of the archipelago includes plains, lowlands, uplands, plateaus, highlands, and mountains. Most elevated surfaces are rolling and dissected by valleys and deep channels. Physiography generally expresses the type and structure of the bedrock: extremely resistant granitic and gneissic rocks form mountains, highlands, and uplands; folded sedimentary rocks form ridged mountains and uplands; slightly folded or flat-lying limestones and related rocks form extensive plateaus and lowlands; and weakly consolidated sandstones and shales underlie lowlands and plains.

Geology and geomorphology of the archipelago

Mountains and highlands lie along the eastern coasts of Baffin, Devon, and Ellesmere islands to form a high eastern edge of the archipelago. The surface slopes generally westward through plateaus and lowlands interrupted by upland and some mountain areas, then drops away to the low Arctic Coastal Plain that lies along the northwestern margin of the region. The eastern mountain belt, in which the highest land attains an elevation of about 7,000 feet,

is deeply dissected by fjords to form a region of high relief and rugged scenery. Another belt of rugged mountains forms Axel Heiberg Island and the northern edge of Ellesmere Island. The highest point of the Canadian Arctic Islands lies in northern Ellesmere Island: Barbeau Peak has an elevation of 8,544 feet (2,604 metres). Plateaus in the central part of the archipelago reach average elevations of 800 to 2,000 feet, with hills and mountains of about 2,500 feet.

Structural regions. The Canadian Arctic comprises two geological provinces: the Central Stable region in the south and the Innuitian orogenic region (a once mobile belt) in the north. The Central Stable region includes the vast crystalline complex of the Canadian Shield, formed during the Precambrian (with radiogenic dates between 2,800,000,000 and 600,000,000 years ago), and large areas of overlying, little-deformed sedimentary rocks of Late Precambrian and Paleozoic (more than 225,000,000 years ago) eras. The Innuitian region is a sinuous belt of slightly to severely folded rocks ranging in age from the Late Precambrian Era to the Tertiary Period (*i.e.,* up to as little as 2,500,000 years ago).

The Innuitian region was the site of two major sedimentary basins: the Franklinian Geosyncline and the Sverdrup Basin. The Franklinian Geosyncline, the earlier of these tectonic depressions, or regional downwarps, received sediments nearly continuously from Late Precambrian to Late Devonian times (about 345,000,000 years ago). An episode of folding and faulting of the geosyncline in latest Devonian time is named the Ellesmerian orogeny, or mountain-building episode. Following this orogeny, much of the area was again depressed to form the Sverdrup Basin, which received sediments between Late Mississippian and Early Tertiary times. The rocks of the Sverdrup Basin were deformed during the Early Tertiary, some 65,000,000 years ago, by compressive forces of the Eurekan orogeny.

Ages and thicknesses of geological formations. The Canadian Shield of Precambrian metamorphic rocks forms a "basement" complex along the southern and eastern edges of the archipelago. Rocks of the Shield are mainly granite, gneiss, and schist. Late Precambrian to Tertiary sedimentary rocks overlie the shield to form northeast-trending belts to the north and northwest; these younger rocks include limestone, dolomite, shale, sandstone, gypsum, and volcanic rocks such as andesite and basalt. Bedded rocks of Late Precambrian Era attain thicknesses of 12,000 to 19,000 feet on Victoria and Baffin islands but are absent over large areas due to later erosion. More widely exposed on the southern islands are Early Paleozoic sedimentary formations with a maximum total thickness of 10,000 feet. The Early Paleozoic formations pass northward into a 40,000-foot-thick, mainly marine sedimentary sequence that occupies the Franklinian Geosyncline. The geosynclinal sequence passes from marine to nonmarine in the uppermost (Devonian) part, in which thin coal seams, as well as plant and fish remains, occur in sandstones. On northern Axel Heiberg and Ellesmere islands, volcanic rocks and related sediments occur along with a metamorphic complex that in part predates the geosyncline but may in part derive from it. The Sverdrup Basin, overlying the folded and partially eroded geosyncline, contains some 50,000 feet of mainly marine sedimentary rocks. Nonmarine strata in the upper part of the sequence include coal and basaltic flows. Late Paleozoic anhydrite deep in the basin was forcibly intruded into overlying beds by orogenic pressures or by the weight of the superimposed rocks and now appears at the surface as conspicuous gypsum–anhydrite diapirs, or "domes."

The Arctic Coastal Plain is a narrow belt of very low relief underlain by Late Tertiary and Early Quaternary sand and gravel and is presumably continuous with the continental shelf, which extends about 100 miles into the Arctic Ocean.

Final shaping and drainage. The form and drainage pattern of the archipelago probably dates from Late Tertiary and Pleistocene times, or since about 25,000,000 years ago. The bedrock was deeply eroded after the Early Tertiary orogeny, and remnants of the erosion surface now stand at various elevations. Faults evidently border many islands, and many channels and major valleys probably are related to Tertiary faults. The subsequent drainage system was modified by Pleistocene glaciation and now is largely submerged in marine waters.

Glacial geology. Successive glacial advances and interglacial intervals of the Pleistocene Epoch followed worldwide climatic changes some 2,000,000 to 3,000,000 years ago. The Arctic Islands were largely ice covered during this epoch, as was most of the rest of Canada. The effects of the latest, or classical Wisconsin, glacial invasion (climaxing about 20,000 years ago) are most abundant and fresh, but earlier advances and interglacial events are recorded at many places. Some areas of the westernmost islands lack obvious evidence of glacial activity and may have escaped glaciation entirely.

Two main ice sheets occupied the islands: the Laurentide Ice Sheet flowing northward from the mainland to override the southern islands, and the Innuitian Ice Sheet, a coalescing series of glacier complexes that were especially well developed in the highlands and mountains of the eastern part of the archipelago. Markedly linear glacial features, such as drumlins (hills of glacier-deposited drift) and eskers (glacial stream deposits), are abundant on the lowlands of the southern islands, while glacial valleys, rounded rock surfaces, and other alpine glacial features are evident in the higher eastern islands.

Effects of the ice sheets

Layered sediments containing plant and woody material occur at several localities beneath glacial till, between till layers, and overlying older till beyond the recognized limit of the late Wisconsin Ice Sheet. Radiocarbon ages greater than 30,000 years have been obtained from the organic matter, which includes pollen of grasses and sedges, other herbaceous plants, and alder, willow, and spruce trees. Willow and spruce wood, some beaver gnawed, is found. Evidently the interglacial climate was somewhat more favourable than at present.

Marine shells yielding radiocarbon ages from 20,000 to more than 40,000 years occur in stratified deposits up to a few hundred feet above sea level and as scattered debris at higher elevations, as on upland surfaces. The dates are minimums, and the "old" shells indicate marine events prior to the last glaciation. Some of the older shells are broken and abraded and must have been transported and dispersed by glacial ice.

Deglaciation in the Arctic Islands was well advanced by about 9,000 years ago, as dated by shells from the highest beaches. Deglaciation of the islands, as elsewhere, appears to have been marked by fluctuations of the ice front. Scattered proglacial lakes formed as the ice abated, and the sea invaded the land, which was then still depressed because of the former ice load. Remarkable flights of raised beaches, now standing from a few feet to about 500 feet or more above sea level, attest to postglacial uplift following removal of the ice load. The greatest uplift occurs in a linear, nearly central zone trending northeasterly through the islands from Boothia Peninsula. The emerged marine features, however, present a complex pattern because of the rise in sea level during deglaciation, the variation in time of ice retreat—and hence of marine incursion—from area to area, and possible deformation or possible fault movement along some coasts.

Climate. The southern boundary of the Arctic climate region is usually placed at the tree line, which lies on the Canadian mainland. North of this distinct natural boundary, the growing season is too short and cold to allow tree growth.

The climate of the Arctic Islands is unique in many respects because of the high latitude, the barren nature of the land surface, and the maritime influence of the Arctic Ocean, Baffin Bay, and the interisland waterways. The winters are longer and the summers cooler than more southerly parts of Canada, but conditions in the Arctic are usually not so fierce or extreme as some popular accounts would suggest.

The Arctic region is snow and ice covered for more than half the year, subjected to a cold, dry climate. The daily average temperatures of the three coldest months range from −20° F (−29° C) in southern sections to −30° F (−34° C) in the north. The average temperatures generally

Temperature ranges

remain below 0° F (−18° C) during six cold months and rise above the melting point only during June, July, and August. The record low temperatures of −55° to −60° F (−48° to −51° C) at Arctic stations are not as low as the −70° F (−57° C) and colder temperatures reported for parts of the mainland. The summer temperatures are cool, generally below 45° F (7° C), with occasional brief sunny intervals of 65° to 70° F (18° to 21° C). Winter sets in about the end of August in the highest latitude and in early or mid-September in the southern islands. The lowest temperatures occur between December and April, with January the most severe month in southern parts and February in the high latitudes. Thawing begins in late May to early June, and the snow has disappeared from much of the land by early July. The breaking up of sea ice, which attains thicknesses of about five to seven feet, occurs considerably later than the beginning of the thaw, but by late July it is well advanced in the southern and eastern periphery of the archipelago. Open water suited to ship navigation advances slowly into the archipelago during August and September, but the channels between the northwestern islands are ice choked the whole season (June to September).

The prolonged Arctic cold season is a result of the far northern latitude: incoming solar radiation is minimal during daylight because of the low angle of the sun's rays and is absent during the long Arctic night. In addition, although the amount of solar energy available during the periods of long days and continuous daylight is even greater than in southern latitudes, only a small percentage reaches the ground because of the high reflectivity of the snow and ice surfaces and cloud layers. The sustained cooling results in a continuous winter ice cover on the seas and channels, generally stable, cloud-free air, and consequent light snowfall. During the thaw season the open water and wet, thawing land result in high humidity and notorious, low-lying stratus clouds and coastal fogs.

Annual precipitation is low—about three to seven inches—throughout the archipelago except along the mountainous eastern coastal sections, where the height of the land and the proximity to the open waters of Baffin Bay, Davis Strait, and the North Atlantic result in anomalously high precipitation and persistent ice caps.

The worst climate in the Arctic Islands is that of the Hudson Strait region. Dominated by open water and frequent cyclonic activity, this region has the highest average temperature, the heaviest snowfall, the highest average wind speeds, and the greatest number of days of summer fog in the Canadian Arctic. (R.L.Ch.)

Plant and animal life. Plants and animals of the Canadian Arctic Islands are relatively few for so large an area but are of great interest through their adaptations to extremely severe conditions. Spitsbergen, lying far north of the Scandinavian Peninsula but bathed by the Gulf Stream, supports many temperate plants that cannot grow in more southerly Canadian islands, which are surrounded by cold water flowing from the Arctic Ocean. The continuous ice and smooth wind-packed snow cause seeds and other reproductive structures to be blown great distances in winter gales. Many plants are circumpolar without local variation, indicating dispersal between Old and New World lands. There is evidence that some insects may be carried similarly.

Environmental hazards and adaptive mechanisms. Flowering plants must flower and fruit rapidly in a short, cold summer, freeze with impunity at any time, resume full growth as soon as they thaw, and withstand abrasion by gritty snow crystals in winter gales. All other Arctic plants—ferns, horsetails, mosses and liverworts, lichens, algae, and fungi—as well as insects—the primitive arthropods known as springtails, spiders, and mites—share the ability to freeze at any time and resume full metabolism as soon as the temperature permits it. This is the most important characteristic of all Arctic organisms except birds and mammals. A few crucifers, members of the mustard family, may mature their seed in a second year if snow cover is complete, while many fungi and insects mature gradually in two to several years, stopping growth at any stage and continuing the next spring.

Suspension and reactivation of metabolic functions

Although small animals may take shelter from gales, plants must either grow in sheltered sites or tolerate snow abrasion. A common protection is assumption of a compact cushion, in which the winter buds lie below the level of the old stems and leaf tips. The projecting parts cause eddies that slow down snow particles.

Mammals of or above the size of Arctic hares can carry fur heavy enough to give them excellent insulation, and they need little shelter. Such coats are unmanageable for such small species as ermine and lemmings, which must spend most of the winter below ground or below the snow. Snowfall is often very light, and large areas may be swept bare by wind. Soil temperature may then be so depressed that lemmings fail to reach breeding condition.

Native species. The most widespread mammals are Arctic fox, wolf, ermine, polar bear, Arctic hare, collared lemming, caribou, and musk-ox. The brown lemming (*Lemmus sibiricus*) occurs on the southern islands, while the red fox and wolverine occur on Baffin Island.

Most birds leave the Arctic Islands before winter begins, but a few willow ptarmigan, snowy owls, and ravens (*Corvus corax*) remain. Common birds in the high-latitude Queen Elizabeth Islands include the red-throated loon, snow goose, old-squaw and king eider ducks, rock ptarmigan, a few sandpipers, long-billed jaeger, glaucous gull, Arctic tern, and a few buntings. About 35 other species, notably plovers and gulls, breed regularly in the southern islands.

The most important fish in the archipelago region is the Arctic char, a staple food item of the Eskimo and increasingly popular for shipment south. No reptiles or amphibians reach the islands.

The insect life is of major consequence. A few moths and bumblebees reach northern Ellesmere Island, but the true flies, or Diptera, are by far the most abundant order. They include many midges, a few mosquitoes, and some muscoid flies. By their great numbers, they are important in pollination, and they are the main food for various birds. Beetles are few, except in the southern islands. Other orders are scarce or absent.

About 325 species of flowering plants grow in the region. Grasses, sedges, and rushes are common, but most other monocotyledons are absent. Dwarf birch and several willows reach the southern islands, but only one prostrate willow reaches northern Ellesmere. Other conspicuous species or groups include mountain sorrel, bistort, various Caryophyllaceae, buttercups, various Cruciferae, poppies, saxifrages, cinquefoils, mountain avens, crowberry, fireweed, louseworts, fleabanes, and dandelions.

Several ferns, horsetails, and clubmosses are recorded, all southern species that survive on sterile sites where competition is lacking. Mosses, liverworts, and some groups of algae are plentiful; but few species are endemic to the Arctic, and some are nearly cosmopolitan. Lichens are numerous and often conspicuous. They are important in delaying water loss from gravel, thus helping to establish mosses and flowering plants. Large lichens are major items in the diet of caribou.

Fungi are of great ecological importance, for they largely replace bacteria in breaking down plant remains. Various mushrooms and a few puffballs and large cup fungi grow in soil; and nearly 300 species of microfungi are parasitic or saprophytic on leaves and stems.

Plant cover is governed by summer climate and by soil and rock type. The numbers of vascular plants correlate closely with the July mean temperature; the flora diminishes precipitously as the freezing point in summer is approached. In many of the western islands, low relief and rapidly weathering sandstone or limestone give extremely barren substrates with much less than 1 percent ground cover.

Total productivity is not known for any site, but breeding-bird censuses give rough comparisons. Isachsen yielded 9.61 birds per square mile; Hazen Camp, 24.99; and grassy tundra in southwestern Baffin, 391 and 544. (In contrast, temperate forests and swamps range from 1,000 to 11,785.) Such figures emphasize the low productivity and the precarious balance of Arctic ecosystems.

(D.B.O.S.)

HISTORY

In the 19th century British interest in the Northwest Passage brought Sir John Ross (1818) and Sir William Edward Parry (1820) to the Arctic archipelago and led to the discovery of the magnetic pole on the Boothia Peninsula by Ross's nephew, Sir James Ross (1831). Sir John Franklin's ill-starred expedition by sea (1845–48), his disappearance with all his men, and the loss of his ships occasioned the famous search (1848–59) that incidentally added much to the knowledge of the Arctic coast.

Until the mid-20th century the total white population consisted of about 50 on Baffin Island and perhaps 20 more people scattered through adjacent islands. With the advent of air services and the extension of meteorological and radio research, this number increased considerably. A notable event was the Canadian military expedition ("Exercise Musk-ox") early in 1946, in which a group of army men travelled by tractors from Churchill to Cambridge Bay on Victoria Island, returning via Coppermine and Norman Wells. Of great interest also were the voyages of "St. Roch," which twice traversed the Northwest Passage. In 1941 and 1942 it made the journey from Cambridge Bay to Pond Inlet in Baffin island, wintering near the magnetic pole. In 1944 Sgt. Henry Larsen made the return journey and traversed the difficult portion of the journey—from Pond Inlet to the Beaufort Sea—in 18 days. He found the narrow Prince of Wales Strait (north of Victoria Island) practically free of ice and so was able to accomplish this feat of navigation in record time. The "St. Roch" was the only vessel to traverse the historic passage both ways. An expedition in 1947 investigated the site of the north magnetic pole. Survey by the aeroplane "Aries" seemed to show that it had moved north from Boothia Peninsula by perhaps as much as 300 miles. About that time the United States cooperated with Canada in establishing weather and loran (long range radar aid to air navigation) stations on various islands of the archipelago, including such locations as Eureka Sound, Winter Harbour, and Cambridge Bay, and on Resolution and Nottingham islands.

(Ed.)

Greenland

Greenland (Danish Grønland; Greenlandic Kalaallit Nunaat), the largest island in the world and an integral part of the Kingdom of Denmark, lies in the Northern Hemisphere, on a northeastern extension of the structural shelf fringing the continent of North America. Its icy, inhospitable environment is mostly within the Arctic Circle and extends to within less than 500 miles (800 kilometres) of the North Pole.

Physical dimensions

Greenland forms a wedge-shaped mass of about 840,000 square miles (2,175,600 square kilometres), of which more than 700,000 square miles are ice covered. Its maximum north–south extension is about 1,660 miles, and it is some 650 miles across at its widest point, at about 70° N. The length of its indented coastline has been estimated at 24,430 miles—almost exactly equal to that of the circumference of the Earth at its Equator. Its southernmost tip is Kap Farvel (Uummannarsuaq), at 59°46′ N, while Kap Morris Jesup, at 83°39′ N, was, until a recalculation in 1969 of the position of Kaffeklubben Island, some 20 nautical miles to the east, considered as the nearest land to the North Pole. The westernmost point, Kap Alexander, at 73°08′ W, is farther west than the U.S. city of Boston, while its easternmost point, Nordostrundingen, at 12° W, is almost as far east as Ireland. At its narrowest point, Greenland is only 16 miles from Ellesmere Island in the Canadian north.

A submarine ridge, no deeper than 600 feet, connects the island physically with North America, from which Greenland is separated at sea level, from north to south, by the Nares Strait, the Robeson Channel, the Kennedy Channel, the Kane Basin, Smith Sound, Baffin Bay, Davis Strait, and the Labrador Sea. To the north lies the Lincoln Sea and the ice masses of the Arctic Ocean, while on the east and south lie the Greenland Sea, Denmark Strait, and the Atlantic Ocean. At a depth of about 2,000 feet, the Yermak Plateau, another submarine ridge, connects Greenland with Spitsbergen, while at the same depth the Faeroe-Iceland Ridge snakes across the ocean floor to Scotland and the European continent.

PHYSICAL AND HUMAN GEOGRAPHY

The land. *Relief and geology.* Structurally, Greenland is an extension of the Canadian Shield, the rough plateau of the Canadian north, made up of hard Precambrian gneiss and granite rocks. In addition, the northern shoulders of Greenland (the mountains of Peary Land and their westward extension to Ellesmere Island) are structurally linked, via Spitsbergen, to the Caledonian orogeny of Europe.

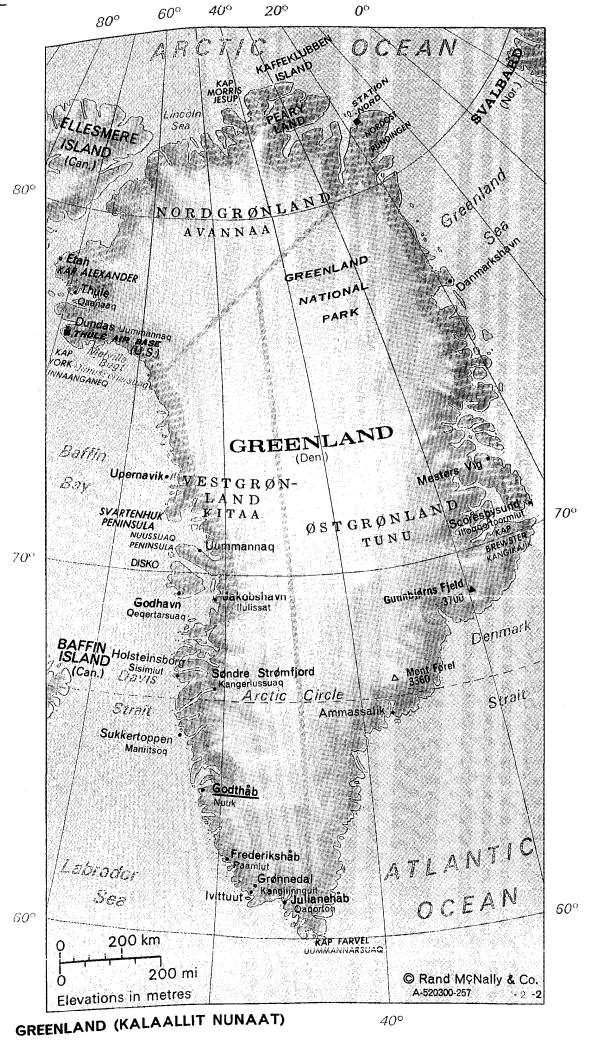
From the dawn of geological time, the vast mass of Greenland, now largely obscured by ice, has been molded by a series of volcanic and orogenic processes. At the earliest period, volcanic activity took place along the eastern edge. Later, sand and clay were washed out from ancient mountains into a long structural trough along the east coast. This became filled with layers of sandstone, limestone, and shale many thousands of feet thick, which were then uplifted to form the fold mountains of East Greenland. In more recent times, sandstones were laid down on both the east and west coasts, and coal of the Cretaceous Period, about 100,000,000 years ago, is found in the coastal areas of Disko Island (Qeqertarssuaq) and the Nuussuaq Peninsula. Later, about 50,000,000 years ago, further volcanic activity formed horizontal layers of hardened lava streams, which can be seen as black bands interlacing exposed rock faces. Basaltic survivals of this period are at the Scoresbysund (Scoresby Sound) hot springs area in the east, at Julianehåb (Qaqortoq) in the south, and on Disko Island and Nuussuaq and Svartenhuk peninsulas in the west. The fossils in the sandstones and coal deposits testify that a temperate climate existed in the Cretaceous Period.

The onset of the Ice Age, 1,000,000 or so years ago, totally changed the Greenland environment. Vast accumulations of ice slid from the interior to the coasts, scouring away the rock masses beneath through debris frozen to the base of the slowly moving ice. Lakes formed on the coastal periphery, and sand and clay plains also developed from sediments washed out from under the advancing ice. In the 20th century the indented, island-strewn coast forms, in part, a narrow, ice-free fringing strip. Long, deep fjords reach far into both the east and west coasts in complex systems, offering magnificent, if desolate, scenery. The Scoresbysund network on the east coast is the largest, with a length of about 185 miles and a breadth of 125 miles. A range of mountains 7,000 feet high also runs along the east coast, with Mt. Gunnbjorn, 12,139 feet, marking the highest point in Greenland. Along many parts of the coast, the ice sheet fronts directly on the sea, with large chunks breaking off the tongueing glaciers and sliding into the waters as icebergs.

The ice sheet

The ice sheet itself, the major feature of the Greenland landscape, is the largest ice mass found outside of Antarctica. It is contained by coastal mountains on the east and, to a lesser extent, on the west, although the rock floor far beneath its surface is at, or slightly beneath, current sea level. The average depth of the ice is 5,000 feet, with a maximum depth of 10,000 feet. Layers of snow falling on its barren, windswept surface become compressed into ice layers, which constantly move outward to the peripheral glaciers—the Jakobshavn Glacier, often moving 100 feet a day, is among the world's fastest glaciers. Nunataks—lofty, isolated peaks—emerge occasionally around the rim of the ice sheet, whose total area is 708,100 square miles.

Climate. The climate of Greenland is bleak and Arctic, modified only by the slight influence of the Gulf Stream in the southwest. Rapid changes, from dazzling sunshine to impenetrable blizzards, are common and result from the eastward progression of low-pressure air masses over a permanent layer of cold air above the ice. January average temperatures at Prins Christianssund, the southernmost station, are 21° F (−6° C) with July readings of 45° F (7° C), while at Station Nord, the northernmost station, the corresponding figures are −31° F (−35° C) and 39° F (4° C). Average monthly precipitation decreases from 9.3 inches (238 millimetres) in the south to 0.6 inch (15



GREENLAND (KALAALLIT NUNAAT)

millimetres) in the north, with summer rainfall concentrated in the southwest; snow can, and does, fall in any month. Although summers can be quite pleasant on the southwest coast, the inland ice is uniformly cold, with a July average of 10° F (−12° C) and a February mean of −53 F (−47° C).

In 1966 U.S. Army engineers drilled a 4,600-foot (1,400-metre) ice core near Dundas (Uummannaq), and in 1974 another 1,325-foot core was drilled in central Greenland. The cores have been studied to determine and date climatic fluctuations up to 100,000 years ago and to establish correlations with the fossil "raised beaches" of the coasts, which indicate relative changes in sea and land levels. The climate became noticeably warmer during the early part of the 20th century, with a 1930 maximum, associated with northward seal movements and an increase in the cod population. By the late 20th century, sea temperature was dropping again, causing a serious decrease in cod fishing.

Recent climate changes

Plant life. The vegetation of Greenland is represented mainly by tundra types, with heather, birch, willow, and alder scrub together with sedge, cotton grass, and lichen. The Greenland summer is rich in plant life. Of the 400 or more species, 300 are of North American origin, 50 or so were imported by the earliest Norse inhabitants, and the rest, on the isolated nunataks, have survived the ice ages. In southernmost Greenland, winter hay is grown for sheep, and along the western coast such vegetables as radishes, cabbage, and lettuce are raised, with potatoes in the extreme south.

Animal life. The rich animal life of the surrounding seas is the basis of existence for Greenlanders. The sea mammals—seals and whales—were formerly the main sources of nourishment, with ring- and black-sided seals predominating. Land mammals are represented by seven species—polar bears, musk-oxen, reindeer, arctic foxes, snow hares, ermines, and lemmings. Half of the breeding birds are native, and most of the remainder are from North America. The most important sea birds are eiders, guillemots, auks, wild geese, ducks, and gulls, while land birds include ptarmigans, ravens, white-tailed eagles, gyrfalcons, peregrine falcons, snowy owls, snow buntings, and longspurs. Among insects, mosquitoes appear in summer. Salmon and trout are found in the rivers, while cod, salmon, flounder, halibut, and *angmagssat* (capelin) are important saltwater fish.

The people. Ancient and modern influences had created by the late 20th century a uniquely blended society in Greenland. The strongest element stems from Eskimo culture, whose representatives, many with admixtures of European blood, are known as Greenlanders. By the late 20th century they accounted for more than three-quarters of the total population. The pure Eskimo are now found only in the far north, near Thule, and in East Greenland. In terms of origin, the Eskimo are believed to have crossed from North America to northwest Greenland, using the islands of the Canadian Arctic as stepping stones in a series of migrations that stretched from 4000 BC to AD 1000. Each wave of migration was represented by different cultures, beginning with Independence cultures (named after a nearby fjord), which followed the musk-ox and reindeer north to Peary Land. About the same time, the Sarqaq culture spread along the west coast and even reached east Greenland, although the lack of hunting grounds impeded movement around the northern rim of the island. Later migrations brought the Dorset, Dundas (Thule), and Inugsuk cultures.

The Eskimo contribution

Greenlanders today retain a clear linguistic and cultural identity (the Eskimo refer to themselves collectively as Inuit, or Innuit, meaning "the people"). The Eskimo dialects, taken together, form a language group differing from all others. A Greenlander grammar (1851) and a dictionary (1871) built up an Eskimo literary language, and Eskimo is the predominant spoken language of Greenland. The material culture, especially in the north, includes such beautifully constructed practical hunting equipment as the kayak, often considered the most elegant boat in the world.

The second element in Greenland society is represented by European, specifically Danish, influence. Apart from the early Norse settlers—whose colonies became extinct

MAP INDEX

by the early 15th century as a result of climatic deterioration and who left only an archaeological legacy—the Eskimo remained the sole inhabitants of the island until 1721, when a Danish missionary initiated new settlement. Greenland was a closed society until 1951, and Danish immigration was small, consisting largely of administrators. Faeroese and Danish settlement increased during the 1950s and 1960s. The modernization of Greenland society was given great impetus by developments in air communications and global strategy during and following World War II. Lying between the great superpowers of the Northern Hemisphere, Greenland became of prime importance in radar and air communications and in weather observations. One result was the construction of the giant United States base at Thule in 1951, although American personnel had little direct contact with, or influence on, the indigenous population.

Permanent settlement is extremely dispersed and entirely confined to the coastal fringe. About three-fifths of the population is urban, a proportion that is growing as a continuation of trends visible after World War II: concentration and industrialization of the fishery industry in the towns, the growth of job opportunities in these towns, and the concomitant decline of village-based fishing and hunting. The capital, Godthåb (Nûk), contains about a fifth of the urban population. Demographically, since mid-century the population has experienced a decline of about two-thirds in its death rate, a consequence of improved health care and nutrition, and a decline of more than half in its birth rate, due to increasing sale and use of contraceptives.

The economy. Seal hunting, once the mainstay of the Greenland economy, declined drastically in the early part of the 20th century, partly as a result of overintensive hunting and partly due to climatic amelioration. By the late 20th century sealing was still the occupation of a small portion of the population, based mainly along the western coast near Ûmának, Upernavik, and Thule, but also along the east coast of Greenland. Fishing, which flourished in the warmer waters, replaced seal hunting as the most important economic activity, and cod fishing, canning, and freezing have played major roles in modernizing Greenland's economy and sustaining population growth. The state owns most fishing plants and several fishing trawlers. There are also a small number of private plants and a large number of private trawlers and smaller fishing boats. The combination of lowered water temperatures and overfishing, however, had reduced landings of cod and other species by the mid-1980s to the point that employment, export earnings, and control of inflation were at risk. These risks, however, were probably short-term, pending determination by marine biologists of sustainable yields for fish stocks that would permit control by quota.

Hunting and fishing

Animal husbandry is marginally possible in the milder southwestern portion of Greenland, where reindeer, first introduced from Norway in 1953, and sheep are raised.

Mining

Mining operations have been erratic: deposits may be rich, but extraction and transportation costs have often been high. The Ivigtut cryolite mine, opened in 1864, closed down its production in 1963, when it proved no longer economically feasible to operate what had been the world's largest natural deposit of this mineral. The coal deposits on Disko Island and the zinc and lead mines at Mesters Vig have been exhausted or are uneconomical to mine. On the other hand, extensive explorations by international companies have discovered considerable potential in terms of uranium, molybdenum, zinc, lead, and other important deposits, and copper deposits were discovered on the east coast of central Greenland. Lead and zinc mining began along the western coast in the late 20th century.

Greenland's trade has continued to fall largely into government hands, in spite of the ending of the official state trading monopoly in 1951. The Royal Greenland Trade Department (Den Kongelige Grønlandske Handel; KGH) continues to supply the country with necessary goods by controlling the import-export business. There is a great excess of imports over exports. The state employs the largest number of Greenlanders—more than one-third of the labour force—and a large contingent of Danes are economically active in the public and private sectors. Greenland's industrial enterprises, like its trade, are either private or government concerns. The latter are run by the KGH or the the Greenland Technical Organization (Grønlands Tekniske Organisation; GTO), which transfers its small factories, workshops, and other enterprises to local, private ownership after setting them up. A good proportion are now in Greenlander, as opposed to Danish, ownership.

Transportation. By the late 20th century the transportation system of Greenland, both internal and external, had undergone a complete transformation linked with the development of a market economy and with technological improvements, notably in air communications. Freight continues to be largely shipped by sea, although access to the ports is seasonally interrupted, while mail and passenger traffic are usually airborne. The airport at Søndre Strømfjord (Kangerlussuaq), in addition to its function as a stopover on transpolar flights to North America and Asia, handles passengers and mail travelling between Greenland and North America and Denmark. Flights to and from Denmark—most often via Iceland—also use airports in East Greenland and in Narssaq in southernmost West Greenland. Helicopters provide internal service between the western towns and the airports at Søndre Strømfjord and Godthåb; a passenger plane has also operated out of Godthåb since 1979.

Traditional dog-drawn sledges remain indispensable to travel in East, North, and northernmost West Greenland. Almost three-fourths of Greenland's motor vehicles and all of its motorcycles are privately owned. There is a high number of vehicles in relation to the island's small road network.

Greenland has a sophisticated telephone, telegram, telex, and radiotelephone network, as well as a military communications network associated with NATO and the North American radar defense system.

Administrative and social conditions. *Government.* On November 29, 1978, the Danish parliament approved legislation granting home rule to Greenland. The legislation was approved by Greenland in a referendum held in January 1979, and it became effective on May 1, 1979. In accordance with home rule, Greenland remains under the Danish crown, which retains jurisdiction over foreign affairs. Each Greenlander is a Danish citizen, enjoying equal rights with all other Danes.

Links with Denmark

Greenland's parliament, called the Landsting, is composed of 21 members elected to four-year terms by all adults 18 years of age or older. The leader of the majority party in the Landsting becomes the chairman; he forms an administration, the Landsstyre, that is composed of himself (as chairman) and four other members of his party from the Landsting. Meetings of the Landsting are scheduled twice each year, each session of about one month's duration; committees are constituted to act as intermediaries between the Landsting and the Landsstyre.

Some matters of government were immediately transferred to the Landsting from Denmark's Ministry for Greenland, while others were to be transferred on a gradual, scheduled basis. The Ministry for Greenland continues to act as a link between the Greenland government and that of Denmark. The two governments are also linked by the two members of the Danish Parliament elected from Greenland by all Danish subjects (Greenlanders and Danes) 18 years of age or older who are resident in Greenland. The local representative of the Danish government in Greenland is the Rigsombudsmand (High Commissioner), whose duties are similar to those of a county prefect.

Greenland is divided into three administrative districts—Nordgrønland (North Greenland), Østgrønland (East Greenland), and Vestgrønland (West Greenland). There are, in addition, a number of municipalities, each with a council elected by adult suffrage every four years and chaired by its locally elected burgomaster.

Justice. The High Court (Landsret) consists of a judge, a graduate of the faculty of law of the University of Copenhagen, and two lay judges. The district courts are headed by lay judges, from whom appeal can be made to the High Court, and, ultimately, to the Supreme Court

in Copenhagen. The police districts are headed by a chief constable.

Education. The educational system is administered by the Director of Education for Greenland, with inspectors for every school district. Preschool and kindergarten facilities are either public or attached to factories. Since 1979, education has been compulsory for the nine years of elementary school, but attendance is voluntary for the next four years—two years in "continuation" school and two in "course" school. There is a teacher-training school in Godthåb, and occupational and vocational training is available at several schools. Special vocational training is available in Denmark, where Greenland's students may also prepare for, and study at, universities and other higher educational institutions. An educational board was established by the Landsstyre in Copenhagen to supervise the education of Greenlanders in Denmark.

Health and welfare. The health service is state supported and without charge, utilizing Danish-trained staff, a doctor, and a small hospital in each district. The central hospital is located in Godthåb. An intensive campaign has nearly eliminated tuberculosis, which was a leading cause of death in the mid-20th century.

The social welfare service is under the jurisdiction of the Landsstyre, which acts in cooperation with the municipal councils. Welfare services are paid for by the Danish government in an annual grant to the Landsstyre. Despite a general improvement in the standard of living in Greenland, the range of social welfare services is wide. A major problem is that of unemployment, which varies from one municipality to another and with the seasons. The Social Welfare Department operates a branch in Copenhagen to care for Greenlanders living in Denmark.

Cultural life. The cultural heritage of the Greenlanders finds most prominent expression in the Eskimo material culture: kayaks, umiaks, sledges, harpoons, and soapstone lamps, all of which have a close relationship to the environment. Cultural values, in the wider sense, arose from the design concepts developed during the construction of these artifacts and found expression in sculptures (some made for the commercial market) in the mediums of ivory, wood, and soapstone. Eskimo legends maintain a relationship to the past and to other Eskimo groups around the Arctic (see above *Arctic cultures*). Modern poetry attempts to revive old traditions but points to a break between Greenlandic and Danish culture, with the Greenlander feeling the danger of losing his cultural identity in the developing market economy and modern European values. A folk high school attempts to support Greenlandic traditions.

Crisis in cultural identity

The press of Greenland dates back to the 1860s, when the newspaper *Atuagagdliutit* (literally "Something to Read") made an appearance in Godthåb. It is still published, in Greenlandic and Danish. There is also an embryonic local press, a product of the movement in the late 20th century to the towns. Broadcasting, which started in World War II, has expanded considerably and provides a wide range of programming.

HISTORY

The history of Greenland is bound up with the history of Arctic exploration. Since the 10th century people have not only settled on the coast but have visited and explored the island in the course of their journeys into the polar regions.

Discovery and exploration. *Exploration of the coasts.* In the beginning of the 10th century the Norwegian Gunnbjorn Ulfsson (son of Ulf Kraka) is reported to have found islands to the west of Iceland. In about 982 Erik the Red sailed from Iceland to find Gunnbjorn's land and spent three years on Greenland's southwestern coasts. On his return in 985 he called the land Greenland in order to make people more willing to go there; and in 986 he started again with 25 ships, of which 14 reached Greenland, where a colony was founded on the southwest coast (see below *Colonization and political developments*). Communication between the Norse settlements and Norway, however, was broken off in the 15th century. In the following century Danish and Norwegian expeditions tried in vain to reestablish the communication, and the rediscovery of Greenland was made by the English navigator Sir Martin Frobisher, who landed on the west coast of Greenland in 1578.

Other explorers, including Gaspar Côrte-Real, had meanwhile seen it, and the work of John Davis (1586–88), Henry Hudson (1610), and William Baffin (1616) afforded further knowledge of the west coast. The east coast was sighted by Hudson at about 73°30′ N in 1607 and in 1617 by the Dutchman Joris Carolus at about 66° N. During the 17th century the coasts were probably visited by many whalers who finished the rediscovery of the country's outline.

The first Danish settlement

Exploration was impossible without bases on the coast and was not started until 1721, when the Danish-Norwegian missionary Hans Egede founded a settlement near Godthåb on the west coast. Egede studied the nature and the people, travelling northward and southward from Godthåb, and in 1752 Peder Olsen Walloe reached a point at 61° N on the east coast.

In the 19th century scientific exploration was accelerated. The southern west coast was mainly explored by Danes from the Danish settlements there, but in the northern region of the west coast English and American explorers were leading. John Ross (1818) found the polar Eskimos at Cape York; E.A. Inglefield (1852) sailed into Smith sound; E.K. Kane (1853–55) worked northward through Smith sound into Kane basin; and C.F. Hall (1871) explored Kennedy strait and Robeson channels to the north of Kane basin. The first to give more accurate information of the east coast was William Scoresby (1822), who made the first fairly trustworthy map of the coast between 69° and 75° N. Captains E. Sabine and D. Clavering visited this coast in 1823 and met the only Eskimo ever seen in this part of Greenland. The German K. Koldewey expedition reached 77° N (Cape Bismarck) in 1870, and the Duke of Orléans penetrated to about 78°16′ N in 1905. The rest of the northern east coast was explored by the Danish L. Mylius-Erichsen–J.P. Koch expedition (1906–08), which discovered Northeast Foreland, the easternmost point. E. Mikkelsen (1909–12) mapped those regions. The southern part of the east coast was first explored by the Dane W.A. Graah (1929–30), and other Danes, G. Holm and T.V. Garde (1883–85) and C. Ryder (1891–92), mapped, respectively, the coast from Cape Farewell to 65°16′ N and the Scoresby Sound. The Dane G. Amdrup (1899–1900) explored the still unknown coast between Angmagssalik and 69°10′ N and the Swede A.G. Nathorst (1899) discovered the large King Oscar Fjord.

Toward the close of the 19th century several explorers visited north Greenland, including L.A. Beaumont of the Nares Expedition (1876), J.B. Lockwood of the Greely Expedition (1882), and R.E. Peary on several journeys (1892, 1895, and 1901). The Danish exploration of north Greenland began in 1910 with the foundation of the station of Thule in North Star bay (at 76°32′ N) by K. Rasmussen. It was the base of five Danish expeditons under Rasmussen and L. Koch.

In 1924–34 expeditions under Rasmussen, Koch, and Mikkelsen continued researches on the east coast, which was also explored by the English–Danish expedition under A. Courtauld and E. Munck (1935–36) and by E. Knuth's expedition in 1938–39. After World War II Koch and Knuth continued their work on the east coast. Great areas were mapped from airplanes.

Exploration of the ice cap. Exploration of the great ice cap, or inland ice, which covers the whole of the interior of Greenland, was attempted in the 18th century but failed and so did E. Whymper's and R. Brown's attempt in 1867. Jens Jensen reached, in 1878, the Jensen nunataks (5,512 feet above the sea) about 45 miles from the western margin at 62°50′ N. A.E. Nordenskiöld penetrated, in 1883, about 80 miles inland at 68°20′ N and two Lapps of his expedition went still farther on skis to about 43° W, at an elevation of 6,600 feet. Peary and C. Maigaard reached, in 1886, about 100 miles inland, a height of 7,500 feet at 69°30′ N. The Norwegian Fridtjof Nansen, with five companions, in 1888 made the first complete crossing of the inland ice, working from east to west,

about 64°23′ N, and reached a height of 8,922 feet. Peary and E. Astrup, in 1892, crossed the northern part of the inland ice between 78° and 82° N and determined the northern termination of the ice covering. Mylius-Erichsen crossed the northeastern corner of the inland ice in 1907, as did Mikkelsen in 1910. In 1912 K. Rasmussen and P. Freuchen crossed from Inglefield Gulf to Danmarkfjord and back and verified that Peary Land is an integral part of Greenland and not, as previously supposed, separated by a strait. In 1912 the Swiss A. de Quervain crossed the inland ice from Disko Bay to Angmagssalik and J.P. Koch and A. Wegener crossed from Louise Land on the northeast coast to Upernavik on the west coast in 1913. The first crossings of the inland ice by airplane were made in 1931 by the German W. von Gronau from Scoresby Sound to Godthåb and by the American Parker Cramer from Holsteinborg to Angmagssalik.

The development of modern air traffic and the fact that Greenland weather is of fundamental importance for predicting conditions on the North Atlantic and in western Europe have meant that since about 1930 weather observations were an important reason for Greenland explorations. In 1930–31 a British expedition under H.G. Watkins made weather observations high on the inland ice 40 miles north of the Arctic Circle; at the same time a German expedition under A. Wegener wintered 300 miles farther north. In 1933 a University of Michigan expedition went up still farther. During World War II Allied military weather stations were at work, and after the war there were weather stations from the south to the far north.

F. Johnstrup first advocated the foundation of a permanent geological survey of Greenland in 1944. Seven years later, in 1951, Grønlands Geologiske Undersogelse became a permanent institution, and since then exploration has been centred on geology, structure, movement of the inland ice, and hydrography. Both Danish and foreign expeditions have undertaken this work. (H.Ln.)

Colonization and political development. *Early settlement.* After discovering Greenland in about 982, Erik the Red settled just north of the present Julianehåb. Soon two colonies had been formed, Eystribygd, in the present district of Julianehåb, and Vestribygd, farther north, in the present district of Godthåb. At the height of their prosperity the colonists numbered about 3,000 on 280 farms. Numerous ruins indicate the location of these colonies. Somewhat later the colonists met the Eskimos farther north in the neighbourhood of Disko Bay, where the Norsemen went to catch seals and walrus. The Eskimos were probably migrating south at that time. Christianity was introduced by Leif (Leifr) Eriksson about 1000, and in 1126 Greenland was assigned its own bishop, who lived at Gardar on Igalikofjord.

Greenland allegiance to Norway

Greenland was a republic until 1261, when the colonists swore allegiance to the king of Norway, who in return charged himself with supplying them with commodities. In the 14th century deterioration in Greenland's climate made it difficult to breed cattle, the colonists' main livelihood. Moreover, the monopolistic trading policy, forced on the kings by the Hanseatic merchants, caused grave trouble to the Greenlanders' trade, and in consequence the colonists diminished in number, and in the 15th century the settlement became extinct. The last vessel from Greenland returned to Norway in 1410, but vessels in the Icelandic fish trade may have visited Greenland until about 1500. Excavations in the Norse burial grounds show 15th-century European influence in the style and texture of the clothes; there is no indication of absorption into the Eskimo groups or of destruction by Eskimo onslaught. Skeletons that show malformation suggest extermination by excessive intermarriage and adverse conditions of life. (B.S.B.)

Recolonization. Norway had been united with Denmark in 1380–81, and in the 16th and 17th centuries the Danish kings several times planned to resume communication with Greenland. The recolonization of Greenland began in 1721 with the voyage of the missionary Hans Egede to Godthåb. He thought that he would find descendants of the Norsemen, but he found only Eskimos. He was deeply disappointed but stayed and started medical and missionary work among them. It was his idea that the economic activities of the colony were to be entirely subordinated to his missionary work. This, however, proved impossible, and a few years later the Danish state had to come to his assistance. Later the Greenland trade was assigned to private interests, but in 1774 it was again taken over by the state and was carried on as a government monopoly until 1951, when Greenland was opened to private Danish enterprise, though the government-owned Royal Greenland Trading company continued its activities. During the period of the government monopoly, Denmark aimed to aid the Greenlanders in cultural respects so as to enable them gradually to establish contact with the outside world without becoming subject to exploitation. Consequently Greenland was shut off from free contact with the outside world, and all resources were reserved for the Greenlanders, who were to sell their surplus production in return for goods required for maintaining and further developing their standard of life.

Establishment of Danish sovereignty. At the dissolution of the union between Norway and Denmark in 1814, Greenland was retained by Denmark. Until 1916 Denmark's sovereignty extended only over the west coast between Cape Farewell and 74°30′ N and the one trading station of Angmagssalik on the east coast, founded in 1894. In 1916, however, the United States declared that it had no objection to the extension of Denmark's political and economical interests to the whole of Greenland. Similar declarations were made by other countries, including Great Britain, and in 1921 Danish sovereignty was extended to embrace the whole island, which led to a dispute with Norway regarding hunting and sealing rights in the uncolonized areas of the east coast. In 1931 some Norwegian hunters, on their own initiative, occupied the east coast between 71° N and 75° N in the name of the Norwegian king, and after some hesitation the Norwegian government recognized the occupation. Denmark at once summoned Norway into the International Court of Justice at The Hague, and in 1933 the occupation was found invalid, the premises stating that Danish sovereignty extended over both the colonized and the uncolonized areas of Greenland. In 1924 a colony had been founded on Scoresby Sound on the east coast, and the privately founded colony of Thule in the far north was taken over by the government in 1937.

Temporary U.S. protectorate

On April 9, 1941, a year after the German occupation of Denmark, Henrik Kauffmann, Danish minister to the United States, signed an agreement that made Greenland a temporary protectorate of the United States. Danish sovereignty was recognized, and the arrangement was to last only for the war emergency; the United States obtained the right to build bases for aircraft and radio and weather stations and to "do any and all things necessary" to hold these positions. New York City displaced Copenhagen as the key point for export and import arrangements, but local Danish officials and the existing Danish government system continued almost unaltered. After World War II the communication between Greenland and Denmark was fully resumed, but U.S. forces remained in some bases. On April 27, 1951, however, an agreement was signed in Copenhagen between Denmark and the United States for the joint defense of Greenland, concluded within the framework of the North Atlantic Treaty Organization (NATO) and replacing the provisional agreement of April 9, 1941. In the areas under U.S. command the United States would enjoy certain rights of use without impairing Danish sovereignty; all defense areas could be used by the ships, aircraft, or armed forces of other NATO countries; U.S. forces in Greenland would respect Danish laws and administration concerning the indigenous population. In accordance with this agreement the United States, in 1951, started the establishment of the great air and radar base at Thule. In 1947 a special committee of Greenlanders and Danes was set up to examine measures to modernize Greenland's political, social, and economic life in accordance with the Greenlanders' desires and the necessity of bringing their way of life to conformity with the new epoch that had come to Greenland as a consequence of modern air strategy and air communications. The examinations

resulted in the abolition of the monopoly of the Royal Greenland Trading Company in 1951 and of the colonial status of Greenland in 1953, when Greenland became an integral part of the kingdom of Denmark. In the following years housing and health services were improved, and harbour and canning facilities increased the possibilities of the fishing trade. (H.Ln.)

Novaya Zemlya

Novaya Zemlya is an Arctic archipelago off the coast of the Russian Soviet Federated Socialist Republic, U.S.S.R. Administered as part of Archangel *oblast*, it consists of two large islands separated by a narrow winding channel 56 miles long, the Matochkin Shar (strait), and many small islands. It lies between latitudes 70°26′ and 77° N and between longitudes 51°26′ and 69°12′ E and forms an elongated crescent, being more than 600 miles long, with a width of 25 to 68 miles and an area of about 31,382 square miles. It separates the Barents Sea on the west from Kara Sea on the east. Along with Vaygach Island, 30 miles to the south, Novaya Zemlya (New Land) forms a continuation from the mainland of the Khrebet Pai-Khao, a branch of the Ural folds.

PHYSICAL GEOGRAPHY

The greatest heights occur near Matochkin Shar (about 3,500 feet). Ice covers about a quarter of the total area, lying mostly in the north island.

A central zone of upper Cambrian and Devonian rocks extends along the islands. These quartzites, conglomerates, and dolomites are flanked by carboniferous shales and limestones. Copper and other metallic ores are known. The coast is severely indented and has raised beaches that give good landing places.

Novaya Zemlya is colder than Spitsbergen (which lies more to the north), as in some degree it shares in the continental conditions of northern Russia and Siberia. The middle and northern parts of the west coast are not so cold as the east. Temperatures at Karmakuly on the west coast (the warmest part) are 2° F (−17° C) in February and 43° F (6° C) in July. Snow is universal from October to May.

Climate

Vegetation is solely tundra and decreases from south to north. It is most luxuriant in the southwest. There are few trees or bushes. The flowering plants number about 200.

In the ice-free areas there are foxes, lemmings, bears, and reindeer. Insects are numerous near the coast. Countless birds come from the south for the breeding season, and at certain parts of the seacoast the rocks are covered with millions of guillemots, while great flocks of ducks, geese, and swans swarm every summer on the valleys and lakes of the south. Whales, walruses, and various seals are frequently seen. The Arctic char and some salmonate species appear in the rivers, and cod frequent offshore waters.

The abundance of sea mammals and birds attracted Russian hunters, and even in the 16th century they had extended their huts to the extreme north of the island. Many of them wintered for years on Novaya Zemlya. Because of the ice in the White Sea, Russian hunters found Novaya Zemlya less accessible than did the Norwegians. But about 1877 systematic attempts at settlement were begun by the Russian government, several families of Nenets (Samoyeds) being established at stations on the west coast of the south island, including Pomorka Bay and Belushya Bay, and Malye Karmakuly on Moller Bay.

HISTORY

Novaya Zemlya was probably known to Novgorod hunters in the 13th century and to Norse hunters earlier still. In 1556 Stephen Borough reached the southern extremity of Novaya Zemlya, being the first western European to do so. Willem Barents touched the island (1594) at Sukhoy Nos (latitude 73°46′) and followed the coast north to the Orange Islands (Ostrova Oranskiye) and south to the Kostin Shar. In 1596, after his discovery of Spitsbergen, Barents wintered at Ice Haven, 76°12′ N. In 1760 Savva Loshkin cruised along the east coast; he spent two winters there and in the next year returned along the west coast, thus accomplishing the first circumnavigation; but the records of his voyage have been lost. In 1768 Lieut. F.F. Rozmyslov explored Matochkin Shar, where he spent the winter. The first scientific information about the island is attributable to the expeditions (1821–24) of F.P. Litke (Lütke; 1797–1882). Nearly all the west coast as far as Cape Nassau, as well as Matochkin Shar, was mapped and valuable scientific information obtained. In 1832 and 1835 Lieut. P.K. Pakhtusov mapped the east coast as far as 74°24′. Karl Ernst von Baer made further investigations in 1837, and A.K. Tsivolka in 1838–39. Expeditions have become less dramatic and more numerous since the middle of the 19th century, particularly since 1920. A weather station has functioned at Karmakuly since 1896; to this was added an observatory at Matochkin Shar (1923), and further weather stations at Cape Zhelaniya (1931), Russkaya Gavan (1932), Capes Stolbovoy and Vykhodnoi (1934), and Cape Menshikova (1953). The settlements in the south island continued to exist, maintained by trapping, raising reindeer, and collecting eiderdown. (T.E.Ar.)

Svalbard

Svalbard is part of the kingdom of Norway and comprises all islands in the Arctic Ocean between longitudes 10° and 35° E and latitudes 74° and 81° N. The main group of islands, known as Spitsbergen, consists of Vestspitsbergen, Nordaustlandet (North East Land), Edgeøya (Edge Island), Barentsøya (Barents Island), Prins Karls Forland (Prince Charles Foreland), and numerous smaller islands. Surrounding the Spitsbergen group from north to south on the east are Kvitøya (White Island, or Gilles Land), Kong Karls Land (King Charles Land, or Wiche Islands), and Hopen (Hope Island). Farther to the south is Bjørnøya (Bear Island). The total area of Svalbard is 23,958 square miles, that of the Spitsbergen group is 23,641 square miles, with Vestspitsbergen 15,075 square miles. Other areas are: Nordaustlandet 5,610 square miles; Edgeøya 1,942 square miles; Barentsøya 514 square miles; Prins Karls Forland 247 square miles; Kong Karls Land 128 square miles; Kvitøya 102 square miles; Bjørnøya 69 square miles; and Hopen 18 square miles. There are no indigenous inhabitants. The population, consisting of miners and administrative staff living at the Norwegian and Soviet mining towns (see below), and some trappers and radio operators, changes seasonally. The administrative centre is Longyearbyen in Vestspitsbergen.

PHYSICAL GEOGRAPHY

The land. *Relief and geology.* More than half of Svalbard is covered by ice. The mountains along the west coast of Vestspitsbergen are rather wild with sharp ridges and peaks formed by folded and metamorphic pre-Devonian sediments. The eastern part of the island and Edgeøya and Barentsøya have plateau-formed mountains built up of nearly horizontal, younger sediments. In the northern part of Vestspitsbergen and Nordaustlandet there are more rounded mountains built up partly of granites and gneisses. The highest peaks are in Ny Friesland, where Newtontoppen reaches 5,633 feet. In the southern part of Vestspitsbergen, Hornsundtind (4,695 feet) is the highest mountain. Many large fjords penetrate the west and north coasts of Vestspitsbergen. The most important are Hornsund, Bellsund with Van Mijenfjorden and Van Keulenfjorden, Isfjorden with several branches, Kongsfjorden and Krossfjorden on the west coast, and Woodfjorden and Wijdefjorden on the north coast. The west and north coasts of Nordaustlandet are also indented by fjords, but the east coast of this island is formed by the front of the inland ice. Many of the glaciers reach the sea, but in Vestspitsbergen there are also large ice-free valleys, such as Sassendalen, Adventdalen, Colesdalen, and Reindalen. In many parts of the coast there are extensive coastal plains, formed by the sea when its level was higher. Bjørnøya consists of an ice-free plain with many lakes, rising to Miseryfjellet (1,759 feet) in the southeast.

Svalbard's mountains

The principal features of the geology are known. Most formations from early Paleozoic, and perhaps Archean, to Recent, occur. The oldest rocks appear chiefly on the west and north, including Prins Karls Forland and

Nordaustlandet. They are the Heclahuken (Hecla Hoek) series of Precambrian, Cambrian, and the Ordovician dolomites, limestones, shales, and quartzites that form the Caledonian folds and overthrusts of the west. The folds can be traced as far east as the west of Nordaustlandet. Granite and gneiss were probably involved in the Caledonian foldings. The metamorphosed Hecla Hoek series and eruptives are in the northwest, uncomformably overlain by Devonian sandstones and shales. These are unconformably followed by Lower Carboniferous sandstones and shales with some coal, Middle Carboniferous or Upper Carboniferous limestones, Permocarboniferous cherts, and Permian sandstones and shales. Next come Triassic, Jurassic, and Cretaceous sandstones, and shales with some small coal seams. Unconformably, Tertiary sandstones and shales with several coal seams follow. Tertiary folding is obvious in the west against the Hecla Hoek beds. In the central part of southern Vestspitsbergen the Tertiary beds form a large syncline. Heavy faulting also occurred in Tertiary times. Intrusions of dolerites and basalt were probably of Cretaceous (Neocomian) date. An extinct volcano and several hot springs (28° C [82° F]) in Bock Bay are Quaternary. The *strandflat* is well developed up to 180 feet; and postglacial raised beaches are marked. Glacial and postglacial debris on low ground generally mask the solid rock, and loose screes form fans on the lower slopes. Many Devonian and later deposits in Svalbard are extremely rich in fossils.

Climate. The sea around Spitsbergen is shallow, and ice readily accumulates around the shores. Pack ice prevents access to most shores except for a few months in the year. The warm North Atlantic drift, however, sends a branch to the western shores of Spitsbergen, moderating its climate and leaving an open passage that permits vessels to approach the western coast during most months of the year. The fjords are frozen from October or November to May or June.

February and March temperature means are about 5° F (−15° C), and July means are about 43° F (6° C). Extreme temperatures may rise to more than 70° F (21° C) in summer and fall below −40° F (−40° C) in winter. In September, autumn sets in. The Arctic night begins in October and lasts until the end of February. The annual precipitation at Grønfjorden (Green Harbour) is 11.6 inches; precipitation is often less in the interior. Winds are generally light except on the west coast. There is mist in the west.

Plant and animal life. The only trees are the polar willow, which does not exceed two inches in height, and the rare dwarf birch (*Betula nana*). The vegetation consists mostly of lichens and mosses, variegated with the golden-yellow flowers of the ranunculus, the large-leaved scurvy grass, the cuckoo flower, many saxifrages, foxtail grass, etc.; while on the driest spots yellow poppies, whitlow grasses, and mountain avens are found—poppies sometimes even on the higher slopes, 2,500 feet or more above the sea. Most species of flowering plants are also found in Europe and many are circumpolar. In the ice-free valleys of Vestspitsbergen there is a comparatively rich vegetation of *Salix polaris,* grasses, moors with cotton grass, and Arctic flowers.

According to the explorer William Scoresby, no fewer than 57,590 whales were killed between 1669 and 1775. Reckless extermination of seals also took place. Walruses are now rarely seen in the water west of Spitsbergen. There is a rich bird life. The fulmars, the glaucous gull or burgomaster, little auks, black guillemots, ivory gulls, and kittiwake gulls breed on the cliffs, while Arctic (Brünnich's) guillemots, puffins, pink-footed brant and barnacle geese, purple sandpipers, red-necked phalarope and other waders, and Arctic terns frequent the tundra and its pools. The commonest passerine bird is the snow bunting. The eider duck breeds on the islands, but its numbers have become noticeably reduced. These birds, however, are only summer visitors, the ptarmigan being the only species that stays permanently. Red char are the only fish found in the rivers and lakes. Land mammals are the polar bear, reindeer, and Arctic fox (both blue and white). There was heavy slaughter of reindeer before they became protected in 1925. The musk-ox, introduced from Greenland in 1929, is also protected and has thrived, unlike the Arctic hare, another import from Greenland.

The economy. Except for coal deposits, ore and mineral deposits are rather poor and scanty. Deposits of anhydrite with gypsum, marble, iron ore, asbestos, galena, zinc blende, and pyrites have been found. Coal deposits were discovered in 1610, but it was not until the beginning of the 20th century that the deposits were thoroughly surveyed. From 1898 to 1920 many areas, most of them containing coal seams, were claimed by companies and persons from various countries. When Norway took over the sovereignty, a Danish commissioner was nominated to decide upon the ownership of these claims. He recognized a total area of 1,631 square miles as private property, of which by the late 20th century the majority belonged to Norway and the remainder to the Soviet Union. All other land is the property of the Norwegian government and may not be transferred into private possession, but the subjects of all signatory powers to the treaty of 1920 have equal rights to exploit mineral deposits. Coal mining on a commercial scale was started by John M. Longyear's Arctic Coal Company (Boston) in Longyeardalen, south of Adventfjorden, in 1906 and by an English company on the north side of this fjord. The U.S. company exported 147,000 tons of coal between 1907 and 1915. In 1916 the property was sold to a Norwegian group that established Store Norske Spitsbergen Kulkompani A/S. A Swedish mine that was worked during 1917–25 at Sveagruva on the north side of Braganzavagen was sold to Store Norske in 1933. A Dutch company worked the mines at Barentsburg on the east side of Grønfjorden until 1926, but sold the property to the Soviet company Arktikugol in 1932. This company also took over the properties at Grumantbyen and Pyramiden. In the late 20th century the Store Norske's in Adventdalen and Arktikugol's at Barentsburg and Pyramiden were still being mined.

Mineral claims

Apart from mining, the only economic activity is trapping and hunting of fox, polar bear, and seal, and increasing activity by U.S., Soviet, and Norwegian oil prospectors.

Besides the coal steamers, a mail and passenger steamer makes trips during the summer from Tromsø to the Vestspitsbergen harbours. In summer Svalbard is also visited by large tourist steamers and many sealers. There are navigation lights and radio beacons in Bellsund-Van Mijenfjorden, Isfjorden, and Kongsfjorden, and Isfjord Radio at Kapp Linné provides a radar service. Svalbard Radio, situated in Longyearbyen, receives transmissions from Norway; the other mining towns have local stations.

Administration. Svalbard is administered by a governor (*sysselmann*), who is also district judge, chief of police, notary public, and revenue officer, and has his residence in Longyearbyen. An inspector of mines (*bergmester*) supervises mining activities. An income tax is payable by Russians as well as by Norwegians. There are primary schools for children.

HISTORY

Exploration. It is probable that the Svalbard (Cold Coast) discovered in 1194, according to Icelandic annals, and mentioned in the *Landnámabók* was Spitsbergen; modern knowledge of Svalbard, however, dates from its discovery by Willem Barents and Jacob van Heemskerck in June 1596. They first discovered Bjørnøya and later saw the northwest coast of Vestspitsbergen. After visits by Stephen Bennett in 1603, Henry Hudson in 1607, and Thomas Marmaduke in 1609, the Muscovy company sent an expedition in 1610, which resulted in the establishment of the whaling industry in 1611, in which English and Dutch whalers took part. Later, French, Hanseatic, Danish, and Norwegian whalers also participated, and the king of Norway-Denmark laid claim to "this part of Greenland," as it was then considered. Quarrels among the whalers resulted in the division of the coast. The English had whaling rights south of latitude 79° N, whereas the Dutch erected their cookeries at Smeerenburg on Amsterdamøya, where more than 1,000 persons might be assembled in summer. The French were driven away from the coast in 1632 and had to boil their blubber on

The whaling industry

board their ships. After about 1640 the whales began to withdraw from the fjords and had to be caught off the coast. English whaling at Spitsbergen ceased in 1660, and Dutch whaling in about 1800, by which time the Greenland whale had been exterminated. English whalers were the first to winter in Spitsbergen, in 1630–31. Probably it was not until about 1715 that the Russians first visited Spitsbergen, wintering there to hunt walrus, seal, bear, fox, and reindeer. Many of them were sent out by the Solovetskiy monastery. Russian activity ceased in about 1850, but already at the end of the 18th century Norwegian sealers from Tromsø, Hammerfest, and other towns had been hunting in the Spitsbergen waters, and from 1822 they also wintered there to trap fur-bearing animals. This wintering was, however, not regular until the last years of the 19th century.

Many expeditions have made Spitsbergen their base for polar exploration, as C.J. Phipps in 1773, D. Buchan and John Franklin in 1818, D.C. Clavering and Edward Sabine in 1823, and William Edward Parry, who in 1827 reached 82°45′ N. In the same year the Norwegian geologist B.M. Keilhau visited the islands, and in 1837 Swedish exploration began with Sven Lovén's expedition. This was followed by many Swedish expeditions, including those of Otto Torell in 1858, A.E. Nordenskiöld in 1864, 1868, and 1872–73, A.G. Nathorst, Gerard De Geer, and others, which were the foundation for later exploration. Other 19th-century exploration included the German expedition under K. Koldewey to the eastern part in 1868, that of Sir Martin Conway, who crossed Vestspitsbergen in 1896–97, and the Swedish–Russian expeditions of 1898–1902. In 1896 the Norwegians built a hotel on Hotellneset at Adventfjorden and started weekly trips by tourist steamers. In the first half of the 20th century Spitsbergen was visited by numerous scientific and exploring expeditions. In 1906 and 1907 the Prince of Monaco sent a party, led by W.S. Bruce, which worked on Prins Karls Forland, and another party, led by Gunnar Isachsen, began mapping the northwestern part of Vestspitsbergen. This led to Isachsen's expeditions of 1909–10, and later to Norwegian expeditions under the leadership of Adolf Hoel, A. Staxrud, and others. After 1948 this work was continued by the Norsk Polarinstitutt, Oslo. From the early 1920s many British expeditions, associated particularly with the Oxford University Exploration Club, visited Spitsbergen, and some wintered there.

Spitsbergen has also been used as a base for polar flight. The first (fatal) attempt was in 1897, when the Swedish scientist S.A. Andrée and two companions set off by balloon from Danskøya; their remains were found on Kvitøya 33 years later. Roald Amundsen started from Ny-Ålesund on his unsuccessful flight to the pole in 1925 and on his flight with Lincoln Ellsworth and Umberto Nobile in the dirigible "Norge" across the pole in 1926. Richard E. Byrd also started from Ny-Ålesund on his transpolar flight in 1926. Sir George Hubert Wilkins landed at the entrance of Isfjorden in April 1928 after his flight from Alaska, and in May Nobile started there in the dirigible "Italia," which met disaster north of Nordaustlandet.

Political development. Despite diverse interest in, and claims to, the islands by British, Dutch, Norwegians, Swedes, Danes, Russians, and Americans, the question of sovereignty was long unsolved, and Spitsbergen was a *terra nullius.* Norway initiated a conference on the matter in Christiania (Oslo) in 1910, which was followed by others in 1912 and 1914, all without result. In 1919 F.H.H. Wedel Jarlsberg persuaded the Allied Supreme Council to grant Norway sovereignty over Spitsbergen, Bear Island, and all the lands lying between 74° and 81° N and between 10° and 35° E. This was put into effect by a treaty of February 9, 1920, signed by Great Britain and the British dominions, the United States, France, Italy, Japan, The Netherlands, Denmark, Norway, and Sweden; the Soviet Union adhered later. Norway took formal possession on August 14, 1925, and declared the region a part of the kingdom of Norway, under the old name Svalbard.

The 1920 treaty forbade the erection of any naval base or fortress within Svalbard. During World War II Spitsbergen was, nevertheless, the scene of serious operations. In August–September 1941 an Allied force destroyed the radio stations, power stations, stocks of coal and oil, etc., and evacuated all Russians to the Soviet Union, and Norwegians to Scotland. In the autumn the Germans erected a meteorological station in Longyearbyen. In May 1942 two vessels, "Isbjørn" and "Selis," were sent from England to prevent the mines from falling into German hands, but on reaching Grønfjorden they were destroyed by German aircraft. A small Norwegian force from England landed in July 1942 and occupied Barentsberg, Longyearbyen, and Sveagruva. After a German meteorological station had been destroyed in Krossfjorden in 1943, a large German force consisting of the "Tirpitz," the "Scharnhorst," and about 10 destroyers entered Isfjorden on September 8, 1943, and totally demolished the mining towns of Barentsburg, Grumantbyen, and Longyearbyen. In 1944 a German submarine set fire to Sveagruva. The Norwegian garrison was, however, maintained throughout the war. After the war all towns were rebuilt. (A.K.O.)

World War II Norwegian garrison

BIBLIOGRAPHY

Physical and human geography: P.J. AMARIA *et al., Arctic Systems* (1977); TERRENCE ARMSTRONG, GEORGE ROGERS, and GRAHAM ROWLEY, *The Circumpolar North: A Political and Economic Geography of the Arctic and Sub-Arctic* (1978); CORA CHENEY, *Crown of the World: A View of the Inner Arctic* (1979); M. DUNBAR and K.R. GREENAWAY, *Arctic Canada from the Air* (1956); S.W. MULLER, *Permafrost: or, Permanently Frozen Ground and Related Engineering Problems* (1947); A.C. O'DELL, *The Scandinavian World* (1957); R.W. RAE, *Climate of the Canadian Arctic Archipelago* (1951); JACK D. IVES and ROGER G. BARRY (eds.), *Arctic and Alpine Environments* (1974); DAVID SUGDEN, *Arctic and Antarctic: A Modern Geographical Synthesis* (1982); J.C.F. TEDROW, *Soils of the Polar Landscapes* (1977); D. ROWLEY (ed.), *Arctic Research* (Special Publication, Arctic Institute of North America, no. 2; 1955); H. WILLIAMS (ed.), *Landscapes of Alaska* (1958); P.D. BAIRD, *The Polar World* (1964); J.B. BIRD, *The Physiography of Arctic Canada* (1967). The periodicals *Polar Record* (3/yr), *Polar Geography and Geology* (quarterly), and *Arctic* (quarterly) contain numerous relevant articles.

T.E. ARMSTRONG, *The Northern Sea Route: Soviet Exploitation of the North East Passage* (1952); L.H. NEATBY, *In Quest of the North West Passage* (1958); N.A. OSTENSO, *Geophysical Investigations of the Arctic Ocean Basin* (1962). AMERICAN ASSOCIATION OF PETROLEUM GEOLOGISTS, *Proceedings of the Second International Symposium on Arctic Geology* (1971); N.A. OSTENSO and R.J. WOLD, "Aeromagnetic Survey of the Arctic Ocean: Techniques and Interpretation," *Mar. Geophys. Res.,* 1:178–219 (1971). NATIONAL ACADEMY OF SCIENCE, *Polar Research: A Survey* (1970); ARCTIC INSTITUTE OF NORTH AMERICA, *Arctic Basin Symposium, 1962* (1963); *The Arctic Basin,* rev. ed. (1969); J.E. SATER, *Arctic Drifting Stations* (1968); Y.Y. GAKKEL and V.D. DIBNAR, "Bottom of the Arctic Ocean," in S.K. RUNCORN *et al.* (eds.), *International Dictionary of Geophysics,* vol. 1, pp. 152–165 (1967); and R.M. DEMENITSKAYA and K.L. HUNKINS, "Shape and Structure of the Arctic Ocean," in A.E. MAXWELL (ed.), *The Sea,* vol. 4, pp. 223–249 (1971). ARCTIC SEA ICE CONFERENCE, *Arctic Sea Ice* (1958); E.R. POUNDER, *The Physics of Ice* (1965); ARCTIC INSTITUTE OF NORTH AMERICA, *Naval Arctic Manual ATP-17 (A), Part 1* (1967). P.D. BAIRD, *The Polar World* (1964); *Arctic,* vol. 22, no. 3 (1969). ALAN COOKE and CLIVE HOLLAND, *The Exploration of Northern Canada: 500 to 1920, A Chronology* (1978); JEAN MALAURIE (ed.), *Arctic Oil and Gas: Problems and Possibilities* (1975); G.H. DENTON and T.J. HUGHES (eds.), *The Last Great Ice Sheet* (1981); YVONNE ROSENBERG-HERMAN, *Marine Geology and Oceanography of the Arctic Sea* (1974).

Arctic cultures: H.B. COLLINS, *Arctic Area* (1954), contains a general archaeological summary and an ethnographic sketch of the American Arctic. For general introductions to the entire Arctic area and to the culture area orientation, see M.G. LEVIN and L.P. POTAPOV (eds.), *The Peoples of Siberia* (1964; originally published in Russian, 1956); M.A. CZAPLICKA, *Aboriginal Siberia* (1914); WALDEMAR JOCHELSON, *Peoples of Asiatic Russia* (1928); WALDEMAR BOGORAS, "Elements of the Culture of the Circumpolar Zone," *A. Rep. Smithson. Instn. 1930,* pp. 465–482 (1931), and "New Data on Types and Distribution of Reindeer Breeding in Northern Eurasia," in *Proc. 23rd Int. Congr. Americanists,* pp. 403–410 (1930); A.L. KROEBER, *Cultural and Natural Areas of Native North America* (1939); and VILHJALMUR STEFANSSON, *The Northward Course of Empire* (1922). RONALD ST.J. MACDONALD (ed.), *The Arctic Frontier* (1966), contains an authoritative symposium. CARLETON S. COON, *The Origin of Races* (1962), and, with EDWARD E. HUNT, *The Living Races of Man* (1965); and E.A. HOOTON, *Up From the Ape,* rev. ed. (1946), are good for information on physical

types in the Arctic in context of all the world's peoples. R.F. SPENCER *et al., The Native Americans: Prehistory and Ethnology of the North American Indians* (1965), includes a summary of Eskimo archaeology on pp. 88–91 and a summary of Eskimo ethnography on pp. 120–154. WALDEMAR BOGORAS, *The Chukchee,* 3 pt. (1904–09, reprinted 1966), and *The Eskimo of Siberia* (1913), are among the many monumental, basic ethnographic monographs on Arctic peoples. ELMAN R. SERVICE, "The Copper Eskimo" and "The Reindeer Tungus of Siberia," in his *Profiles in Ethnology,* rev. ed. (1971), contain some very readable secondary ethnographies of individual Arctic cultures. For aspects of the contemporary Arctic, see TERENCE ARMSTRONG, "The Administration of Northern Peoples: The USSR," and MARGARET LANTIS, "The Administration of Northern Peoples: Canada and Alaska," in RONALD ST.J. MACDONALD (*op.cit.*), pp. 57–119; TERENCE ARMSTRONG, *Russian Settlement in the North* (1965); and DIAMOND JENNESS, *Eskimo Administration,* pt. 1–3, *Tech. Pap. Arct. Inst. N. Am.* (1962–65). Of the ethnological monographs, the following are primary: И. ВЕНЯМИНОВ, *Записки об островах уналашихинскаго отдела,* 3 vol. (1840), still the classic source on the Aleut; KAJ BIRKET-SMITH, *The Chugach Eskimo* (1953); MARGARET LANTIS, *The Social Culture of Nunivak Eskimo* (1946); E.W. NELSON, *The Eskimo About Bering Strait* (1899, reprinted 1971); JOHN MURDOUCH, *Ethnological Results of the Point Barrow Expedition* (1892); R.F. SPENCER, *The North Alaskan Eskimo: A Study in Ecology and Society* (1959, reprinted 1969); NICHOLAS J. GUBSER, *The Nunamiut Eskimos: Hunters of Caribou* (1965); H.B.S. OSTERMANN (ed.), *The Mackenzie Eskimos After Knud Rasmussen's Posthumous Notes* (1942); DIAMOND JENNESS, *The Life of the Copper Eskimos* (1922, reprinted 1970), and *The People of the Twilight* (1928, reprinted 1959); VILHJALMUR STEFANSSON, *The Stefánsson-Anderson Arctic Expedition* (1914); KNUD RASMUSSEN, *Intellectual Culture of the Copper Eskimos: Report of the Fifth Thule Expedition, 1921–24* vol. 9 (Eng. trans. 1932). KAJ BIRKET-SMITH, *The Caribou Eskimos: Report of the Fifth Thule Expedition, 1921–24* (Eng. trans. 1929); KNUD RASMUSSEN, *Observations on the Intellectual Culture of the Caribou Eskimos: Report of the Fifth Thule Expedition, 1921–24,* vol. 7, no. 2 (Eng. trans. 1930), *The Netsilik Eskimos: Social Life and Spiritual Culture: Report of the Fifth Thule Expedition, 1921–24,* vol. 8 (Eng. trans. 1931), and *Intellectual Culture of the Iglulik Eskimos: Report of the Fifth Thule Expedition, 1921–24,* vol. 7 (Eng. trans. 1929); DAVID DAMAS, *Igluligniut Kinship and Local Groupings: A Structural Approach* (1963); FRANZ BOAS, *The Central Eskimo* (1888). WILLIAM THALBITZER (ed.), *The Amassilik Eskimo,* in *Meddelelser om Grøland,* vol. 39–40 (1914–23); KAJ BIRKET-SMITH, *Ethnography of the Egedsminde District with Aspects of the General Culture of West Greenland,* in *Meddelelser om Grønland,* vol. 66 (1924), and *Eskimoerne* (1927; Eng. trans., *The Eskimos,* 2nd ed., 1959, reprinted 1971); E.W. WEYER, *The Eskimos: Their Environment and Folkways* (1932, reprinted 1971). (*Religion*): MARGARET LANTIS, "The Alaskan Whale Cult and Its Affinities," *Am. Anthrop.,* 40:438–464 (1938); "The Religion of the Eskimos," in V.T. FERM (ed.), *Forgotten Religions,* pp. 309–340 (1950). (*Contemporary problems*): C.C. HUGHES, "Under Four Flags: Recent Culture Change Among the Eskimos," *Curr. Anthrop.,* 6:3–54, 64–69 (1965). For a discussion of art, see HENRY B. COLLINS *et al., The Far North: 2000 Years of American Eskimo and Indian Art* (1977). Other cultural studies include ROBERT COLES, *The Last and First Eskimos* (1978); D.E. DUMOND, *The Eskimos and Aleuts* (1977); JEAN MALAURIE, *The Last Kings of Thule: With the Eskimos as They Face Their Destiny* (1982).

Siberian ethnographies include BOGORAS, *op.cit.,* and JOCHELSON, *The Koryak* (1905–08), *The Yukaghir and the Yukaghirized Tungus* (1910–26), and *The Yakut* (1933). An early report from the Samoyeds is KAI DONNER, *Bei den Samojeden in Sibirien* (1926; Eng. trans., *Among the Samoyed in Siberia,* 1954). A work on the Tungus is S.M. SHIROKOGOROFF, *Social Organization of the Northern Tungus* (1933, reprinted 1966); the author, however, is concerned with Tungus of the Manchurian–Chinese–Russian border area. Important reviews of the contemporary situation are TERENCE ARMSTRONG, "The Administration of Northern Peoples: The USSR," in R.ST.J. MACDONALD, *op.cit.;* STEPHEN and ETHEL DUNN, "The Transformation of Economy and Culture in the Soviet North," *Arctic Anthrop.,* 1:1–28 (1963); and ETHEL DUNN, "Educating the Small Peoples of the Soviet North: The Limits of Cultural Change," *ibid.,* 5:1–31 (1968). A great deal of ethnography has been written about the Lapp of Scandinavia. Two general works are BJORN COLLINDER, *The Lapps* (1949), particularly strong on the culture of the Lapps of Sweden; and ORNULV VORREN and ERNST MANKER, *Samekulturen,* 2nd ed. (1958; Eng. trans., *Lapp Life and Customs: A Survey,* 1962), an overview of all the Lapp groups of Scandinavia with particular attention to the various ecologic divisions. The contemporary situation of the Lapp in Finland is outlined in KARL NICKUL, *Report on Lapp Affairs* (1952). HARALD EIDHEIM, *Aspects of the Lappish Situation* (1971), provides a social anthropological analysis of the Lapp minority situation in Norway.

History: For archaeological summaries, see JESSE D. JENNINGS, *Prehistory of North America,* pp. 287–320 (1968); and GORDON R. WILLEY, *An Introduction to American Archaeology,* vol. 1, pp. 69–72 and 410–453 (1966). H.B. COLLINS, *Archaeology of St. Lawrence Island, Alaska* (1937); HELGE LARSEN and FROELICH RAINEY, *Ipiutak and the Arctic Whale Hunting Culture* (1948); and S.I. RUDENKO, *The Ancient Culture of the Bering Sea and the Eskimo Problem* (1961; originally published in Russian, 1947), are basic archaeological monographs.

On exploration, see T.E. ARMSTRONG, *op.cit.;* ANDREW CROFT, *Polar Exploration* (1939); RICHARD CYRIAX, *Sir John Franklin's Last Arctic Expedition* (1939); MOIRA DUNBAR and KEITH R. GREENAWAY, *Arctic Canada from the Air* (1956); A.W. GREELY, *The Polar Regions in the Twentieth Century* (1928); F. MOWAT (ed.), *Ordeal by Ice* (1960); P. FREUCHEN, *Men of the Frozen North* (1962); JEANNETTE MIRSKY, *To the Arctic!: The Story of Northern Exploration from Earliest Times to the Present* (1948); FRIDTJOF NANSEN, *In Northern Mists* (1911); LESLIE H. NEATBY, *In Quest of the North West Passage* (1958); LOUIS SEGAL, *Conquest of the Arctic* (1939); VILHJALMUR STEFANSSON, *The Northward Course of Empire* (1922), and *Unsolved Mysteries of the Arctic* (1939).

The Arctic Islands: CANADIAN HYDROGRAPHIC SERVICE, *Sailing Directions: Arctic Canada,* 3rd ed., 3 vol. (1978–82), a general account of geography, history of exploration, ice characteristics, and climate for navigators; ANDREW TAYLOR, *Geographical Discovery and Exploration in the Queen Elizabeth Islands* (1955), a résumé of geographical discovery and exploration in the northern islands to about 1940; R. THORSTEINSSON and E.T. TOZER, "Geology of the Arctic Archipelago," in the *Geological Survey of Canada, Geology and Economic Minerals of Canada* (1970), a summary account; H.S. BOSTOCK, "Physiographic Subdivisions of Canada," *ibid.;* H.A. THOMPSON, *Climate of the Canadian Arctic,* an augmented reprint from the *Canada Yearbook* (1967). KEITH J. CROWE, *A History of the Original Peoples of Northern Canada* (1974); T. FENGE *et al., Land Use Programs in Canada: Northwest Territories* (1979), provides a historical and contemporary discussion; MORRIS ZASLOW, *A Century of Canada's Arctic 1880–1980* (1981). J.A. DOWNES, "Arctic Insects and Their Environment," *Can. Ent.,* 96:279–307 (1964), a review of several studies; W.E. GODFREY, *The Birds of Canada* (1966); J.D. MCPHAIL and C.C. LINDSEY, *Freshwater Fishes of Northwestern Canada and Alaska* (1970); A.H. MACPHERSON, "The Origin of Diversity in Mammals of the Canadian Arctic Tundra," *Syst. Zool.,* 14:153–173 (1965), with range maps; A.E. PORSILD, *Illustrated Flora of the Canadian Arctic Archipelago,* 2nd ed. (1964), keys, descriptions, illustrations, and range maps; D.B.O. SAVILE, "The Botany of the Northwestern Queen Elizabeth Islands," *Can. J. Bot.,* 39:909–942 (1961), on glacial and postglacial plant geography, and *Arctic Adaptations in Plants* (Canada Department of Agriculture, 1972).

Greenland: DANISH COMMISSION FOR THE DIRECTION OF THE GEOLOGICAL AND GEOGRAPHICAL INVESTIGATIONS IN GREENLAND, *Meddelelser om Grønland,* 189 vol., containing papers in English, German, Danish, French, and Greenland Eskimo (1879–), is the largest existing work about Greenland, written by prominent specialists, and mostly covering geological, geographical, and ethnographic subjects. *Greenland,* 3 vol. (1928), a useful source of information, notably in matters concerning the physical environment and ethnography; KAJ BIRKET-SMITH, *The Eskimos,* 2nd ed. (Eng. trans. 1959, reprinted 1971), the most authoritative work on the Eskimos and their culture; PETER FREUCHEN, *Book of Eskimos* (Eng. trans. 1961), an entertaining account by a famous Danish explorer; VILHJALMUR STEFANSSON, *Greenland* (1942), a readable work on Eskimo culture by a famous American explorer, now mainly of historic interest; SVEND KLITGAARD, *Greenland* (Eng. trans. 1970), an extensive volume covering all aspects of Greenland life. For history, see POUL NORLUND, *De gamle Nordbobygder ved Verdens Ende,* 2nd ed. (1936; trans. by W.E. CALVERT, *Viking Settlers in Greenland and Their Descendants During Five Hundred Years,* 1936); FINN GADD, *The History of Greenland,* 3 vol. (1971–82; originally published in Danish, 1967–75); and C.W. SCHULTZ-LORENTZEN, *Greenland,* vol. 2, *The Intellectual Culture of the Greenlanders* (1942). Current information may be found in the *Danish Foreign Office Journal,* and in publications of the Danish Ministry of Foreign Affairs. HUBERT J.C. SCHUURMAN, *Canada's Eastern Neighbor: A View on Change in Greenland* (1976), a discussion covering the time since World War II; IVARS SILIS, *Greenland Today* (1982), a series of vignettes on contemporary Greenlandic life; ARNE GAARN BAK, *Atlashandbog over Grønland* (1978), an atlas with city and town maps, relief maps, and place-name index.

(J.B.Bd./M.J.Du./N.A.O./W.K.C./R.F.S./R.P.B.P./
M.Dr./R.L.Ch./D.B.O.S./M.Ga./B.S.B.)

Argentina

Argentina (República Argentina) occupies most of the southern portion of South America. It is the eighth largest country in the world, with an area of 1,073,399 square miles (2,780,092 square kilometres). Shaped like an inverted triangle with its base at the top, it is only about 884 miles (1,423 kilometres) across at its widest from east to west, but it stretches 2,360 miles from the subtropical north to the subantarctic south. This great length embraces regions of striking diversity, including the Andes Mountains; the thorny scrubland and seasonal swamps of the Gran Chaco; the broad, fertile plains of the Pampa; the stark tableland of Patagonia; and an undulating Atlantic coastline of 2,936 miles. Argentina also claims a 49° wedge of Antarctica and several islands in the South Atlantic, including the British-ruled Falklands (Islas Malvinas). It is bounded by Chile on the south and west, Bolivia and Paraguay on the north, and Brazil, Uruguay, and the Atlantic Ocean on the east.

For many foreigners, especially Europeans, Argentina has presented the traditional New World image of a land of romance and opportunity. It received its name, roughly translated as "land of silver" or "silvery one," from Spanish explorers of the 16th century who were lured there by rumours that portrayed the presence of vast mineral wealth. In the 19th century the former colony of Spain was the land of gauchos, the lone horsemen of the pampas, and of estancieros, ranchers who lived like kings on estancias the size of small countries. In the last part of the 19th century and the first quarter of the 20th, Argentina became for the poor of Europe a place where they could earn a decent living on the expansive farmland of the interior or in the growing cities of the coast. During this period millions of immigrants came to Argentina, bringing skills that helped transform it into a modern country whose agriculture and industry remain among the most productive of Latin America.

Historically, Argentina has often been ruled by a caudillo figure, a strong leader, often of the military, who dominated the nation, usually until he was deposed or died. During its periods of democratic rule Argentina has been administered as a federation of autonomous states with a republican system of government. In the late 20th century Argentina has been set back by failed policies that have at times brought hyperinflation and led the country into misadventures such as the Falkland Islands war of the early 1980s, a venture that cost dearly in terms of both morale and finances. Its recovery from such disappointments and its future development are based on the potential manifested in its excellent resources and its well-educated populace.

This article is divided into the following parts:

Physical and human geography

THE LAND

Major landform regions

Argentina encompasses a variety of major landforms that are often grouped together into four major regions: the Andes, the North, the Pampa, and Patagonia. The Andean region extends some 2,300 miles along the western edge of the country, from Bolivia to southern Patagonia, forming most of the natural boundary with Chile. It is commonly subdivided into two parts: the Northwest and the Patagonian Andes, the latter of which is discussed here under Patagonia. The North is commonly described in terms of its two main divisions: the Gran Chaco, or Chaco, comprising the dry lowlands between the Andes and the Paraná River, and Mesopotamia, an area between the Paraná and Uruguay rivers. The centrally located plains, or Pampa, are grasslands subdivided into arid western and more humid eastern parts called, respectively, the Dry Pampa and the Humid Pampa. Patagonia is the cold, parched, windy region that extends some 1,200 miles south of the Pampa, from the Colorado River to Tierra del Fuego.

Relief. *The Northwest.* This part of the Andes region includes the northern half of the main mountain mass in Argentina and the transitional terrain, or piedmont, merging with the eastern lowlands. The Andean system of north–south-trending mountain ranges, 16,000 to 22,000 feet (4,900 to 6,700 metres) high and separated by flat-floored structural valleys ranging from 10,000 to 13,400 feet, gradually decreases in size and elevation southward from Bolivia to the upper Colorado River, which forms

the southern border of the region. South America's highest mountain, Aconcagua (22,831 feet [6,959 metres]), lies in the Northwest, together with a number of other peaks that reach over 21,000 feet. Some of these mountains are volcanic. In the northern part of the Argentine Andes are the high undissected structural valleys referred to as puna.

To the south and east, where the parallel to subparallel ranges become lower and form isolated, compact north–south-trending units, the flat valleys between are called *bolsones* (basins). This southeastern section of the Northwest is often referred to as the Pampean Sierras, a complex that has been compared to the Basin and Range region of the Southwestern United States. They are characterized by west-facing escarpments and gentler east-facing backslopes such as are found in the spectacular Sierra de Córdoba at the eastern edge of the Pampean Sierras. The Pampean Sierras have variable elevations; ranges begin at 2,300 feet in the Sierra de Mogotes in the east and rise to 20,500 feet in the Sierra de Famatina in the west.

The Pampean Sierras

The Gran Chaco. The western sector of the North region, the Gran Chaco, extends beyond the international border at the Pilcomayo River into Paraguay, where it is called the Chaco Boreal (North Chaco) by Argentines. The Argentine sector between the Pilcomayo River and the Bermejo River is known as the Chaco Central. Southward to the 30th parallel, where the Pampa begins, Argentines have named the area the Chaco Austral (South Chaco). The Gran Chaco in Argentina, though descending in flat steps from west to east, is poorly drained and has such a difficult combination of physical conditions that it remains one of the least inhabited parts of the nation. Its subtropical climate is characterized by some of Latin America's hottest weather; it is largely covered by a thorny vegetation; and it is subject to summer flooding.

Mesopotamia. East of the Chaco lies a narrow depression, 60 to 180 miles wide, called Mesopotamia, which is bordered on the north by the Brazilian Shield. The narrow lowland stretches for 1,000 miles southward, finally merging with the Pampa south of the Río de la Plata. Its designation as Mesopotamia (between the rivers) reflects the fact that its western and eastern borders are large rivers, namely the Paraná and Uruguay. In the south there are small hills, around which the rivers flow. The northeastern part of this region, Misiones Province, between the Upper Paraná and the Uruguay rivers, is higher in elevation.

The Pampa. Pampa is a Quechua Indian term meaning "flat surface." As such it has a broad applicability in southeastern South America beginning in Uruguay where grass-covered plains commence south of the Brazilian massif. In Argentina the plains broaden out west of the Río de la Plata to meet the Andean forelands, blending imperceptibly on the north with the Chaco Austral and southern Mesopotamia and extending southward to the Colorado River. The eastern boundary is the Atlantic coastline.

The flat plains

The main landmass of the Pampa is flat. Thick deposits of silt interrupted only by occasional caps of alluvium and volcanic ash rest on its surface. In the area of the southern Pampa, the landscape rises gradually. There the morphology is changed by the emergence of sierras formed from old sediments and crystalline rocks.

Patagonia. This region consists of an Andean zone (also called Western Patagonia) and the main Patagonian plateau south of the Pampa, which extends to the tip of South America. The surface of Patagonia descends east of the Andes in a series of broad, flat steps extending to the Atlantic coast. Evidently, tectonic activity related to the formation of the Andes created the gigantic platforms and coastal terraces. The coastline is cliffed along its entire length as a result. The northern cliffs are rather low but tend to rise in the south, where they reach heights over 150 feet. The landscape is cut by eastward-flowing rivers—some of them of glacial origin in the Andes—that have created broad valleys and steep-walled canyons.

Other features of Patagonia include a series of basins, some of which contain lakes, nestled between the Patagonian Andes and the plateau, and volcanic hills in the central plateau west of the city of Río Gallegos. These hills and accompanying lava fields are characterized by dark colours spotted with lighter-coloured bunchgrass, which creates a leopard-skin effect that intensifies the desolate, windswept appearance of the Patagonian landscape. A peculiar type of rounded gravel called *grava patagónica* lies on level landforms, including isolated mesas. Glacial ice in the past extended beyond the Andes only in the extreme south, where there are large moraines.

Leopard-skin coloration

Drainage. The largest river basin in Argentina is that of the Río de la Plata (often called the River Plate in English). It drains an area of 1,600,000 square miles, which includes, in addition to northern Argentina, the whole of Paraguay, eastern Bolivia, most of Uruguay, and a large part of Brazil. The river is actually an estuary formed by the confluence of the Paraná and Uruguay rivers; its name, meaning "River of Silver," was given it in colonial times before explorers found that there was neither a single river nor silver upstream from its mouth. Other tributaries of this system are the Iguazú (or Iguaçu), Paraguay, Pilcomayo, Bermejo, Salado, and Carcarañá. Before the Iguazú River joins the Paraná at the point where the borders of Argentina, Brazil, and Paraguay meet, it plunges over the escarpment of the Brazilian massif, creating the Iguazú Falls—one of the world's most spectacular natural attractions.

The "River of Silver"

Because most of Argentina is flat or only gently sloping from the Andes to the Atlantic, the country's river drainage is poor. Grand rivers flow across the Gran Chaco flatlands, but their shallow nature rarely permits navigation, and never with regularity. Moreover, long-lasting summer floods cover vast areas and leave behind ephemeral swamplands that gradually dry out in winter. During winter, most rivers of the Gran Chaco dry up, the air chills, and the land seems visibly to shrink. Only three of the region's numerous rivers (the Pilcomayo, Bermejo, and Salado) manage to flow from the Andes to the Paraguay–Paraná system in the east without dying somewhere en route and forming salt pans (salinas) due to evaporation. The region's largest rivers follow a veritable maze of courses during flood season, however.

In the Northwest the Desaguadero River, the upper course of the Salado, and its tributaries in the Andes Mountains water the sandy deserts of Mendoza Province. The principal tributaries are the Jáchal, Zanjón, San Juan, Mendoza, Tunuyán, Diamante, and Atuel. In the northern Pampa, Lake Mar Chiquita, the largest lake in Argentina, receives the waters of the Dulce, Primero, and Segundo rivers but has no outlet. Its name, meaning "Little Sea" Lake, refers to the high salt content of its waters. Rivers that cross Patagonia from west to east diminish in volume as they travel through the arid land. The Colorado and Negro rivers produce major floods after seasonal snow and ice melt in the Andes. In the south, the Santa Cruz River is a notable stream that flows eastward out of the glacial Lake Argentino in the Andean foothills to the Atlantic.

Soils. Soil types in Argentina range from the light-coloured saline formations of the high puna to the dark, nutritive, and friable type found in the Pampa. Golden-brown loess soils of the Gran Chaco, sometimes lighter where salinity is excessive, turn darker toward the east in the Mesopotamian border zone where poor drainage occurs. These give way to soils ranging from rust to deep red colorations in Misiones. South of the Pampa, grass-covered dark soils begin to turn brown as the drier parts of northern Patagonia are reached. Light-tan arid soils of varying texture cover the rest of this region. Grayish podzolic types and dark-brown forest soils characterize the Andean slopes.

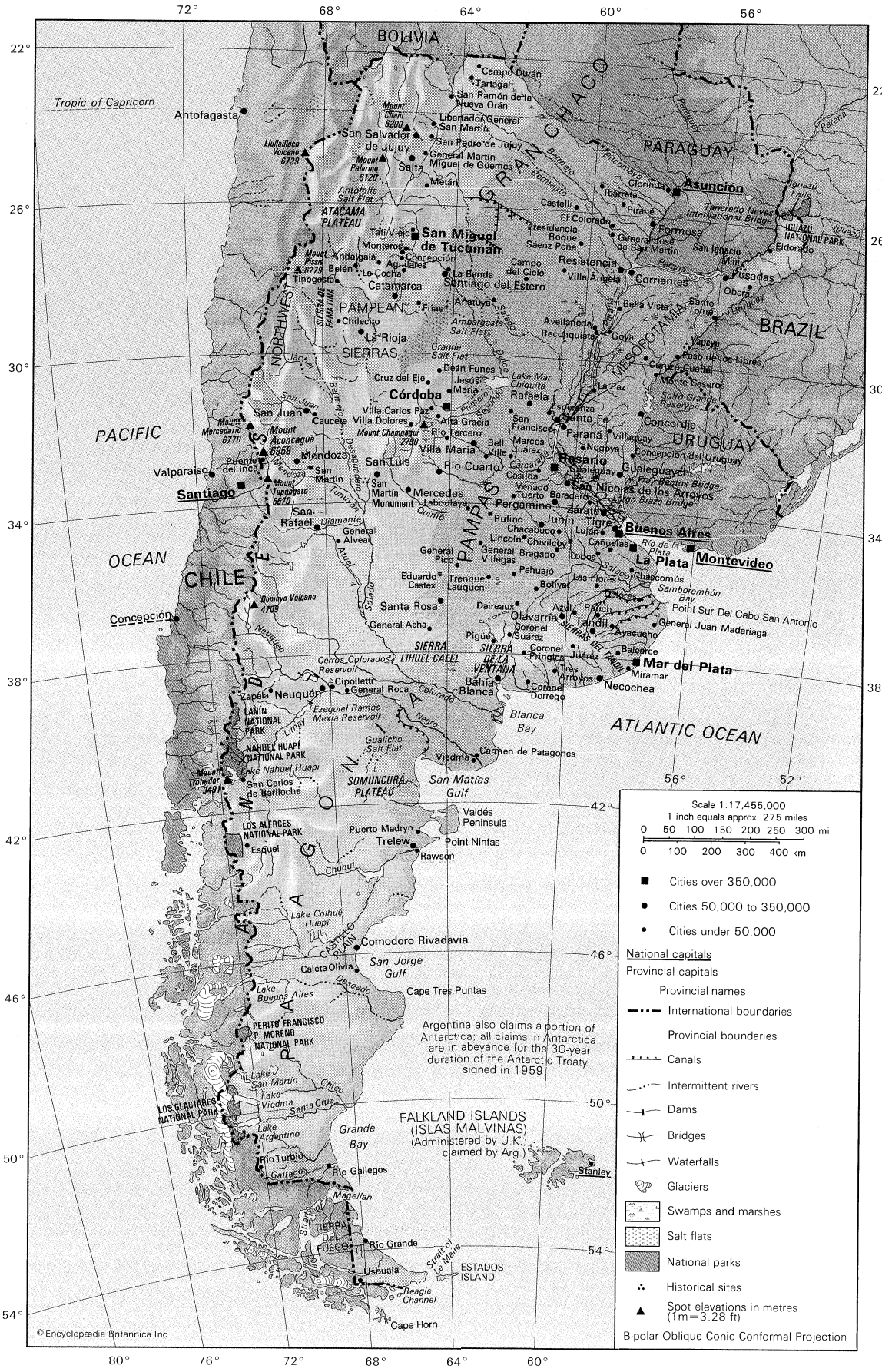
Climate. Argentina lies almost entirely within the temperate zone of the Southern Hemisphere, unlike the rest of the continent to the north, which lies within the tropical zone. Only occasional tropical conditions invade the provinces of Formosa and Misiones in the extreme north. On the other hand, a landmass extending to 55° south latitude might be expected to possess a continental climate with at least one month having an average temperature below freezing, similar to comparable latitudes in North America. The continent narrows so rapidly toward its southern tip, however, that the moderating effect of both the Pacific and Atlantic oceans influences a predominantly temperate climate despite the high latitudes. The

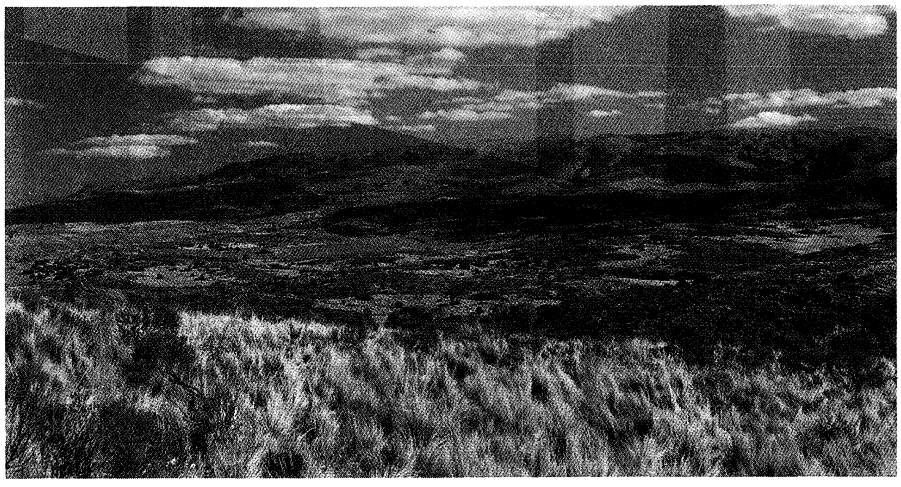
The temperate climate

MAP INDEX

temperate climate is interrupted by a long, narrow north–south band of semiarid to arid conditions, which makes Argentina the only place in the Southern Hemisphere with an extensive portion of arid eastern coastline. Zones of tundra climate and polar climate predominate in the high Andes and in southern portions of Tierra del Fuego. The temperate regions of the nation have an average cold month temperature from 50° to 65° F (10° to 18° C).

Precipitation is moderate to light throughout most of the country, with the driest areas in the far northwest and


BOLIVIA
PARAGUAY
BRAZIL
URUGUAY
CHILE
PACIFIC OCEAN
ATLANTIC OCEAN
Tropic of Capricorn
Antofagasta
Campo Durán
Tartagal
San Ramón de la Nueva Orán
Libertador General San Martín
San Salvador de Jujuy
San Pedro de Jujuy
General Martín Miguel de Güemes
Salta
Metán
Mount Cerro Cabi 6200
Llullaillaco Volcano 6739
Mount Palermo 6120
Antofalla Salt Flat
ATACAMA PLATEAU
GRAN CHACO
Pilcomayo
Bermejo
Falcondra
Paraguay
Paraná
Clorinda
Asunción
Ibarreta
Pirané
Castelli
El Colorado
Formosa
Tancredo Neves International Bridge
Iguazú Falls
IGUAZÚ NATIONAL PARK
Eldorado
Posadas
Oberá
Santo Tomé
Uruguay
Tafí Viejo
San Miguel de Tucumán
Monteros
Concepción
Aguilares
Famaillá
Presidencia Roque Sáenz Peña
General José de San Martín
San Ignacio
Mini
Resistencia
Campo del Cielo
Villa Angela
Corrientes
Mount Missa 6779
Andalgalá
Belén
La Cocha
Tinogasta
La Banda
Santiago del Estero
Catamarca
Añatuya
Frías
Bella Vista
Avellaneda
Reconquista
Goya
Vapeyú
Paso de los Libres
Carlos Casado
Monte Caseros
Ambargasta Salt Flats
Salado
Dulce
Grande Salt Flat
Chilecito
La Rioja
SIERRAS PAMPEAN
SIERRAS MARIANA
NORTHWEST
Jáchal
Bermejo
San Juan
Cruz del Eje
Déan Funes
Jesús María
Córdoba
Lake Mar Chiquita
Primero
Segundo
Rafaela
La Paz
Esperanza
Santa Fe
Concordia
Salto Grande Reservoir
Mount Mercedario 6770
Mount Aconcagua 6959
San Juan
Caucete
Villa Carlos Paz
Villa Dolores
Alta Gracia
San Francisco
Río Tercero
Mount Champaqui 2790
Villa María
Marcos Juárez
Bell Ville
Paraná
Nogoyá
Villaguay
Concepción del Uruguay
Puente del Inca
Mendoza
San Luis
Río Cuarto
Carcarañá
Rosario
Gualeguaychú
Prof. Santos Bridge
Valparaíso
Santiago
Mount Tupungato 6570
San Martín
Mercedes
Monument
Laboulaye
Casilda
Venado Tuerto
Baradero
San Nicolás de los Arroyos
Largo Brazo Bridge
San Rafael
Diamante
Quinto
Pergamino
Rufino
Zárate
Junín
Tigre
Chacabuco
Lincoln
Chivilcoy
Lujan
Buenos Aires
Río de la Plata
General Alvear
Anguil
General Pico
General Villegas
General Bragado
Cañuelas
La Plata
Montevideo
Eduardo Castex
Trenque Lauquen
Pehuajó
Lobos
Las Flores
Chascomús
Samborombón Bay
Point Sur Del Cabo San Antonio
Bolívar
Dolores
Domuyo Volcano 4709
Concepción
Santa Rosa
Daireaux
Olavarría
Azul
Rauch
Tandil
Ayacucho
General Juan Madariaga
General Acha
Pigüé
Coronel Suárez
SIERRA DE LA VENTANA
SIERRAS DEL TANDIL
Coronel Pringles
Juárez
Balcarce
Mar del Plata
Miramar
Tres Arroyos
Necochea
Cerros Colorados Reservoir
LIHUEL-CALEL
Bahía Blanca
Coronel Dorrego
Neuquén
Zapala
Cipolletti
General Roca
Colorado
Blanca Bay
Ezequiel Ramos Mexía Reservoir
LANÍN NATIONAL PARK
Limay
Negro
Gualicho Salt Flat
Viedma
Carmen de Patagones
NAHUEL HUAPÍ NATIONAL PARK
Lake Nahuel Huapi
Mount Tronador 3491
San Carlos de Bariloche
SOMUNCURÁ PLATEAU
San Matías Gulf
Valdés Peninsula
LOS ALERCES NATIONAL PARK
Esquel
Puerto Madryn
Trelew
Point Ninfas
Rawson
Chubut
Lake Colhué Huapi
CASTILLO PLAIN
Comodoro Rivadavia
San Jorge Gulf
Caleta Olivia
Deseado
Cape Tres Puntas
Lake Buenos Aires
PERITO FRANCISCO P. MORENO NATIONAL PARK
Lake San Martín
Lake Viedma
Chico
Santa Cruz
LOS GLACIARES NATIONAL PARK
Lake Argentino
Grande Bay
Río Turbio
Río Gallegos
Magellan
Strait of
TIERRA DEL FUEGO
Río Grande
Ushuaia
Strait of Le Maire
ESTADOS ISLAND
Beagle Channel
Cape Horn
Argentina also claims a portion of Antarctica; all claims in Antarctica are in abeyance for the 30-year duration of the Antarctic Treaty signed in 1959.
FALKLAND ISLANDS (ISLAS MALVINAS)
(Administered by U.K.; claimed by Arg.)
Stanley
Scale 1:17,455,000
1 inch equals approx. 275 miles
0 50 100 150 200 250 300 mi
0 100 200 300 400 km
Cities over 350,000
Cities 50,000 to 350,000
Cities under 50,000
National capitals
Provincial capitals
Provincial names
International boundaries
Provincial boundaries
Canals
Intermittent rivers
Dams
Bridges
Waterfalls
Glaciers
Swamps and marshes
Salt flats
National parks
Historical sites
Spot elevations in metres (1m=3.28 ft)
Bipolar Oblique Conic Conformal Projection
© Encyclopædia Britannica Inc.



Sierra de Córdoba, facing east, overlooking an escarpment in the Pampa.
Chip and Rosa Maria Peterson

in the southern part of Patagonia. Most rainfall occurs in the northeast, in the Humid Pampa, Mesopotamia, and the eastern Chaco. Thunderstorms, called *pamperos,* with lightning and hail are common. Weak fronts create another mechanism that produces rain, often bringing long periods of winter moisture. Dull, gray days and damp weather characterize this season, especially in the Pampa. Between winter storms, tropical air masses make incursions southward and bring mild relief from the damp cold.

The *pamperos*

In the Andean Northwest region, there is in some places an annual temperature range of more than 36° F (20° C), and occasional continental climatic conditions occur. Winter temperatures sometimes fall below freezing on cloudless days and nights.

Tundra climate

The high-elevation, cold climatic phenomenon in Argentina is sometimes referred to as tundra climate and, in even colder mountaintop areas, as polar. Generally, the tundra climate occupies the mountain zones where average annual temperatures are below 50° F (10° C); in the north this occurs above 11,500 feet. Moving southward, tundra climate occurs at gradually decreasing elevations until it reaches sea level in southern Tierra del Fuego. The highest Andean peaks may have permanent snow or ice cover.

A longitudinal rain-shadow zone (created when winds lose their moisture in passing over high mountains) on the east side of the Andes causes an arid to semiarid climate in the interior of Argentina. The zone begins in the Andean Northwest and extends along the eastern slopes of the Andes southward to, but not including, Tierra del Fuego. The rain-shadow area has a central core of arid or desert climate rimmed by semiarid, or steppe, conditions. The steppe areas have about twice the annual precipitation found in the arid zones, but evaporation exceeds precipitation in both zones, which therefore remain treeless. Most of the arid region is subjected to strong winds that carry abrasive sand and dust. This is particularly true in Patagonia, where the windblown dust creates a continuous haze that considerably reduces visibility.

The Pampa is a transitional area between high summer temperatures to the north and cooler summers to the south. Buenos Aires, located on the north edge of the Pampa, has a climate similar to cities in the southeastern United States, with hot, humid summers and cool, mild winters. Mean temperatures for summer months (December to February) range from about 72° to 75° F (22° to 24° C) and for winter months (June to August) about 46° to 55° F (8° to 13° C). In the Humid Pampa the rainfall varies from 39 inches (990 millimetres) in the east to 20 inches in areas near the Andes—about the minimum for nonirrigated crops. Cold fronts that move northward from Patagonia, chiefly in July, bring occasional frosts and snow to the Pampa and Mesopotamia. On rare occasions a dusting of snow covers Buenos Aires itself.

Plant and animal life. Argentina's fauna and flora vary widely from the country's mountainous zones to its dry and humid plains and its subpolar regions. In heavily settled regions, the natural animal and plant life have been profoundly modified by man.

Northwest. Vegetation in the Northwest region includes the high puna desert, with mostly exposed soil and bunchgrass, the forested slopes of the Andes, and the subtropical scrub forests of the Pampean Sierras, which merge with the deciduous scrub woodland in the Gran Chaco. The vegetation of the puna consists of dwarf shrubs and tough grasses; these and other plants in the region are almost the brown colour of the ground itself. The region is the land of the guanaco and its near relatives, the llama, alpaca, and vicuña.

Forests continue along the eastern border of the puna region southward to the colder Andean zones so that they cover many slopes in this part of the mountains. Above 1,650 feet, the so-called mistol forest thrives. There is a subtropical rain forest at elevations of about 4,000 feet, composed of laurels, cedars, and other species. Above 3,300 feet, the giant cedars and some other tree species disappear. The tree heights are less impressive, and the growth becomes more of a cloud forest type; myrtles and poorly developed laurels predominate, and pine trees reach above 7,000 feet. Above this forest grow the *queñoa,* small crooked trees that in places extend to the tree line at 11,500 feet.

Southeast of the Andean region described above, xerophytic scrub forests, called *monte,* and intervening grasslands spread across the Pampean Sierras. Vegetation includes mimosas and acacias, and there is a smattering of cactus. Hares, skunks, and small deer abound in this part of the Northwest.

The thorny Gran Chaco

Gran Chaco. The western Gran Chaco has growths of thorn forest dominated by algarroba in the drier and often saline zones. Quebracho is present, but not to the extent that it is farther east. In areas with finer salt at the surface, even halophytic plants, which normally grow in salty soils, do not survive. Coarse bunchgrasses are common in the dry steppe. Overall, dense scrub forests intermixed with prickly pear, barrel, and many other types of cactus and trees blend with the steppe areas.

The vegetation of the Chaco becomes increasingly lush as it spreads to the east. The thorn forest gradually becomes dominated by dense forest with another, less valuable quebracho species; there are some pure stands of algarroba. Some 90 miles west of the Paraná River, some of the giant trees that are also found in the much more humid Misiones Province begin to appear. The rich wildlife of the Chaco includes deer, peccaries, monkeys, tapir, jaguars, pumas, ocelots, armadillos, capybaras, and agoutis. The vast birdlife includes a refuge for the rhea; streams harbour numerous fish species, including the piranha; and snakes and reptiles abound.

Mesopotamia. In Mesopotamia, thin stands of tall wax palms occupy the flood zones. Groups of trees and grassy areas form a park landscape of noted beauty. The quebra-

cho from which tannin has been extracted since colonial times, *urunday,* and *guayacán,* used for tannin and lumber, are common. Gallery forests growing along rivers become denser and taller in Misiones Province. At higher elevations, the Paraná pine appears. Mesopotamia is the habitat for jaguars, monkeys, deer, tapir, peccaries, many snake varieties, and numerous birds, such as the toucan and hummingbird, as well as stingless honeybees.

Pampa. The principal Pampa vegetation is *monte* in the Dry Pampa and grassland in the Humid Pampa. The north–south boundary between the Dry and Humid pampas lies approximately along the 64th meridian. In the most humid areas, knee-high grasses appear. To the north, west, and south, where precipitation decreases, tougher grasses give way to the *monte* of the Dry Pampa. Planted grasses and trees have replaced much of the original flora. Pampa wildlife includes the rhea and the guanaco. Both animals are fleet-footed, which is probably the reason for development of the bola, a device used by the Indians to trip the animals. Small deer, introduced hares, and viscacha, a burrowing rodent, are common.

Monte vegetation of the Dry Pampa

Patagonia. Patagonian vegetation consists of deciduous Andean forests and the steppe and desert zones east of the Andes. The largest area—the steppe region—lies in northern Patagonia between the Colorado River and the port city of Comodoro Rivadavia. This zone represents an extension southward of the *monte,* which gives way gradually to a xerophytic shrub region without trees except along river banks. In the extreme west on the Andean border, small stands of araucaria survive, and clumps of wiry grasses are also present. South of Comodoro Rivadavia to the tip of the continent, low scrub vegetation and green grass steppe alternate. Wildlife in the region includes the now rare guanaco and rhea, as well as eagles and herons, burrowing rodents and the Patagonian hare, mountain cats and pumas, and various poisonous reptiles.

In Tierra del Fuego, it appears that grasses first covered glaciated zones, but after volcanic ash settled there, forests advanced. Antarctic beech colonized rapidly in valleys and grows along with cypress on steep slopes. A phenomenon in the southern tip of the continent is the existence of species of parrots and canaries, both of which are more associated with the tropics than with Patagonia.

The Patagonian Andes have coniferous and broad-leaved forests that spread into Chile. Antarctic beech and needle-leaf trees mixed with araucaria are common. The Patagonian Andes do not have a flourishing animal life. The smallest known deer, the pudu, dwells there, and wild pigs, introduced by Europeans, have multiplied.

Settlement patterns. The varied topography, climate, and natural resources of Argentina shaped the pattern of settlement in its major regions from the beginning of colonization. Although modern transportation and industry have partly effaced regional differences, the organization of life in both city and country still follows patterns that were set in the years of the first European settlement.

The Northwest. Numerous archaeological sites in the region indicate the presence—before the Spanish invasion—of permanently settled Indians with irrigated and terraced farming in the oasis-like valleys. The Spanish, coming from Peru, first established Santiago del Estero in 1553, eastward on the lowland plain in the Chaco away from belligerent natives. Not long afterward, forts arose in the Northwest at Tucumán, Salta, Jujuy, and San Luis; Córdoba, to the south, was founded in 1573. Meanwhile, the Northwest received colonists still farther south from Chile. Chileans founded the cities of Mendoza and San Juan in the early 1560s.

Early settlement

The cities in the Northwest were founded originally as centres for agriculture and livestock raising and to serve as support for the silver mines of the Viceroyalty of Peru, particularly at Potosí (now in Bolivia). Later, as Buenos Aires developed and the mines became less profitable, the country's orientation switched to the southeast. The Spanish established a trade route between Chile and Buenos Aires that went through Córdoba and Mendoza, both of which thrived. This northward path was chosen because of the need to avoid the Pampa Indians, and it remains an important transportation route. In the 600-mile-long rain-shadow zone east of the Andes, beginning just south of San Miguel de Tucumán, settlement took place in river oases stretching to San Rafael, south of Mendoza.

Rail transportation, which tied Mendoza to the Pampa in 1885, brought about an emphasis on viticulture in the Mendoza region. With access to Buenos Aires came new capital, more settlers, better grape stock, and larger markets. Once the problem of a labour shortage was solved with European immigrants, Mendoza and oases such as San Rafael expanded. Tucumán, lying amid more humid Andean foothills outside the rain shadow, responded to the new markets across the Pampa by increasing sugar production, which had had a modest beginning during early colonial times. Major change followed the first direct rail link between Tucumán and the Pampa in 1875, which provided access to expanding sugar markets and more modern machinery. Most of the tens of thousands of workers needed to harvest the crop came to live year-round on the large plantations, making Tucumán Province the most densely settled in Argentina.

Growth of Tucumán

The Gran Chaco. Although the Gran Chaco is still considered to be a frontier region, its settlement has often followed efforts supported by the government to exploit its economic potential. Agricultural colonies and cities grew first along the Paraná–Paraguay water route and then along railroads built to serve the quebracho industry. Resistencia was founded in 1878 and Formosa in 1879.

The harsh physical conditions of the Gran Chaco explain why its native peoples did not engage in sophisticated agriculture and why their sporadic attempts at farming were unsuccessful. Early Spanish expeditions aiming to conquer the Chaco came from Santiago del Estero to the west, Santa Fé to the southeast, and Asunción, across the Paraguay River to the northeast. None of these succeeded in subduing the warlike Indians, however.

Chaco Indian resistance

Settlement in the Chaco ultimately took place from Santiago del Estero, where irrigated cotton was successfully produced as early as the mid-16th century, and from Santa Fe, where cattle ranchers had purchased enormous acreages on which to raise tough criollo (or Creole) cattle, which had survived from earlier expeditions. Ranchers reached the northern frontier of the Argentine Chaco around the Bermejo River after subduing the Indians in 1885. Logging operations followed the ranchers and helped open parts of the Chaco—particularly in the east, where tannin from the quebracho tree met the demand of the Argentine leather industry. At the start of the 20th century, European settlers in the eastern Chaco began raising cotton, a crop that could withstand the long drought period. Small cotton areas spread westward nearly to Tucumán, north to the Paraguayan border at the Pilcomayo River, and east into Mesopotamia.

Mesopotamia. The northern part of the Mesopotamian region was first settled by Spaniards from Asunción (now the capital of Paraguay), who in 1588 founded the city of Corrientes near the confluence of the Paraná and Paraguay rivers. In the south, settlers from Santa Fe crossed the Paraná and established what became the city of Paraná. Having founded towns along navigable rivers, the Spanish secured the water route to the Río de la Plata estuary.

When the Spanish first entered the Mesopotamian region, supply distances between settlements were so great that the settlers found it necessary to produce their own subsistence crops. This they accomplished mainly by subjugating the remaining Indians under the encomienda system, which granted settlers the use of Indian labour on lands awarded by the crown. Indian rebellions became so serious, however, that the Spanish had to resort to military force to keep peace, and the conflicts forced many Indians to flee. Finally, in the early 17th century, the crown turned to the Jesuits to restore peace and protect the natives. Within a century the Jesuits had built numerous *reducciones,* or mission settlements, in Mesopotamia, which later acquired the name Misiones. Under Jesuit rule northern Mesopotamia became the most important centre of colonization in eastern South America.

The Jesuit *reducciones*

The Territory of Misiones was created in the early 1880s, and Europeans, particularly Germans, began to settle the forested zone in the north. Yerba maté, citrus, and veg-

etables, as well as tung trees, tea, and sugarcane, were grown on small farms. Outside the agricultural zones of Mesopotamia, cattle ranching came to dominate.

The Pampa. The Pampa was originally inhabited by Indians such as the Querandí, who reportedly had no agriculture but were hunters who developed the bola, for entangling the fleet-footed guanaco and rhea. Fierce attacks by the Querandí forced Spanish settlers in Buenos Aires to flee upriver to Asunción in 1541. After Buenos Aires reemerged in 1580, the Spanish showed less interest in opening up the southern Pampa than in keeping open the northern trade route to Santa Fe, Asunción, and Alto Perú; as a result the estancias, huge cattle ranches, were first established northwest of Buenos Aires.

The estancias

The estancias became one of the most important institutions in the economy, politics, and culture of Argentina. They began as gigantic tracts of land, often measuring in the hundreds of square miles, that were sold or granted to the Creole descendants of Spanish settlers during the 17th century. Herds of criollo cattle and horses ran half-wild on these tracts. To manage the herds, the estancia owners hired gauchos. These ranch hands worked the estancias until the open ranges disappeared late in the 19th century.

On the estancias widely dispersed *ranchos,* or simple adobe houses with dooryard gardens, marked the headquarters of the estancieros. More primitive huts, or leantos, housed the gauchos. In addition, there were small pulperias, centrally located inns where marketing, banking, eating and drinking, and other functions took place. Some pulperias grew into villages. Gradually the estancia region of the Pampa spread west and south of Buenos Aires.

Buenos Aires and Santa Fe survived as small, sparsely populated towns until the mid-19th century. After that time rapid growth in agriculture changed the face of the Pampa. The world market for food products was increasing, and estancieros modernized their operations to meet the demand. Sheep and more productive breeds of English cattle replaced the criollo. The new cattle, unable to live on the Pampa grass, had to be fed with alfalfa. Gauchos were not numerous or willing enough to cultivate this crop, obliging their employers to contract European immigrants as tenant farmers. The southern frontier of the Pampa was pushed back so that by 1880 no Indian interference to settlement remained north of the Negro River in Patagonia. By 1914 several million European workers had arrived to work ranches and farms. Gradually, small farming and tenant farming operations spread west and south from Santa Fe and Entre Ríos provinces.

The growth of agriculture spurred the growth of cities in the Pampa. Railroads radiating from Buenos Aires penetrated the interior of the Pampa, forming the densest network in the country. After the late 1800s foreign-owned frigorificos, meat-packing plants for the export of beef and mutton, were established on the Río de la Plata estuary. Efforts by the government in the 20th century to encourage the growth of manufacturing favoured the port cities, attracting most immigrants as well as many workers from the countryside. Buenos Aires subsequently became one of the most populous and cosmopolitan cities of the world, and the Humid Pampa became the most prosperous industrial and agricultural region of Argentina.

Patagonia. Most approaches to Patagonia from the sea were hampered by inhospitable coastal cliffs and by high tides. With the Pampa Indians acting as a buffer against Europeans to the north, the Patagonian Indians thus remained unmolested until the mid-19th century. The Conquest of the Desert, as the Patagonian Indian wars were called, ended in 1879, smashing the Indian resistance to settlement. Argentines began to colonize Patagonia, with soldiers and financial contributors to the Desert War receiving large land grants. Settlement proceeded south from the Pampean port city of Bahía Blanca and from Neuquén in the Andean foothills. Pioneers came from other countries as well. Chileans from Punta Arenas settled in Tierra del Fuego. Welsh, Scottish, and English immigrants spread along the coast and inland, with the result that English is still spoken in parts of Patagonia.

Subjugation of the Patagonian Indians

The southernmost city in the world, Ushuaia, on Tierra del Fuego, began as a missionary settlement and is still reached only by ship or airplane. About the end of the 19th century sheep ranching began along the rail line connecting the port of Río Gallegos with coal deposits at Río Turbio. Comodoro Rivadavia became an important oil and natural gas centre, and the Negro River fruit region began to develop in 1886 when the area east of Neuquén was settled by soldiers of the Desert War and by others.

THE PEOPLE

Heavy immigration, particularly from Spain and Italy, has produced in Argentina a people who are almost all white and of European ancestry. In the colonial period, though, the Spanish explorers and settlers encountered a number of native peoples. Among these were the Diaguita tribes of the Andean Northwest, a town-dwelling, agricultural people who were forced into labour after they had been conquered. They were divided by the Spanish into small groups and were sent to work in Peru and the Río de la Plata area. In the Mesopotamian region the agricultural Guaraní were also forced into labour.

Most other Argentine Indians belonged to hunting tribes who fought the Spanish tenaciously but were eventually exterminated or driven away. In the Gran Chaco were the Guaycuruan tribes among others. The Araucanian Indians came over the mountains from Chile and raided Spanish settlements in the southern Pampa until the Conquest of the Desert in the 1870s. Another Pampa Indian tribe was the Querandí, who inhabited the region of Buenos Aires. In Patagonia the largest group was the Tehuelche, and on Tierra del Fuego the Ona.

Ethnic mix in 1810

Population estimates of the colonial period suggest that by 1810 Argentina had more than 400,000 people. Of these perhaps 30 percent were Indian. Ten percent were black and mulatto, either slaves or descendants of slaves who had been smuggled into the country through Buenos Aires, and there was a large element of mestizos (white and Indian mixture). The whites were in the minority.

It was the great wave of European immigration after the mid-1800s that molded the present-day ethnic and racial character of Argentina. The Indians and mestizos were pushed aside or absorbed, and the blacks and mulattos disappeared, apparently also absorbed into the dominant population. Mestizos from Chile, Bolivia, and Paraguay have grown numerous in bordering regions, but the urban areas remain largely white.

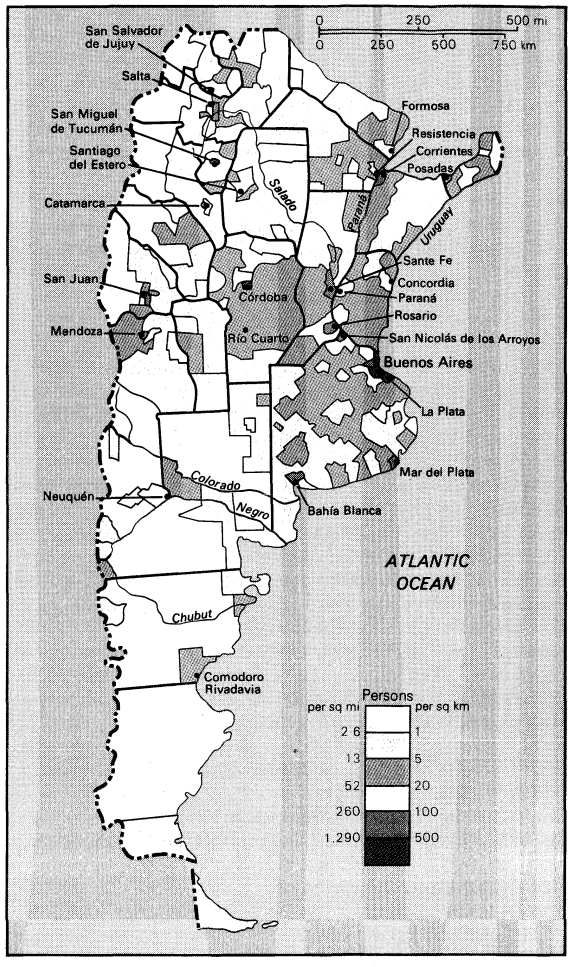
Almost half of the European immigrants in the late 19th and early 20th centuries were Italian and about one-third were Spanish. Substantial numbers also came from France, Poland, Russia, and Germany. In 1869 the foreign-born made up 12 percent of the population; this had grown to about one-third by 1914, and in large cities foreigners outnumbered natives by as much as two to one. As immigration slowed later in the 20th century, the proportion of foreign-born Argentines dropped.

Influence of European cultures

The Italian influence on Argentine culture became the most important of any immigrant group. Other major foreign influences have come from the Spanish and Polish groups. Smaller groups have also made notable contributions, however. English capital and management, for example, built railroads and created the meat-processing industry; Germans established farm settlements and cooperatives; the French contributed their viticultural expertise; and Japanese invested in business.

The children of immigrants were quick to identify themselves as Argentines, so that the people were not divided into antagonistic ethnic groups. But Argentine society developed a serious division between the rural interior and the urban coast. Many rural people grew to resent the wealth, political power, and cultural affectations of the *porteños,* the "people of the port," in the Buenos Aires region, while many *porteños* looked upon residents of the interior as ignorant peasants. These divisions became deeply rooted in the politics of the country.

Spanish is the national language, although in Argentina it is spoken in many accents and has absorbed many words from other languages, especially Italian. Numerous foreign languages and dialects can be heard, from Basque and Sicilian to Welsh and Gaelic. Toward the end of the 19th century an underworld language called *lunfardo* de-



Population density of Argentina.

veloped in Buenos Aires. It was composed of words from many languages—among them Italian, Portuguese, Spanish, French, German, and African. Most of the Argentine people are adherents of Roman Catholicism, the official state religion. Of the remainder about equally small percentages are Protestants and Jews. Roman Catholic influence is strongly reflected in government and society, and the religion is officially recognized in the constitution, although freedom of worship is professed.

Religion

The population of Argentina has increased by more than 17 times since 1,800,000 people were recorded there by the first census in 1869. This rapid growth began to decline after the early part of the 20th century as both the birthrate and immigration began to drop off. By the late 20th century Argentina's birthrate and growth rate were among South America's lowest; the percentage of young people has declined proportionally. The nation's population density is also among the continent's lowest, although certain areas are quite heavily populated, including the Humid Pampa, Mesopotamia, and parts of the eastern Northwest. The population is growing faster in urban areas—especially Buenos Aires—than in the rest of the country. More than four-fifths of the people live in urban areas, about a third in greater Buenos Aires alone.

THE ECONOMY

In the 60 years after the founding of the farming colony at Esperanza in 1856 the base of Argentine agriculture shifted from livestock to crops. The spread of wheat, corn (maize), and flax roughly conformed to the estancia region of the Pampa. Although agriculture there did not become as intensive as in North America, soils were good and land was abundant. Argentine industry became important when manufacturers (mostly foreign-dominated) developed food products for export. The growth trend continued well into the 20th century as Argentina became one of the most prosperous countries in Latin America. Meat and grain were exported to the expanding markets of Europe in exchange for fuel and manufactured products.

In the early decades of the 20th century Argentina became the world's leading exporter of corn, flax, and meat. Prosperity was curtailed, however, by World War I and the Great Depression of the 1930s, which considerably damaged the Argentine economy by reducing foreign trade. In response, successive governments over the next 40 years followed an import-substitution strategy designed to transform Argentina into a country self-sufficient in industry as well as agriculture. This was accomplished mainly by imposing high tariffs on imports, thereby sheltering Argentine textile, leather, and home-appliance manufacturers from foreign competition. The government's encouragement of industrial growth, however, diverted investment from agriculture, causing agricultural production to fall dramatically. Fruits, vegetables, oilseeds such as soybeans and sunflowers, and industrial crops such as sugarcane and cotton increased their share of total agricultural production at the expense of the dominant grain crops.

By 1960 manufacturing contributed more to the country's wealth than did agriculture. Argentina had become largely self-sufficient in consumer goods, but it also was more dependent than ever on imported fuel and heavy machinery. In response, the government invested heavily in such basic industries as petroleum, natural gas, steel, petrochemicals, and transport; it also invited investment by foreign companies. By the mid-1970s Argentina was producing most of its own oil, steel, and automobiles and also was exporting a number of manufactured products. Manufacturing became the largest single component of the gross domestic product (GDP). The country also had become self-sufficient in fuel while remaining one of the world's major agricultural producers.

At the same time, growing government spending, large wage raises, and inefficient production created a chronic inflation that rose until, in the 1980s, it briefly exceeded an annual rate of 1,000 percent. Successive regimes tried to control inflation through various combinations of wage and price controls, cuts in public spending, and restriction of the money supply. With the peso quickly losing value to inflation, a new currency, the peso argentino, was introduced in 1983, only to be replaced by the austral in 1985.

The era of import substitution ended in 1976, when an economic adjustment program lowered import barriers, liberalized restrictions on foreign borrowing, and gave the peso support against foreign currencies. These measures were intended to control inflation and increase efficiency by forcing competition upon the sheltered Argentine economy. Instead, they made imported products so cheap that many Argentine manufacturers could not compete with them. Between 1975 and 1981 manufacturing's share of the GDP declined from about 32 to 25 percent.

The economic adjustment program brought about another difficult economic problem: a huge foreign debt. Within a decade, borrowing from foreign creditors for many state and private-sector industrial schemes had quintupled Argentina's foreign debt. The overvalued peso made exports expensive, so that export earnings could not keep pace with the growing debt. Even as the government dismantled parts of the economic adjustment program, it had to address fundamental defects in Argentina's economy. The country had agricultural and industrial sectors similar to those of developed countries, but they were considerably less efficient. And despite a high standard of living by South American standards, it had a foreign debt comparable to that of Third World countries.

Resources. Argentine industry is well served by the country's abundance of energy resources. By the late 20th century the country was self-sufficient in power, even to the degree that it could become a power exporter. The two main sources of power are petroleum and hydroelectricity. Petroleum deposits are scattered throughout the country. The basin around the Patagonian port of Comodoro Rivadavia is estimated to hold some two-thirds of the country's onshore reserves. Other deposits are located in Jujuy and Salta provinces, in Mendoza and Neuquén provinces, and at the tip of Patagonia and Tierra del Fuego.

While most of the country's power has traditionally been

derived from petroleum, there has also been emphasis on the need to develop other energy resources. It has been estimated that half of the country's energy reserves lie in the potential of its rivers. Power from hydroelectric projects was increased by 10 times in the 1970s, and plans for the completion of more projects have been made. Such plans also involve the greater realization of abundant natural gas supplies and the production of nuclear energy, the latter being an area in which Argentina, with several nuclear plants, is a Latin-American leader. The main natural gas fields are in the Northwest, near Campo Durán (Salta Province) and Mendoza, and in Patagonia, near Neuquén and Comodoro Rivadavia. Argentina mines some coal, but most of its needs are met by imports. The chief coal deposits are in southern Patagonia.

Nuclear energy

With the exception of oil and natural gas, exploitable mineral reserves are generally small and widely scattered. Reserves of iron ore, uranium, lead, zinc, silver, copper, manganese, and tungsten are worked. A wide range of nonmetallic minerals is found throughout the country. Salt deposits are located on the western and southwestern edges of the Pampa. Materials such as clay, limestone, granite, and marble supply the construction industries.

Agriculture, forestry, and fishing. Argentina is one of the world's major exporters of soybeans and wheat, as well as meat. It is also one of the largest producers of wool and wine, most of the wine being consumed domestically.

Wheat is Argentina's largest crop in harvested land area, and it is the dominant crop in the cattle-raising southern Pampa of Buenos Aires and La Pampa provinces. Wheat and corn dominate in the north. Planting of corn began simultaneously with wheat in the northern Pampa. By the end of World War II, however, corn production had been cut in half due to worldwide competition, and production has only gradually increased since then. About half of the corn produced is used for livestock feed. The total area of the Pampa planted in sorghum and soybeans has grown since 1960 to rank just behind that of wheat and corn. These crops serve primarily as livestock feed and are valuable for export. Another crop of the northern Pampa is flax.

More than 90 percent of the country's grape vines are planted in the Northwest provinces of Mendoza and San Juan; most of the crop is for wine making. Table grapes are a specialty in La Rioja. The warmer northern provinces of Tucumán, Salta, and Jujuy make up the sugarcane-growing region of Argentina. The sugarcane provinces have also introduced citrus crops, because of the unreliable sugar market, and Salta and Jujuy grow tobacco. The best area for cotton growing lies mainly west of the Paraná River, between the Bermejo and Dulce rivers. Most of the crop goes to the Argentine textile industry.

Sugarcane growing in the Northwest

In Mesopotamia yerba maté is the most important product of Misiones Province, although since 1940 inroads have been made by tea farming and by the cultivation of tung trees, from which tung oil is derived. Citrus has also become important. Farther south in Mesopotamia, in the truck-farming area that supports Buenos Aires, oranges, grapefruit, mandarins, and numerous vegetables are grown. The Negro River irrigation district in Patagonia has become one of Argentina's major fruit-producing regions, particularly of apples and pears.

As the source of the country's most valuable export commodity, beef cattle dominate the Pampa. Estancieros have proved quick to adapt to changing markets, switching breeds and supplementing alfalfa feed with grain sorghum in order to produce leaner meat. Most of Argentina's hogs are raised in the Pampa, principally for domestic consumption. The cool, moist area of the southeastern Pampa, between Buenos Aires and the city of Mar del Plata, is an important dairy and sheep-raising district. Corrientes and Entre Ríos remain important cattle-raising provinces, ranking just behind those of the Pampa. Chaco Province began as grazing ground for criollo cattle, but modern breeds seem inevitably to succumb to disease, so that the cattle economy there remains backward. Patagonia has at least half of the country's sheep, most of which are sheared for their wool.

The forestry industry does not supply all of Argentina's needs. Most of the harvest goes to lumber, with smaller amounts to firewood and charcoal. In Mesopotamia the Paraná pine is harvested for lumber; there are also plantations of poplar and willow. The Northwest highlands produce pine and cedar, used for pulp and industry. The red quebracho of the Chaco region is valuable for its tannin, and the white quebracho is used for lumber and charcoal. Scattered stands of algarraba provide local firewood and cabinet wood in the Pampa.

The fishing industry is comparatively small, owing in part to the overwhelming preference among Argentines for beef in their diet. Most coastal and deep-sea fishing is done in the Buenos Aires sector, from the Río de la Plata to the Gulf of San Matías; the major ports are Mar del Plata and Bahía Blanca. Hake, squid, and shrimp make up a large part of the catch, about three-quarters of which is frozen or processed into oil and fish meal for export.

Industry. The product that initiated industrialization in Argentina was beef. This success was due in no small part to refrigeration techniques that, after 1876, made it possible to store and ship fresh meat. By the late 1920s frigoríficos (meat-packing plants) were located in various parts of the country, several of them in the Buenos Aires area. Later, shipments proceeded from La Plata, Rosario, and Bahía Blanca. Frigoríficos at the ports of Patagonia also came to serve the sheep ranches of that region.

Proliferation of frigoríficos

The growth of beef production in Argentina gave rise to a host of associated industries, including those producing tinned beef, meat extracts, tallow, hides, and leather. The Chaco region supplies the necessary tannin, of which it is a major world producer. Argentina has been a consistent world leader in the export of hides. Leather processing occurs locally, and fine leather clothing can be obtained at retail outlets in the cities.

The Argentine grain milling industry has grown in cities that built huge storage silos. As production has increased, grain has become a significant export. Wheat flour produced in the mills that emerged in the silo areas is consumed locally. Certain kinds of food industries based on wheat flour and various pastas have been attracted to the same sites along the Río de la Plata littoral. Smaller but similar activities are emerging in the interior of Argentina, wherever grain production takes place. Textile production in Argentina also developed on the basis of agricultural products, namely wool and cotton. It is concentrated in the cities of the Pampa, where the labour needs can be met and where the largest markets are located.

Textile production

Argentina's refining industry has grown along the coast in Buenos Aires and nearby cities. Tankers and pipelines from Comodoro Rivadavia and Venezuela bring crude oil to this area. The refining industry has also found a base in the petroleum fields north and south of Mendoza, where petrochemical industries have emerged.

The steel industry in Argentina began in the 1940s and grew slowly during the following decades. The Zapla works in Jujuy, the integrated San Nicolás de los Arroyos mill between Rosario and Buenos Aires, and the mill in Rosario produce most of the nation's steel but fall short of supplying domestic demand. A developing automobile industry provides a market for Argentine steel producers.

The Argentine sugar industry of the Northwest is centred mainly in Tucumán, but a few mills also operate in Salta and Jujuy. The sugar mills fulfill domestic demand. Mendoza in the same region is the nation's centre for olive and olive oil production, as well as for wine bottling. Argentine wine shipments to other South American countries and to North America have become common, based on a steadily improving reputation among consumers. The Northwest is also a major area for cement production, and there is a developing aircraft industry at Córdoba.

Finance. Argentina's financial system includes banks owned by the national government, by provincial and municipal authorities, and by private companies. The Banco Central de la República Argentina is the nation's central bank; it issues currency, sets interest and exchange rates, and regulates the money supply by deciding the amount of reserve cash banks must hold.

The central bank

Economic troubles beginning in the 1970s caused the near collapse of the country's financial system. Inflation

made savings deposits almost worthless; liberalized foreign-investment laws and an overvalued peso encouraged excessive borrowing abroad; and defaulted loans caused many bankruptcies. To prevent complete collapse, the government took greater control of the private sector.

Trade. Argentine trade has always been oriented toward Europe and the United States. Following the colonial period, the United Kingdom began acquiring its wheat and meat from the Pampa region. After World War II, Argentina attempted to raise beef and grain prices in order to rebuild lost sterling reserves, but the attempt failed. In addition, Britain's trade was inclined toward those countries seeking consumer goods while, at the same time, Argentina was attempting to become self-sufficient in such needs. Trade with Britain diminished, falling especially after the Falkland Islands war in 1982. Since that war continental European countries (particularly The Netherlands and West Germany) and the Soviet Union have become important trading partners of Argentina, although the United States remains the single most important market for the country's exports. Brazil also ranks high in trade volume with Argentina. Agricultural products are Argentina's major exports, and its imports include nonelectrical and electrical machinery, chemicals, and petroleum and petroleum products.

Major trading partners

Transportation. During the Spanish colonial period, there were three principal overland transportation routes. The most important led from Buenos Aires to the wealthy mining centre in Peru via the northwestern route through Córdoba, Santiago del Estero, Tucumán, and Jujuy. A second route linked Buenos Aires with Chile westward to Villa María, San Luís, and Mendoza. The third route extended north from Buenos Aires to Santa Fe and Corrientes. These and less important side roads were used by mule drivers, horsemen, huge-wheeled oxcarts called carretas, and stagecoaches drawn by six to eight horses.

The transformation of this system did not occur by means of the modernization of roads but rather by the rapid introduction of rail lines during the period just after 1857. British and other foreign capital funded rail networks that led from Buenos Aires in several directions. Road and rail construction continued from that time into the 20th century, extending throughout the nation. By the late 20th century, Argentina had the most extensive transportation system in Latin America. The largest share of surface freight is carried by road, with lesser amounts carried by river, pipeline, and railroad.

Extension of the rail and road network

Small ships that carry passengers and freight have served the coastal cities from Buenos Aires to Río Gallegos since the end of the 19th century. The ocean shipping fleet, however, is not well developed, considering Argentina's extensive export trade. Air transportation provides rapid travel to all regions of the nation. Every major city has a jet airport, and even small, remote centres like Ushuaia in southern Patagonia have good service. The country's most important air transport company, Aerolíneas Argentinas, was founded by the government in 1950 to handle domestic and international traffic. International airports are located in most of the major cities, the most important being Ezeiza outside of Buenos Aires.

GOVERNMENT AND SOCIAL CONDITIONS

Government. Argentina is a federal union of 22 provinces, one territory, and a federal capital, the city of Buenos Aires. Federalism came to Argentina only after a long struggle between proponents of a central government and supporters of provincial interests. The constitution of 1853 was modeled after that of the United States, and, although altered a number of times, the document has sustained Argentina with at least a nominal form of republican, representative, and federal government.

The executive and the legislature

Executive power resides in the office of the president, who is elected with a vice president to a six-year term. The president may not be reelected for consecutive terms. The president is commander in chief of the armed forces and appoints all civil, military, and federal judicial officers. Both the president and the vice president must profess Roman Catholicism. The Argentine legislature, or National Congress, consists of two houses: a Senate, composed of two representatives from each province and the federal capital elected for nine-year terms; and the Chamber of Deputies, whose members are elected for four-year terms, with apportionment based on population. Each province has its own government, with executive, legislative, and judicial branches similar to those of the federal government.

The Argentine judicial system is divided into federal and provincial courts. Supreme Court judges are appointed by the president with approval of the Senate. Federal judges are appointed constitutionally for life, but it has not been unusual for them to serve only as long as the administration that appointed them.

Political parties

The Argentine system of political parties has been volatile, particularly in the 20th century, with numerous parties being formed, taking part in elections, and being disbanded as new factions evolved. Among the major parties operating in Argentina are the Radical Civic Union, a centrist party with moderate leftist leanings; the National Justicialist Movement, the Peronist party (made up of followers of former president Juan Perón), divided into factions and made up largely of nationalists and labourers; and the Union of the Democratic Centre, a coalition of right-wing parties.

Education. Argentina has one of the better educated populations in Latin America, which is reflected in its large number of schools and high literacy rate. Primary education is compulsory and free; secondary and higher education is offered in free public schools and in private schools subsidized by the state. Higher education in Argentina was seriously hampered by the censorship and other strictures of the military government of 1976–83, but efforts to restore the system began after a civilian government was returned to power. The National University of Córdoba, founded in 1613, is the nation's oldest, and the University of Buenos Aires, founded in 1821, is its largest. Other major national universities are at Mendoza, La Plata, Rosario, and Tucumán. The National Technical University is located at Buenos Aires.

Major universities

Health and welfare. An extensive system of hospitals and clinics in Argentina is run by national, provincial, and local authorities as well as by private organizations. Public health and sanitation standards are particularly high in developed areas but can drop off considerably in some of the undeveloped areas. Diseases such as smallpox, cholera, yellow fever, and tuberculosis have been brought under control or eliminated.

Argentina's social welfare services were developed on a large scale during the first presidency of Juan Perón (1946–55). A social security system was set up to provide extensive benefits for all workers. Housing, however, has become a problem in cities because of the movement of workers from rural areas, especially during difficult economic periods. This has produced new living quarters on the outskirts of urban zones that have been built hastily from corrugated iron and scraps of wood, cardboard, and other scavenged materials. These communities are called *villas miserias,* and, despite pressures from various directions, little has been done to improve their condition.

CULTURAL LIFE

Heritage and daily life. Because almost all Argentines are descendants of relatively recent immigrants from Europe, their culture has a more distinctly European orientation than that of their fellow Latin Americans. The people of Buenos Aires, the *porteños,* often call their city the Paris of South America, and, with its culture and glamour, it probably earns that name. But there is another Argentina away from the capital: this is the Argentina of the Pampa and the interior. The interior gave to all Argentines their symbol of national identity, the gaucho, who occupies a position in South American lore similar to that of the cowboy in the United States. Scorned in his heyday of the 18th and 19th centuries as a drinker and vagabond, this mestizo ranch hand rode the open rangeland of the huge estancias in pursuit of wild horses and criollo cattle. Eventually Argentines came to see him as a character whose solitary life in the open taught him self-reliance, courage, indifference to hardship, and love of the land, traits that represented the ideal of their national character.

The gaucho

The tango

Another hybrid of the Old and New Worlds is the tango, which emerged from the poor immigrant quarters of Buenos Aires toward the end of the 19th century and quickly became famous around the world as the Argentine national dance. Influenced by the Spanish tango and, possibly, the Argentine *milonga,* it was originally a high-spirited local dance, but, popularized by such singers as Carlos Gardel, it became an elegant ballroom form danced to melancholy tunes.

The combination of Old and New World cultures is also seen in the Argentine diet. Southern European influences appear especially in the cities, where breakfast is often a light serving of rolls and coffee, and supper is taken, in the Spanish tradition, after nine o'clock at night. The Italian influence is seen in the popularity of pasta dishes. But the New World asserts itself in the Argentine passion for beef, which is overwhelmingly preferred to other meats and fish. Maté, the native tealike beverage brewed from yerba maté leaves, is popular in the countryside.

The arts. The fine arts of Argentina have always found their inspiration in Europe, particularly in France and Spain. In literature the Modernismo movement of the late 19th century and the Ultraísmo of the early 20th were both influenced by the French Symbolist and Parnassian poets. By composing verses of unconventional metre and by using unusual imagery and symbolism, such poets as Leopoldo Lugones and Jorge Luis Borges hoped to draw attention to the beauty of the Spanish language. Borges went on to become one of the most innovative novelists of Latin America. He prepared the way for experimental novelists of the later 20th century.

Music, painting, and sculpture

Composers of the early 20th century, such as Alberto Williams and Carlos López Buchardo, contributed to a nationalist revival in music by adapting folk and gaucho themes to classical forms. A generation later Alberto Ginastera and Juan Carlos Paz experimented with musical forms that were current throughout Europe and the Americas. Painters and sculptors studied in Italy and France and brought the academic, Impressionist, and Cubist styles to Argentina. Later artists were inspired by Mexican murals and by abstract and Pop art in the United States.

Recreation. The most popular sport is football (soccer), introduced by the British in the 19th century; Argentine teams are generally among the best internationally and have won the World Cup. The British also brought polo, at which the equestrian-loving Argentines quickly excelled. There are excellent hiking and fishing areas in the Lake District of the Patagonian Andes, and skiers travel to Andean resorts. In the summer months bathers pack the beaches at resorts, such as Mar del Plata.

Important civic holidays are the Anniversary of the Revolution (May 25) and Independence Day (July 9). Christmas is a national holiday. Regional festivals include the Fiesta del Milagro in Salta, commemorating the salvation of the city from an earthquake in September 1692, the celebration on July 6 of the founding of Córdoba, and the wine festival in Mendoza in March.

Press and broadcasting. The mass media in Argentina are well advanced among Latin American nations. All of the largest newspapers are published in Buenos Aires. The largest daily circulation is claimed by *Clarín;* two other large-circulation dailies, *La Nación* and *La Prensa,* founded in 1870 and 1869, respectively, have high reputations in the Spanish-speaking world as well as among the international press. In the capital various foreign-language papers serve the ethnic groups. The majority of the radio and television stations are privately operated. Throughout the country's postwar history the broadcast media have periodically become agents of state propaganda, only to be returned to some independence by succeeding administrations. This process has also afflicted the press.

For statistical data on the land and people of Argentina, see the *Britannica World Data* section in the BRITANNICA WORLD DATA ANNUAL. (R.C.Ei./Ed.)

History

It has been estimated that the population living in the land area of what is now called Argentina before the arrival of the Europeans totaled some 300,000. As in the rest of the New World, it was composed of Indians believed to have been descendants of Asians who, in prehistoric times, are thought to have migrated across the Bering Strait to the North American continent and who gradually spread across North, South, and Central America. The state of civilization of the Argentine Indian did not match those of the Aztec and the Maya and the empire-building Incas. Some of the tribes, such as those in the Chaco and Patagonia, were nomadic hunters and fishers; but others, such as the Diaguitas of the Northwest and the Araucanians of the Pampa, developed a primitive agriculture, handicrafts, trade, and relatively sophisticated weaponry.

EARLY PERIOD

Discovery and settlement. It is well established that Amerigo Vespucci stimulated the Spanish search for a southern strait, which led to the discovery both of the Río de la Plata by Juan Díaz de Solís in 1516 and of the Strait of Magellan by Ferdinand Magellan in 1520. These two voyagers revealed the main outlines of the Atlantic coast of what is now Argentina. Sailing up the Plata, which he called the Mar Dulce, or Freshwater Sea, Solís, with a small party, landed and was ambushed by Indians. Solís and most of his followers were killed; several disappeared. The survivors of the expedition returned to Spain.

Exploration by Sebastian Cabot

The Río de la Plata was not explored again until Magellan arrived in 1520 and Sebastian Cabot in 1526. The latter picked up two survivors of the Solís expedition, who told glowing tales of the wealth of the region. Cabot discovered the Paraná and Paraguay rivers and established the fort of Sancti Spíritus (the first Spanish settlement in the Plata basin). He also sent home reports of the presence of silver.

In 1528 Cabot met another expedition from Spain under Diego García, commander of a ship from the Solís expedition. Both Cabot and García had been supposed to sail for the Moluccas but altered their courses, influenced by excited tales about an "enchanted City of the Caesars," a variant of the Eldorado legend, which later incited many explorations and conquests in Argentina. While Cabot himself was preparing to search for the fabled city, a surprise attack by the Indians in September 1529 wiped out his Sancti Spíritus base.

Inspired by the conquest of Peru and the threat from Portugal's growing power in Brazil, Spain in 1535 sent an expedition under Pedro de Mendoza (equipped at his own expense) to settle the country. Mendoza was initially successful in founding Santa María del Buen Aire, or Buenos Aires (1536), but lack of food proved fatal. Discouraged by Indian attacks and mortally ill, Mendoza sailed for Spain in 1537, dying on the way.

In the same year a party from Buenos Aires under Juan de Ayolas and Domingo Martínez de Irala, lieutenants of Mendoza, pushed a thousand miles up the Plata and Paraguay rivers. Ayolas was lost on an exploring expedition, but Irala founded Asunción among the Guaraní, a sedentary, agricultural people. In 1541 the few remaining inhabitants of Buenos Aires abandoned it and moved to Asunción, which was the first permanent settlement in this area. In the next half-century Asunción played a major part in the conquest and settlement of northern Argentina. The main population of Argentina was concentrated here until the late 18th century. Buenos Aires, reestablished in 1580 by Juan de Garay with settlers from Asunción, was largely isolated from this northern area. Northern Argentina as well as Buenos Aires was settled mainly by the overflow from neighbouring Spanish colonies—Chile, Peru, and Paraguay (Asunción). There was little direct migration from Spain, probably because the area lacked the attractions of Mexico, Peru, and other Spanish colonies—rich mines, a large supply of tractable Indian labour, accessibility, and the privilege of direct trade with Spain. Nevertheless, in the early communities a simple but vigorous society developed on the basis of Indian labour and the horses, cattle, and sheep imported by the Spaniards, as well as native products such as corn and potatoes; some of the Indians worked as virtual serfs. Missions established by the Roman Catholic Church played a notable role in the colonizing process. Because few Spanish women were

Role of Indian labour

among the settlers, there was much intermarriage with the Indians.

Colonial period. Politically, Argentina was a divided and subordinate part of the Viceroyalty of Peru until 1776, but three of its cities—Tucumán, Córdoba, and Buenos Aires—successively achieved a kind of leadership in the area and thereby sowed the seed of community feeling that much later grew into the idea of Argentine national union.

Tucumán's leadership lasted from the latter part of the 16th through the 17th century. Its political and ecclesiastical jurisdiction extended over most of northern Argentina, including Córdoba. Tucumán also dominated the chief economic activity of supplying the rich silver-mining area of Upper Peru (Bolivia) with foodstuffs and livestock in return for European manufactures and other goods brought from Spain. Under the same economic system Córdoba rose to leadership in the 17th and 18th centuries because the expansion of settlement gave the city a central location and because the University of Córdoba, founded in 1613, put the city in the intellectual forefront in Argentina and in Spanish America in general.

Preeminence of Buenos Aires

On the other hand, Buenos Aires, which rose to leadership in the late 18th century, symbolized three important reorientations of Argentine life from west to east. The first, economic, was the shift from trade with the now declining silver mines of Peru to direct transatlantic trade with Europe. Another reorientation, intellectual, resulted from the rise of interest in the new ideas of the European Enlightenment, which found fertile soil in cosmopolitan Buenos Aires. The third reorientation, political, was brought about by Spain's detaching its possessions in the Plata basin (modern Argentina, Uruguay, Paraguay, and southern Bolivia) from the Viceroyalty of Peru and erecting them into the new Viceroyalty of the Río de la Plata, with Buenos Aires as its capital (1776). Spain's main purpose was to put its Plata dominions in a better defensive position. The chief threat came from Brazil, which was growing rapidly in population, wealth, and military potential. For the first time the port of Buenos Aires was opened to transatlantic trade with Spain and, through Spain, with other countries. This resulted in a great increase in both legal trade and the already flourishing smuggling.

Effect of Napoleon's intervention in Spain

Independence. In Argentina, as in most of Spanish America, the independence movement was precipitated mainly by Napoleon's intervention in Spain, beginning in 1808. The break began when Napoleon's intervention plunged Spain into a civil war between two rival governments. One was set up by Napoleon, who placed his own brother Joseph Bonaparte on the throne. The other was created by patriotic juntas in Spain in the name of the exiled Ferdinand VII and aided by the British. In most of Spanish America, however, while there was general sympathy with the regency, both claims were rejected, mainly on the ground that an interregnum existed and that under ancient principles of Spanish law the king's dominions in America had the right to govern themselves pending the restoration of a lawful king.

This view was sustained in Argentina by the Creoles rather than by the "peninsular" Spaniards, and it was given effect by the Buenos Aires cabildo, or municipal council. This ancient Spanish institution had existed in all the colonies since the 16th century. Its powers were very limited, but it was the only organ that had given the colonists experience in self-government. In emergencies it was converted into an "open" cabildo, a kind of town meeting, which included substantial members of the community. On May 25, 1810, such an "open" cabildo in Buenos Aires set up an autonomous government to administer the Viceroyalty of La Plata in the name of Ferdinand VII, pending his restoration. When Ferdinand was restored in 1814, however, he promptly proved himself one of the worst kings Spain had ever had. Thereupon, an assembly representing most of the country met at San Miguel de Tucumán and, on July 9, 1816, declared the country independent under the name of the United Provinces of the Río de la Plata.

Several years of hard fighting followed before the Spanish royalists were defeated in northern Argentina. But they remained a threat from their base in Peru until it was liberated by Simón Bolívar in 1824. The Buenos Aires government tried to maintain the integrity of the old viceroyalty of the Río de la Plata, but the outlying portions, never effectively controlled, soon were lost: Paraguay in 1814, Bolivia in 1825, and Uruguay in 1828. The remaining territory, modern Argentina, was frequently disunited until 1860. The root cause of the trouble, the question of the relation of Buenos Aires to the rest of the country, was not settled until 1880, and, even after that, it continued to cause dissatisfaction.

Conflict between Buenos Aires and the hinterland

EFFORTS TOWARD RECONSTRUCTION, 1820–29

In 1820 there were only two political organizations that could claim more than strictly local and provincial followings: the revolutionary government in Buenos Aires and the League of Free Peoples (Liga de los Pueblos Libres) that had grown up along the Río de la Plata and its tributaries under the leadership of José Gervasio Artigas. But both organizations collapsed simultaneously in that year, and Buenos Aires seemed on the verge of losing its position as the seat of national government. However, as the city regained its function as an intermediary between the nation and foreign governments, it also regained its national prominence.

Dominance of Buenos Aires. In practically all of the provinces, military leaders had assumed power. Each provincial political regime soon acquired its own character, according to the relative power held by military strongmen and by local political interests. This differentiation was not, however, cause for friction among the provinces; their cleavages were more directly attributable to economic and geographical differences. Buenos Aires was able to make significant advances toward national leadership by prudently taking advantage of the interprovincial rivalries.

Reforms of the Party of Order

Within the Province of Buenos Aires itself, the regime of the so-called Party of Order (Partido del Orden) put into effect reform measures that were viewed with favour. Principal among them was the dismantling of the military apparatus that had remained from the war and the dedication of what remained to the defense of the frontier areas and plains against the Indian threat. This prudence on the part of the government in power won the support of the rural landowners as well as the urban businessmen, whose backing assured victory at the polls.

The political order that seemed to be taking hold in Buenos Aires, however, and in the country as a whole was achieved by setting aside, rather than resolving, certain fundamental difficulties. In particular, the institutional organization of the country was not carried out, and nothing was done about the Banda Oriental (the east bank of the Uruguay River), which was occupied first by Portuguese and then by Brazilian troops. By 1824, both problems were becoming urgent. England was willing to recognize Argentine independence if a government was established that could act for the whole country. In the Banda Oriental, a group of eastern patriots had taken over large sectors of the rural areas and agitated for their reincorporation into the United Provinces of the Río de la Plata, forcing the Buenos Aires government to face the possibility of war with the Brazilian empire.

Presidency of Rivadavia. In the meantime an attempt was made to establish a national government through a constituent assembly that met in December 1824. Overstepping its legal authority, the constituent assembly, in February 1826, created the office of president of the republic and installed the *norteño* ("northerner") Bernardino Rivadavia as its first occupant. Civil war flared up in the interior provinces, soon dominated by Juan Facundo Quiroga—the caudillo from La Rioja who opposed centralization. When the assembly finally drafted a national constitution, the major portion of the country rejected it.

War against Brazil had begun in 1825. While the Argentine forces were able to defeat the Brazilians on the pampas of Uruguay, the Brazilian navy blockaded the Río de la Plata and succeeded in crippling Argentine commerce. Unable to end the war on favourable terms, Rivadavia resigned in July 1827, and the national government dissolved. Leadership of the Province of Buenos Aires was given to a federalist, Col. Manuel Dorrego. Behind Dorrego there were local interest groups whose political spokesman

was the great landowner Juan Manuel de Rosas, who had been named commander of the rural militia. Dorrego was prevailed upon to make peace with Brazil. In 1828 the disputed eastern province was constituted as the independent state of Uruguay; the territory that Rivadavia had considered indispensable to the "national integrity" of Argentina was never to be recovered. In December 1828, troops returning from the war overthrew Dorrego and installed Gen. Juan Lavalle in his place; Dorrego was executed.

While there was little resistance to the new governor in the city of Buenos Aires, uprisings began quite promptly in the outlying areas of the province. In Santa Fe a convention of representatives from the provinces dominated by the federalists, under the leadership of Rosas, called upon the Governor of Santa Fe to take steps against the Lavalle regime. Lavalle finally came to terms with Rosas, and they agreed to hold elections in Buenos Aires for a new provincial legislature. Under the compromise agreement, Rosas and Lavalle appointed a moderate federalist as governor of Buenos Aires, but political tensions were too great for this attempt at reconciliation. Rosas reconvened the old legislature that Lavalle had disbanded when he came to power—a triumph for the most intransigent forces of federalism. The legislature unanimously elected Rosas governor on Dec. 5, 1829.

CONFEDERATION UNDER ROSAS, 1829–52

The regime of Rosas in Buenos Aires enjoyed far broader support than any of its predecessors. Special interest groups, landholders, and export-import merchants (along with the British diplomatic contingent that was identified with these interests) all fell behind the new governor. Practically all the influential sectors in the province identified Rosas' triumph with their own best interests.

Hegemony of Buenos Aires under Rosas

Domestic politics, 1829–35. The new governor, however, saw clearly the ambiguities and dangers of such widespread support. He sought a political formula that would enable him to harness the energies of the various sectors to his own purposes. He found this formula in factionalism. By imposing an orthodox federalism, the exact nature of which he alone would determine according to his own political objectives, Rosas would be assured of control over the masses and would be able to use this influence as a restraint on the economic and social elites, with whose interests Rosas fundamentally identified. This blueprint assigned to Rosas the role of permanent arbiter of a delicate and constantly threatened balance.

By 1832 the opposition to federalism had disappeared throughout the country, and Rosas turned over the reins of the government of Buenos Aires to his legal successor, Gen. Juan Ramón Balcarce. Balcarce's assumption of the office fanned sparks of dissidence among those who had pledged to uphold the principles of federalism. Balcarce was overthrown, and his successor took office with a Cabinet composed of Rosas' friends. They adopted policies that were designed to lead to normality. But it was normality that Rosas feared, since it would have entailed the demobilization of his mass political following. The legislature in Buenos Aires was induced to designate Rosas as governor of the province, but under conditions that Rosas himself successfully imposed: he was granted extraordinary resources, absolute public authority, and an extension of the governor's term of office from three to five years.

Foreign policies. In the field of international relations, Rosas' policies also left no room for anything other than total success or total failure. International difficulties arose as extensions of domestic turmoil. At that time it was not clear to everyone that the neighbouring countries of Bolivia, Paraguay, and Uruguay were destined to be independent states and not part of the federal grouping controlled by Buenos Aires. Gen. Andrés Santa Cruz, who had established a confederation of Peru and Bolivia, supported opponents of Rosas in Argentina. Rosas in turn supported the influential governor of the northern Province of Tucumán, when that governor decided to go to war against the forces of General Santa Cruz. In 1839 the war ended in victory for the northern Argentine forces, in part thanks to an alliance with Chile.

Rosas' involvement in civil struggles in Uruguay proved to be costly. It led to the first open friction with France, which sent warships to blockade Buenos Aires in 1838. This created dissension in the coastal region, which had always depended heavily on export trade. Argentine political exiles in Montevideo received French backing in their efforts to overthrow Rosas, and in the north a league of dissident provinces was formed.

This formidable coalition of adversaries soon fell apart. Faced with other problems, France abandoned its adventure in the Río de la Plata area and left its local allies to fend for themselves against Rosas. At the same time, an army organized by Buenos Aires and commanded by the Uruguayan leader Manuel Oribe gained control of most of the Argentine interior. For the first time since 1820, troops from Buenos Aires had advanced as far as the Bolivian and Chilean frontiers. The hegemony of Buenos Aires under Rosas' system of federalism was not to be challenged again. Oribe went on to conquer most of Uruguay, and his predominantly Argentine army began the siege of Montevideo in February 1843. The city was saved only through the intervention of British warships, and in 1845 an Anglo-French fleet blockaded Buenos Aires and a British fleet sailed up the Paraná River. Eventually the British and French withdrew their aid to Montevideo and made peace with Rosas.

The fact that Rosas was able to conduct a vigorous foreign policy for so many years was partly because of the weakness of Brazil, Argentina's natural rival in the Río de la Plata area, which had been involved in a civil war (1835–45) in Río Grande do Sul. Once the rebellion had been put down, it was only a question of time until Brazil's influence was again a factor in the Río de la Plata region. This influence opposed Rosas. It worked in support of a rebellion by Gen. Justo José de Urquiza, governor of the Province of Entre Ríos. In 1851 Urquiza formed an alliance with Brazil and Uruguay. The allies first forced Rosas' troops to abandon the siege of Montevideo and then defeated his main army in the Battle of Caseros (Feb. 3, 1852), just outside Buenos Aires. Rosas, abandoned by most of his troops as well as his political supporters, was taken to England, where he died in 1877.

Economic development, 1820–50. In the 30 years after 1820, Argentina's society and economy underwent considerable changes. Buenos Aires was the province best adapted to the new era of free trade based upon the export of cattle products in return for consumer goods from overseas. The interior provinces adjusted slowly, replacing their traditional markets in highland Peru with new ones in Chile, where a great expansion of the mining industry was taking place. The coastal provinces fared better, although their livestock industry suffered from the effects of the civil war. For Santa Fe, the decade of the 1830s brought a return to moderate prosperity, and a similar trend began in Entre Ríos and Corrientes provinces in the 1840s.

Growth of a cattle economy

NATIONAL CONSOLIDATION, 1852–80

General Urquiza called a constitutional convention that met in Santa Fe in 1852. Buenos Aires refused to participate, but the convention adopted a constitution for the whole country that went into effect on May 25, 1853. Buenos Aires refused to join the new confederation, the first elected president of which was Urquiza and the first capital of which was Paraná. The dissidence was a serious financial handicap to the state, since Buenos Aires kept for itself all the revenues from customs duties on imports. In 1859 Urquiza incorporated Buenos Aires by armed force, but he had to agree to a constitutional revision that underscored the federal character of the government.

Constitutional convention of 1852

Before the unification took effect, however, Urquiza was succeeded in the presidency by Santiago Derqui. Another civil war broke out, but this time Buenos Aires defeated the forces led by Urquiza. The latter came to an understanding with Gen. Bartolomé Mitre, governor of Buenos Aires. They agreed that Mitre would lead the country but that Urquiza would exercise authority over the provinces of Entre Ríos and Corrientes. Derqui resigned, and Mitre was elected president in 1862 with Buenos Aires as the seat of government.

The authority of the new president (1862–68) was pro-

gressively weakened by opposition within his own province of Buenos Aires. The pressures of this opposition forced Mitre definitely toward intervention in the political struggles of Uruguay and thence into a war with Paraguay. From 1865 to 1870, a triple alliance of Argentina, Brazil, and Uruguay carried on a war to the death against Paraguay, employing modern weapons and tens of thousands of troops.

The war with Paraguay did not disrupt Argentina's commerce, as other wars had. In the 1860s and 1870s foreign capital and waves of European immigrants poured into the country. Railroads were built; alfalfa, barbed wire, better breeds of cattle and sheep, and finally the refrigeration of meat were introduced. The national armed force became one of the cornerstones of the new centralized state.

But if the national state seemed to be taking shape with surprising vigour, the process never resulted in consolidation of the hegemony of the so-called liberal faction that had backed Mitre in 1861–62. Moreover, the army refused to uphold the policies of the President. One of Rosas' nephews rallied the support of the military behind the presidential candidacy of Domingo Faustino Sarmiento, a native of San Juan. His victory was guaranteed by the influence of the military combined with the support of the liberal faction in Buenos Aires that opposed Mitre. The new president (1868–74) held office without a political party of his own. Credit from abroad brought a continuing prosperity that allowed Sarmiento to engage in a costly civil war to put down an uprising in Entre Ríos.

Sarmiento was succeeded by his minister of justice, public education, and worship, Nicolás Avellaneda (1874–80), a native of Tucumán, whose victory was a reaction against the hegemony of Buenos Aires. The new president faced serious financial difficulties engendered by the European economic crisis of 1873. When the external sources of capital dried up, the country once again felt the consequences of Buenos Aires' financial superiority. Avellaneda adopted a policy of austerity, and he undertook to reach an agreement with the declining forces of liberal nationalism.

The Roca presidency

Gen. Julio Argentino Roca, who was also from Tucumán and who had influence in Córdoba, became the next president (1880–86). Roca had led a brilliant military career, during which he had brought the long cycle of Indian wars to a victorious close in 1879. This achievement, opening up the pampas to settlement, made Roca a political hero. His campaign for the presidency provoked a new rebellion in Buenos Aires, but the uprising was quickly put down. The perennial question of the city's status was then settled by making it a federal territory and converting it into the national capital; a new capital for the Province of Buenos Aires was established at La Plata.

THE CONSERVATIVE REGIME, 1880–1916

The entire country was now dominated by the National Autonomist Party, which was composed of an alliance of the various groups supporting Roca. These included many of the big ranchers, as well as commercial and business interests who were more than happy with Roca's formula of "peace and efficient administration." Argentina's economic expansion in this period owed much to British capital, which made possible the building of an extensive rail network linking the upriver provinces to Buenos Aires and the sea. Along with the growth of agriculture and ranching, there was also some expansion in other industries, but the impetus came from large-scale foreign investment. This was also an era of rapid growth in population, largely from immigration. Argentina's population grew from less than 2,000,000 in 1869 to nearly 8,000,000 in 1914.

Economic expansion

The crisis of 1890. The economic expansion led ultimately to inflation, accompanied by the issuance of too much paper currency and a crisis of confidence in the London capital market. The financial crisis was followed by a political one. The government of Roca's successor, Miguel Juárez Celman (1886–90), had been wary to launch a necessarily unpopular anti-inflationary program. Discontent was growing, both within and outside the official party ranks. In July a revolt erupted that had strong support from within the army, but it was defeated by loyal elements. Even so, Juárez Celman was forced to step down in favour of the vice president, Carlos Pellegrini, a solid ally of Roca.

The rise of radicalism. The difficulties of the 1890s brought the rise of a new party, the Radical Civic Union, strongly opposed to the ruling regime and the compromise candidate, Luis Sáenz Peña, accepted by Mitre and the more moderate opponents of the Roca–Juárez Celman regime. In 1898 Roca returned to the presidency for a second term (1898–1904) and began an attempt to bring the more moderate radicals effectively back into the loose alliance of local political groups, which after 1890 controlled the national government. The most intransigent radical factions, headed by Hipólito Irigoyen, who later became president twice, remained in opposition.

Broadening of political life

While political opposition was on the decrease, social unrest was becoming more widespread. Within the government itself there was growing disarray. When Roca broke with Pellegrini, it was a final blow for the National Autonomist Party, and Roca was barely able to avoid being succeeded in office by Pellegrini in 1904. The candidate he finally put into the presidency, Manuel Quintana (1904–06), was far from being one of Roca's stauchest supporters. Quintana was forced to quell a radical revolution in 1905, and his death opened the way to the presidency for the Cordoban José Figueroa Alcorta (1906–10), who turned immediately to the task of destroying Roca's political machine. In 1910, Alcorta installed as his successor Roque Sáenz Peña (1910–14), a brilliant politician who was fully prepared to construct a governing coalition on new foundations.

The course of Argentine politics in the final stages of Roca's career had convinced many of his most influential and militant followers that the country needed electoral reform, which was not seen as excessively dangerous since the Radical opposition seemed to have limited support. In 1912, Sáenz Peña had the Congress pass an electoral-reform law that called for a compulsory, secret ballot for all male citizens. His death in 1914 deprived the national leadership of its guiding force, and the electoral law he had authored opened the gates of power to the Radicals. The interim presidency of Victorino de la Plaza (1914–16) was followed by that of the Radical leader Hipólito Irigoyen (1916–22). He was the first Argentine president who owed his victory to the popular vote rather than to selection by the incumbent president among the members of a ruling oligarchy.

THE RADICAL REGIME, 1916–30

The Radical front was a coalition of heterogeneous social groups, which made it difficult to put through economic and social reforms without upsetting the front's equilibrium. Not surprisingly, Irigoyen preferred to concentrate on curing the political ills he had inherited from the conservative regime. The most urgent measure had to do with political patronage, which had been used by the conservatives to keep their candidates in office. Patronage had to be put to the service of the new party in power, and, thus, the Radical machine made itself virtually unbeatable at the polls in almost every province.

Political patronage and university reform

In other fields the Radical administration also showed concern for expanding its political base. Irigoyen achieved substantial rapport with the more moderate labour unions—a rapport expressed in a generally pro-labour policy. That policy had to be tempered after violent clashes occurred in the capital city during the general strike of January 1919, after which the military aligned itself with conservative interest groups. His administration supported organizations and movements among tenant farmers and also put through a university-reform plan.

Irigoyen's influence was a deciding factor in the election of his successor, Marcelo T. de Alvear (1922–28), who represented a safe choice. He, however, was not content with the restrictions that Irigoyen imposed upon him and became the reluctant leader of a conservative wing hostile to Irigoyen. In the elections of 1928, Irigoyen ran for a second term and was elected by a popular vote margin of two to one, establishing him as head of his party.

If Irigoyen was scarcely a revolutionist, his victory over the economic, social, and political elites of the country

had nonetheless earned him their strong enmity. His political machine, while an excellent mechanism for securing power, proved to be incapable of governing; its inadequacies were starkly revealed in the crisis of 1929, precipitated by the army, which expelled Irigoyen from office on Sept. 6, 1930. This marked the end of the constitutional continuity that had lasted for 68 years and also the end of the long period of economic expansion based on the export of raw materials, which had doubled between 1913 and 1928.

Behind the continuing upward trend lay a shift in economic power to foreign merchants and processors at the expense of the Argentine landowning class. Before 1914, these foreign interests had been concentrated mainly in the grain-growing sector; after 1920 they moved into the cattle-raising industry. Private investment still came mainly from Great Britain, which was also the main market for Argentine exports. The United States provided industrial and transportation equipment and was the government's main source of credit, but it had erected tariff and other barriers to the importation of Argentine goods, prompting Irigoyen to adopt an anti-U.S. and pro-British line.

Economic relations with the United States and Britain

THE CONSERVATIVE RESTORATION, 1930–43

The military coup that expelled Irigoyen installed Gen. José Félix Uriburu in the presidency (1930–32). A descendent of an old northern family, the conservative Uriburu leaned toward the Fascist ideas of the 1930s. His influence with the army, however, was not as great as that of Gen. Augustín Pedro Justo, a former minister of war under Alvear, who favoured a conservative reorientation of the country within constitutional limitations. The Radicals had been reorganized under the leadership of Alvear and had won an unexpected victory in trial elections held in the Province of Buenos Aires in April 1931. But the Radicals, under stringent restrictions, boycotted the national election of 1931. General Justo, with the backing of a coalition of conservatives, anti-Alvear Radicals, and independent Socialists and with only the limited use of electoral fraud, was elected by a large majority.

The new president (1932–38) faced a difficult economic situation. The Roca-Runciman Agreement with Great Britain (1933) guaranteed Argentina a fixed share in the British market for meat and ruled out tariffs on British cereal imports. In return, Argentina agreed to certain stipulations with regard to trade, currency exchange, and the preservation of Britain's commercial interests in the country. The treaty sharpened doubts of Argentines concerning the economic relationship between Argentina and Britain. Other unpopular reforms included a restructuring of the monetary system and the establishment of agencies to control exports. After 1935 the economic climate improved.

The election of 1937, in which the government retained its power, was marked by fraud and violence. The next president, Roberto M. Ortiz (1938–40), returned to more proper electoral procedures, calling for federal intervention in the province of Buenos Aires, where a corrupt conservative machine had been in control. Ortiz' poor health obliged him to resign, and his successor, Ramón S. Castillo (1940–43), restored the conservative coalition to power, gaining the support of General Justo.

At the outbreak of World War II, Argentina declared its neutrality and remained neutral even after the United States entered the conflict in 1941. This issue united all the opposition groups, and Castillo imposed a state of siege. In January 1943 a stabilizing influence disappeared with the death of General Justo. President Castillo fell from power in June 1943 in a coup led by his own minister of war, Gen. Pedro P. Ramírez.

THE PERÓN ERA, 1943–55

Transition period. The military government faced several urgent and difficult problems. Neutrality was advocated by some, and a choice between the restoration of a representative system and the installation of a long-term military dictatorship was also required. Gen. Arturo Rawson was made president but resigned after two days when his anti-conservative stance and his advocacy of the United Nations won no military support.

General Ramírez replaced Rawson as president (1943–44); neutrality was maintained, and opposition from all political groups (except the nationalist right wing and the Fascist sympathizers) increased. Press censorship and the dissolution of political parties reflected emergent authoritarianism. Under pressure from the United States, the regime broke off diplomatic relations with Germany. This deed was not favoured by the military officers, and Ramírez turned the presidency over to Gen. Edelmiro J. Farrell (1944–46). But international pressures grew, and to avert ruin Argentina had to face the return to representative democracy.

The search for a solution ended in the rise of Col. Juan Perón to the office of president. Since October 1943, when he had secured the minor job of running the labour department, Perón had been building a political empire based in the labour unions. He helped the unions win favourable settlements from employers; he pushed through a welfare program that provided vacations, retirement benefits, and severance pay. By 1945 he was also vice president and minister of war. His changes included giving autonomy to universities, reconstructing political parties (including the Communist Party, prohibited since 1936), and declaring war on Germany, thereby making it possible for Argentina to enter the United Nations. But with the return of political freedom came renewed opposition, culminating in a mass demonstration in Buenos Aires in September 1945. Emergency measures were enacted, Perón's military support gave way, and on October 9 he was removed from office and arrested. Neither the political nor the military opposition, however, could agree on what to do. Perón's adherents in the unions organized a strike that found enthusiastic support among the people. He was released on October 17, and his foes were forced to resign.

Rise of Perón

Perón in power. Perón campaigned for the presidency in the elections of 1946. He organized a Labour Party that was resisted by all of the old parties and by the major vested-interest groups. His victory, though narrow, gave him control of both houses of Congress and all the provincial governorships. Perón's political strategy and tactics were authoritarian and personalistic. Education and the courts were politically purified; a state of internal war was declared, to allow an expansion of executive authority; revenues were redistributed in favour of the workers; public services were nationalized; and urban and industrial areas were given preferential treatment over their rural counterparts.

Until 1949 Perón's economic policies were successful, largely because of the prosperity of exporters during and just after the war. As inflation increased and terms of trade worsened, however, it became difficult to finance imports of vital raw materials. The constitutional reform of 1949 allowed Perón to be reelected in 1951; his government took on a more conservative hue, hastened by the death of the President's wife, Eva Perón, in July 1952. She had been a powerful political figure in her own right, burnishing the regime's image of popular democracy. After 1952 Perón incurred the increasing hostility of the church and the students. His efforts to eliminate the influence of the church provoked disaffection in the officer corps, and in September 1955 he was overthrown by Gen. Eduardo Lonardi with support from the Córdoba garrison.

Influence of Eva Perón

ATTEMPTS TO RESTORE CONSTITUTIONALISM, 1955–66

General Lonardi was acting president, but he soon gave way to Gen. Pedro Eugenio Aramburu, who became president in November 1955. The new administration (1955–58) was a military dictatorship seeking to restore constitutional government. Taking a fiercely anti-Peronista stance, it dissolved Perón's old party and placed the labour unions under state administration. The Peronistas wielded considerable influence on the factions that were competing for power and in 1958 supported Arturo Frondizi, a radical leader who promised to readmit them to political life in return for their support. Frondizi won the presidency and majorities in both houses of Congress.

The new president (1958–62) showed a keen interest in reviving the flow of foreign investment. A currency devaluation that favoured exporters and foreign investors, however, had adverse effects on the middle and lower classes.

The campaign against inflation brought restrictions on credit, increasing the difficulties of industry, and Frondizi had to use the military to uphold his unpopular policies.

In March 1962 the reorganized Peronistas gained control of important districts, among them the province of Buenos Aires. The armed forces withdrew support from Frondizi, dissolved Congress, and set up a government in the name of José María Guido, president pro tempore of the Senate. Guido's 18-month administration was one of confusion as two military factions fought for control. The Colorados (Reds) sought a dictatorship that would deal strongly with the Peronistas and extreme leftists. The Azules (Blues), who prevailed, favoured a constitutional government by a coalition including the Peronistas, though the latter were to be confined to a weaker role than that indicated by their voting strength.

The elections of July 1963 resulted in victory for Arturo Illia, the candidate of the People's Radical Civic Union. President Illia (1963–66) inherited Frondizi's economic problems, although the drastic reorientation of the economy had begun to show signs of success. Illia tried without success to split the resurgent Peronistas, who now controlled the labour unions, from their exiled leader. Antagonized, the Peronistas supported a coup in June 1966 that brought to power Gen. Juan Carlos Onganía, a former Azules leader and commander in chief of the army.

Onganía takes power

GOVERNMENT BY THE ARMED FORCES

Adalbert Krieger Vasena, minister of economy and labour, strove for economic stability through yet another currency devaluation and also undertook programs in electric power, steel, roads, and housing. In May 1969, disturbances and riots in the cities of Corrientes, Rosario, and particularly Córdoba rose out of student and labour conflicts; these incidents, later known as the Cordobazo, were identified as resentment toward Krieger Vasena's economic policies. Krieger Vasena was removed, but the Onganía administration was unable to agree on an alternative economic policy, and the Cordobazo decisively affected the political climate. Underground activities were organized by a Trotskyite group, the People's Revolutionary Army (Ejército Revolucionario del Pueblo; ERP), and by Peronista groups. In 1970 one of these Peronista organizations, the Montoneros, besides decimating the moderate Peronista union leadership, captured and killed former president Aramburu, who had been organizing a movement for a return to constitutional rule. The armed forces overthrew the Onganía government in June 1970. Gen. Roberto Marcelo Levingston replaced Onganía; but the return of inflation and his political ambitions caused his overthrow, and he was replaced in March 1971 by Gen. Alejandro Agustín Lanusse.

Perón supported the Peronista underground but also used other means in a new bid for power. He maintained a formal alliance with the Frondizi followers, but the cornerstone of his strategy was an understanding with the largest non-Peronista party, the Radicals. Mindful of the vested interests, he purged his economic proposals of any motives that could alarm the propertied classes. The military government prevented Perón's own candidacy but could not stop the electoral victory of the Peronista coalition, the Justicialist Liberation Front (Frente Justicialista de Liberación), in March 1973.

THE RETURN OF PERONISM

The newly elected president, Héctor J. Cámpora, took office in May 1973. Under his administration the Peronista left wing acquired considerable influence. The final return of Perón in June was the occasion of a battle between right and left. The union leadership and Perón's private secretary, José López Rega (appointed minister of social welfare by Cámpora), launched a violent antileftist campaign that had the discreet support of Perón himself. In July Cámpora resigned, and new elections were called. An interim president, Raúl Lastiri, began a purge of leftist influences in the government.

Perón's second presidency. Perón was elected president on a ticket that included his wife, María Estela (called Isabel) Martínez de Perón, as vice president and took office in October 1973. He continued the campaign against the left, and in May 1974 the victims of the purge acknowledged the break with their leader and passed into (still legal) opposition. Underground activities again became evident, and a right-wing organization, the Argentine Anti-Communist Alliance (Alianza Anticomunista Argentina; AAA), suspected by many to be close to the police and intelligence branches of the administration, began to take a heavy toll of political, student, and union leaders.

Split in the Peronista ranks

Perón's economic policies after May 1973 combined monetary stabilization and rigid control of prices and wages with a gradual redistribution of income in favour of wage earners and a limitation on profits of the agrarian exporters. By 1974, however, the world petroleum crisis caused Europe to reduce its substantial imports of Argentine meat, and the balance of payments suffered.

Perón's legacy. When Perón died on July 1, 1974, he left to his widow, who succeeded him, a deeply compromised inheritance. The transition of power was smooth, however, and President Martínez de Perón, in close alliance with López Rega, became even more inflexibly oriented toward the right. Violence reached new heights. López Rega, who used the rightist crusade to consolidate his power base, favoured labour and army leaders who were committed to support him. This course created hostility among union, political, and military leaders. In the autumn of 1975 a drastic devaluation and a steep drop in real wages were proposed. Martínez de Perón was persuaded to dismiss López Rega, but the unrest deepened. In December there was an uprising by the air force, and on March 24, 1976, military officers deposed the President and took over the government. (T.H.D./Ed.)

THE RETURN OF MILITARY GOVERNMENT

The Videla regime. A three-man military junta filled the presidency five days after the coup with Lieut. Gen. Jorge Rafaél Videla, who served two consecutive terms in office (1976–81). During Videla's regime the economy showed some improvement. Inflation dropped from about 600 percent in 1976 to 138 percent in 1982—still the highest in the world for that year. The balance of foreign trade also improved. This economic relief, however, was accompanied by continued political violence. Thousands of citizens were killed or imprisoned or disappeared. The Argentine government, which maintained that it was fighting a civil war, was subject to much criticism at home and abroad for civil rights violations.

Viola and Galtieri. Videla was succeeded on March 29, 1981, by Gen. Roberto Eduardo Viola, who pursued a policy of moderation. On Dec. 21, 1981, the government announced that Viola was stepping down for reasons of health; he was succeeded the next day by Lieut. Gen. Leopoldo Galtieri. Galtieri was occupied by an economy again on the downswing, by increased civil opposition to military rule, and by Argentina's claims to islands in the Atlantic. The government had refused to accept a decision of the International Court of Justice in 1977 awarding the three Beagle Channel islands to Chile. The matter again went into negotiation, under Vatican auspices from 1979, and was resolved in 1984 when Chile was awarded sovereignty of the islands. In February 1982 Argentina increased pressure on the United Kingdom to relinquish the Falkland Islands. With popular support at home, Argentine troops landed on the Falklands and South Georgia island on April 2 and overcame the U.K. Royal Marines stationed at Port Stanley. For the next three weeks, while a British naval force sailed to the Falklands, negotiations between the two belligerents failed to reach a solution. British forces retook South Georgia on April 30, and the Argentine military governor surrendered the Falklands to them on June 14.

Falkland Islands war

Galtieri resigned as commander in chief of the army and president on June 17, 1982. Maj. Gen. Reynaldo Bignone was installed as president on July 1. The members of the junta representing the air force and the navy resigned in protest over Bignone's appointment, but the junta was reconstituted on September 10. Under Bignone, political parties were allowed to resume activities, and general elections were announced. The Peronista party delayed

choosing a presidential candidate and thus lost ground to the Radical Civic Union, led by Raúl Alfonsín, a civilian lawyer; Alfonsín won the election on Oct. 30, 1983, and the Radicals gained the majority in the National Congress, followed by the Peronistas.

Alfonsín's democracy. Soon after his inauguration on Dec. 10, 1983, Alfonsín, in a reversal of legislation passed under Bignone, announced plans to prosecute several members of the former military government, including the former presidents Videla, Viola, and Galtieri. He also repealed a law granting amnesty to those accused of crimes and human rights violations during what came to be called the "dirty war." Hundreds of military personnel were ordered to stand trial. In the trial of nine former junta members in 1985, five were convicted, including Videla and Viola. Although Galtieri was acquitted in that trial, he was convicted in 1986, along with two other officers, of incompetency in the Falkland Islands war. Rebellions broke out within the armed forces in protest over the prosecutions but most of the armed forces stayed loyal. Massive rallies voiced approval of Alfonsín's democracy, and support poured in from the international community.

Alfonsín launched an austerity program, the Austral Plan, which adopted a new currency (the austral) and invoked wage and price controls and currency devaluations. The measures initially brought a generally downward trend to inflation and restored the confidence of international bankers, who allowed Argentina to refinance and reschedule its foreign debts, the growth of which had reached crisis proportions. The inflation rate began to rise again, however, reaching almost 388 percent at the end of 1988, and the austral began a steady decline against the U.S. dollar. The worsening state of the economy and the Alfonsín government's inability to handle it contributed to the defeat of the Radical presidential candidate, Eduardo Angeloz, in May 1989. (Alfonsín was constitutionally ineligible to succeed himself.) Carlos Saúl Menem, a Peronista, led his party to victory in the presidential and congressional elections. Throughout his campaign Menem had openly cultivated an image recalling his party's founder, and it was his appeal to the poor and working classes, the traditional supporters of Peronism, that clinched his victory.

Election of Menem

With the economy crumbling around him, Alfonsín resigned five months early, and Menem officially took over on July 8. Menem's plan was a free-market economy, with lower tariffs based on a wage–price pact between labour, business, and government and the privatization of inefficient government-owned companies. To help carry it out, he enlisted the aid of former top-level executives from Bunge y Born, one of Argentina's leading corporations.

The continuing discontent of the military over wages, equipment, and the trials of those of their number accused of human rights violations during the dirty war had manifested itself in more rebellions in the last months of Alfonsín's tenure. After a bloody attack at La Tablada barracks outside Buenos Aires, Alfonsín initiated a huge military investment program. Menem in turn sought to quell the military's discontent and to win their support in a time of economic emergency by pardoning those accused of human rights violations, a move that was strongly criticized. Former president Galtieri was also pardoned. In October, Argentina and Great Britain formally agreed to end the hostilities that began with the Falklands war, and they reestablished consular ties. (Ed.)

For later developments in the history of Argentina, see the *Britannica Book of the Year* section in the BRITANNICA WORLD DATA ANNUAL.

For coverage of related topics in the *Macropædia* and *Micropædia,* see the *Propædia,* sections 951, 964, 966, and 974, and the *Index.*

BIBLIOGRAPHY

Physical and human geography: (*The land and the people*): General descriptive information is available in *The South American Handbook* (annual); *El País de los Argentinos* (1977), published by the Centro Editor de América Latina; and JAMES D. RUDOLPH (ed.), *Argentina, a Country Study,* 3rd ed. (1986). Statistical information can be found in *Anuario estadístico de la República Argentina* (annual). Basic geographical information is discussed in PRESTON E. JAMES, C.W. MINKEL, and EILEEN W. JAMES, *Latin America,* 5th ed. (1986); FEDERICO A. DAUS, *Geografía y unidad Argentina,* 2nd ed. (1978); and FRANCISCO DE APARICIO and HORACIO A. DIFRIERI (eds.), *La Argentina: suma de geografía,* 9 vol. (1958–63). Geographical distribution of animals and plants is discussed in E.J. FITTKAU *et al.*, *Biogeography and Ecology in South America,* 2 vol. (1968–69); an early work on the animals of Argentina is W.H. HUDSON, *The Naturalist in La Plata* (1892, reissued 1968). Patterns of settlement are the subject of CARL-CHRISTOPH LISS, *Die Besiedlung und Landnutzung Ostpatagoniens unter besonderer Berücksichtigung der Schafestancien* (1979); ROBERT C. EIDT, *Pioneer Settlement in Northeast Argentina* (1977); MARK JEFFERSON, *Peopling the Argentine Pampa* (1926, reprinted 1971); and RICHARD W. SLATTA, *Gauchos and the Vanishing Frontier* (1983).

(*The economy and administrative and social conditions*): Economic conditions are documented by JONATHAN C. BROWN, *A Socioeconomic History of Argentina, 1776–1860* (1979); LAURA RANDALL, *An Economic History of Argentina in the Twentieth Century* (1978); ROBERTO CORTÉS CONDE and EZEQUIEL GALLO, *La formación de la Argentina moderna,* 2nd ed. (1973); and PIERRE DENIS, *The Argentine Republic: Its Development and Progress* (1922, reprinted 1976; originally published in French, 1920). Agricultural economy is the focus of JAMES R. SCOBIE, *Revolution on the Pampas: A Social History of Argentine Wheat, 1860–1910* (1964); and ROBERTO SCHOPFLOCHER, *Historia de la colonización agrícola en Argentina* (1955). See also *Argentina: Economic Memorandum* (1985), a World Bank country study. Economic policy is discussed in JUAN E. CORRADI, *The Fitful Republic: Economy, Society, and Politics in Argentina* (1985); GARY W. WYNIA, *Argentina in the Postwar Era: Politics and Economic Policy Making in a Divided Society* (1978); and GUIDO DI TELLA and D.C.M. PLATT (eds.), *The Political Economy of Argentina, 1880–1946* (1986).

(*Cultural life*): Works on Argentine literature include JOSÉ ALBERTO SANTIAGO (comp.), *Antología de la poesia argentina* (1973); RICARDO ROJAS, *Historia de la literatura argentina: ensayo filosófico sobre la evolución de la cultura en el Plata,* new ed., 9 vol. (1960); and DAVID WILLIAM FOSTER (comp.), *Argentine Literature: A Research Guide,* 2nd ed., rev. and enl. (1982). Art and music are covered in JOSÉ LEÓN PAGANO, *El arte de los argentinos,* rev. ed. (1981); and VICENTE GESUALDO, *Historia de la música en la Argentina,* 2nd ed., 3 vol. (1978).

History: An excellent summary is THOMAS E. SKIDMORE and PETER H. SMITH, *Modern Latin America,* ch. 3, "Argentina: From Prosperity to Deadlock," pp. 70–112 (1984). Broader treatments can be found in ACADEMIA NACIONAL DE LA HISTORIA, *Historia de la nación argentina: desde los origenes hasta la organización definitiva en 1862,* 3rd ed., 11 vol. in 15 (1961–63), and *Historia argentina contemporánea, 1862–1930,* 4 vol. (1965–67); EDUARDO CRAWLEY, *A House Divided: Argentina, 1880–1980* (1984); JOHN LYNCH, *Argentine Dictator: Juan Manuel De Rosas, 1829–1852* (1981); DAVID ROCK (ed.), *Argentina, 1516–1982: From Spanish Colonization to the Falklands War* (1985); JAMES R. SCOBIE, *Argentina: A City and a Nation,* 2nd ed. (1971); and IONE S. WRIGHT and LISA M. NEKHOM, *Historical Dictionary of Argentina* (1978). The history of Argentine politics is detailed in TULIO HALPERÍN-DONGHI, *Politics, Economics and Society in Argentina in the Revolutionary Period,* trans. from Spanish (1975); DAVID ROCK, *Politics in Argentina, 1890–1930: The Rise and Fall of Radicalism* (1975); JOSÉ LUIS ROMERO, *A History of Argentine Political Thought* (1963, reissued 1968; originally published in Spanish, 3rd ed., 1959); ROBERT A. POTASH, *The Army & Politics in Argentina,* 2 vol. (1969–80); MARK FALCOFF and RONALD H. DOLKART (eds.), *Prologue to Perón: Argentina in Depression and War, 1930–1943* (1975); GUIDO DI TELLA, *Argentina Under Perón, 1973–76: The Nation's Experience with a Labour-Based Government* (1983); and FREDERICK C. TURNER and JOSÉ ENRIQUE MIGUENS (eds.), *Juan Perón and the Reshaping of Argentina* (1983); JOSEPH A. PAGE, *Perón, a Biography* (1983); and NICHOLAS FRASER and MARYSA NAVARRO, *Eva Perón* (1980). Works on the Falkland Islands war of 1982 include RAPHAEL PERL, *The Falkland Islands Dispute in International Law and Politics: A Documentary Sourcebook* (1983); ALEJANDRO DABAT and LUIS LORENZANO, *Argentina, the Malvinas, and the End of Military Rule* (1984; originally published in Spanish, 1982); and MAX HASTINGS and SIMON JENKINS, *The Battle for the Falklands* (1983).

(R.C.Ei./T.H.O./Ed.)

Aristotle and Aristotelianism

Aristotle, more than any other thinker, determined the orientation and the content of Western intellectual history. He was the author of a philosophical and scientific system that through the centuries became the support and vehicle for both medieval Christian and Islāmic scholastic thought: until the end of the 17th century Western culture was Aristotelian. And even after the intellectual revolutions of centuries to follow Aristotelian concepts and ideas remained embedded in Western thinking.

Aristotle's intellectual range was vast, covering most of the sciences and many of the arts. He worked in physics, chemistry, biology, zoology, and botany; in psychology, political theory, and ethics; in logic and metaphysics; in history, literary theory, and rhetoric. His greatest achievements were in two unrelated areas: he invented the study of formal logic, devising for it a finished system, known as Aristotelian syllogistic, that for centuries was regarded as the sum of logic; and he pioneered the study of zoology, both observational and theoretical, in which his work was not surpassed until the 19th century.

Even though Aristotle's zoology is now out-of-date and his thought in the other natural sciences has long been left behind, his importance as a scientist is unparalleled. But it is now of purely historical importance: he, like other scientists of the past, is not read by his successors. As a philosopher Aristotle is equally outstanding. And here he remains more than a museum piece. Although his syllogistic is now recognized to be only a small part of formal logic, his writings in ethical and political theory as well as in metaphysics and in the philosophy of science are read and argued over by modern philosophers. Aristotle's historical importance is second to none, and his work remains a powerful component in current philosophical debate.

This article deals with the man, his achievements, and the Aristotelian tradition. For treatment of Aristotelianism in the full context of Western philosophy, see PHILOSOPHY, THE HISTORY OF WESTERN, and PHILOSOPHICAL SCHOOLS AND DOCTRINES.

For coverage of related topics in the *Macropædia* and *Micropædia,* see the *Propædia,* Part Ten, Division V, especially Section 51.

The article is divided into the following sections:

THE LIFE OF ARISTOTLE

Early years and background

Aristotle was born in the summer of 384 BC in the small Greek township of Stagira (or Stagirus, or Stageirus), on the Chalcidic peninsula of Macedonia, in northern Greece. (For this reason Aristotle is also known as the "Stagirite.") His father, Nicomachus, was court physician to Amyntas III, king of Macedonia, father of Philip II, and grandfather of Alexander the Great. As a doctor's son, Aristotle was heir to a scientific tradition some 200 years old. The case histories contained in the *Epidemics* of Hippocrates, the father of Greek medicine, may have introduced him at an early age to the concepts and practices of Greek medicine and biology. As a physician, Nicomachus was a member of the guild of the Asclepiads, the so-called sons of Asclepius, the legendary founder and god of medicine.

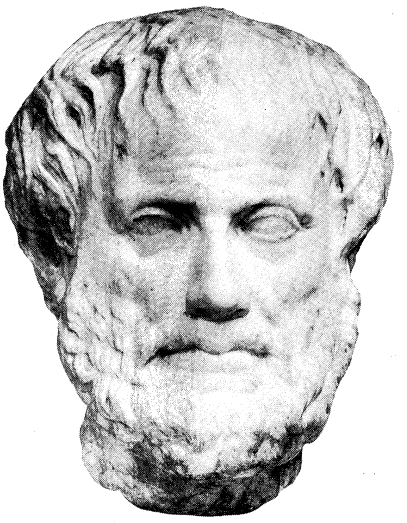
Because medicine was a traditional occupation in certain families, being handed down from father to son, Aristotle in all likelihood learned at home the fundamentals of that practical skill he was afterward to display in his biological researches. Had he been a medical student he would have undergone a rigorous and varied training: he would have studied the role in therapy of diet, drugs, and exercise; he would have learned how to check the flow of blood, apply bandages, fit splints to broken limbs, reset dislocations, and make poultices of flour, oil, and wine. Such, at least, were the skills of the trained physician of his time. It is not known for certain that Aristotle actually acquired these skills; it is known that medicine and its history were later studied in the Lyceum, Aristotle's own institute in Athens, and that later, in a snobbish vein, he considered a man sufficiently educated if he knew the theory of medicine without having gained experience practicing it.

This early connection with medicine and with the rough-living Macedonian court largely explains both the predominantly biological cast of Aristotle's philosophical thought and the intense dislike of princes and courts to which he more than once gave expression.

First period: in the Academy at Athens. While Aristotle was still a youth, his father died, and the young man became a ward of Proxenus, probably a relative of his father. He was sent to the Academy of Plato at Athens in 367 and remained there for 20 years. These years formed the first of three main periods in Aristotle's intellectual development, years dominated by the formative influence of Plato and his colleagues in the Academy. Aristotle doubtless interested himself in the whole range of the Academy's activities. It is known that he devoted some time to the study of rhetoric, and he wrote and spoke for the Academy in its battles against the rival school of Isocrates.

Death of Plato: Aristotle's departure from Athens

After Plato's death in 348/347 his nephew Speusippus was named as head of the Academy. Aristotle shortly thereafter left Athens—in disgust, it is sometimes claimed, at not being appointed Plato's successor. This interpretation of his motive, however, lacks foundation, for evidence suggests that he was ineligible to be the school's head because of his status as a resident alien who could not hold property legally. It is more likely that his departure from Athens may have been linked with an anti-Macedonian feeling that arose in Athens after Philip had sacked the Greek city-



Aristotle, marble bust with a restored nose, Roman copy of a Greek original, last quarter of the 4th century BC. In the Kunsthistorisches Museum, Vienna.

By courtesy of the Kunsthistorisches Museum, Vienna

state of Olynthus in 348. Aristotle's 12-year absence from Athens nevertheless indicates that he valued more the circle of friends who accompanied him on his travels—chief among them Theophrastus of Eresus, his pupil, colleague, and eventual successor as head of the Lyceum—than he did his membership in the Platonic Academy.

Second period: his travels. With him went another Academy member of note, Xenocrates of Chalcedon, whose lethargy became the target of Plato's ridicule. Plato reportedly contrasted it with Aristotle's more energetic manner: "The one needs a spur, the other a bridle .... See what an ass I am training to compete with what a horse." The distinctive characters of the two men, however, seem to have integrated well in establishing a new academy on the Asian side of the Aegean at the newly built town of Assus.

Assus

At Assus, Hermeias of Atarneus, a Greek soldier of fortune, had first acquired fiscal and then political control of northwestern Asia Minor, as a vassal of Persian overlords. After a visit to the Athenian Academy he invited two of Plato's graduates to set up a small branch to help spread Greek rule as well as Greek philosophy to Asian soil. Aristotle came to this new intellectual centre. To this period may belong the first 12 chapters of Book 7 of Aristotle's *Politics.* There he sketches the connection between philosophy and politics, namely, that the highest purpose of a city-state (*polis*) is to secure the conditions in which those who are capable of it can live the philosophical life. Such a life, however, lies only within the capacity of the Greeks, whose superiority qualifies them to employ the non-Greek tribal peoples as serfs or slaves for the performance of all menial labour. Thus, citizenship and service in the armed forces are considered to be the exclusive rights and duties of the Greeks. Aristotle's espousal of an enlightened oligarchy, nonetheless, actually constituted an advance over the political concepts flourishing at the time and it should be viewed in its context as a positive development in the establishment of the noble civilization created by the Greeks.

At about the same time, Aristotle composed the work, now lost, *On Kingship,* in which he clearly distinguishes the function of the philosopher from that of the king. He alters Plato's dictum—for the better, it is said—by teaching that it is

> ... not merely unnecessary for a king to be a philosopher, but even a disadvantage. Rather a king should take the advice of true philosophers. Then he would fill his reign with good deeds, not with good words.

Aristotle thus strove to assure the independent role of the philosopher.

Aristotle was on good terms with his patron, Hermeias, and married his niece, Pythias. She bore Aristotle a daughter, whom he called by her mother's name. In the *Politics,* Aristotle prescribed the ideal ages for marriage—37 for the husband and 18 for the wife. Because Aristotle was himself 37 at this time, it is tempting to guess that Pythias was 18. It is also possible that their own marital relations are reflected in his further, somewhat cryptic, observation: "As for adultery, let it be held disgraceful for any man or woman to be found in any way unfaithful once they are married and call each other husband and wife." In his will Aristotle ordered that "Wherever they bury me, there the bones of Pythias shall be laid, in accordance with her own instructions." Pythias did not live long, however; and after her death Aristotle chose another companion, Herpyllis (whether concubine or wife is uncertain), by whom he then had a son, Nicomachus. She outlived Aristotle, and he made ample and considerate provision for her in his will "in recognition of the steady affection she has shown me."

Mytilene

After three years at the young Assus Academy, Aristotle moved to the nearby island of Lesbos and settled in Mytilene, the capital city. With his friend Theophrastus, a native of that island, he established a philosophical circle patterned after the Athenian Academy. There his centre of interest shifted to biology, in which he undertook pioneering investigations. (The landlocked lagoon of Pyrrha in the centre of Lesbos has been identified as one of his favourite haunts.) He appears to have felt it necessary to justify this new attention to biology by rejecting the arguments that had classed it as an inferior, unattractive study. In his biological researches he focused on a new type of causation, namely teleological. Teleological causation has to do with the aim, or end, of nature, a type that is distinct from mechanical causation but one that is, nonetheless, operative in the inorganic sphere. According to Aristotle, natural organisms—plants and animals—have natural ends or goals, and their structure and development can only be fully explained when these goals are understood. To admit the existence of such ends, or aims, in nature is to argue teleologically (Greek *telos,* "an end") or to admit the idea of a final cause (Latin *finis,* "end"). Teleology, and theory in general, is important in Aristotle's biology; but it is always, in principle at least, subordinate to observation. Thus, confessing his ignorance of the mode of generation of bees, Aristotle wrote in his treatise *On the Generation of Animals:*

> The facts have not yet been sufficiently established. If ever they are, then credit must be given to observation rather than to theories, and to theories only insofar as they are confirmed by the observed facts.

Associated with his researches into plant and animal life were his reflections on the relation of the soul to the body. As revealed by his tract *On the Soul,* Aristotle distanced himself from the Platonic conception of the soul as an independently existing substance that is only temporarily resident in the body. With greater emphasis on the positive value of material existence, he suggested instead that the soul is the vital principle essentially united with the body to form the individual person. With some acknowledgment to Plato, he then proceeded to define the soul as the form of the body and the body as the matter of the soul.

Alexander the Great

In late 343 or early 342 Aristotle, at about the age of 42, was invited by Philip II of Macedon to his capital at Pella to tutor his 13-year-old son, Alexander. As the leading intellectual figure in Greece, Aristotle was commissioned to prepare Alexander for his future role as a military leader. As it turned out, Alexander was to dominate the Greek world and defend it against the Persian Empire. Using the model of the epic Greek hero, as in Homer's *Iliad,* Aristotle attempted to form Alexander as an embodiment of the classical valour of an Ajax or Achilles enlightened by the latest achievement of Greek civilization, philosophy. With his firm conviction of the superiority of Greeks over foreigners, he instructed Alexander to dominate the barbarians—*i.e.,* non-Greeks—and to hold them in servility by refraining from any physical intermixture with them. Despite this advice, however, Alexander later became committed to intermarriage; he chose a wife from the Persian nobility and forced his high-ranking officers (and encouraged his troops) to do likewise.

In other ways too the influence that Aristotle had on Alexander was negligible. Although later, on his return to Athens, Aristotle enjoyed considerable political and economic support from the Macedonians and perhaps received assistance in the organization of his biological researches, it is not likely—as some have held—that Alexander collected and dispatched to Aristotle specimens of rare animals from Persia and India; in fact, Alexander's first penetration of the valley of the Indus did not occur until 328/327, less than six years before Aristotle's death. Indeed, the relation between the two was embittered by the execution of Aristotle's nephew, the historian Callisthenes of Olynthus, who was charged with treason while accompanying Alexander to Persia early in 328 in order to write a chronicle of the campaign. It has even been reported that Alexander meditated revenge on Aristotle himself because he was a blood relative of the victim. But Alexander was diverted by his preoccupation with the invasion of India. Clearly, in matters of political ideology, a gulf separated Aristotle and Alexander. Aristotle showed no awareness of the fundamental changes that Alexander's conquests were bringing to the Greek world; indeed, he was opposed in principle to Alexander's imperial policy because it diminished the importance of the city-state. On the other hand, Alexander gratified his tutor by rebuilding the town of Stagira, Aristotle's birthplace, which Philip II had destroyed earlier.

After three years at the Macedonian court, Aristotle withdrew and returned to his paternal property at Stagira (*c.* 339). There he continued the associations of his philosophical circle, which still included Theophrastus and other pupils of Plato.

Third period: founding and directing of the Lyceum. Aristotle remained in Stagira until 335, when, nearing 50 years of age, he once again returned to Athens. At this time the presidency of the Academy became vacant by the death of Speusippus, and Xenocrates of Chalcedon, his old associate in biological research, was elected to the post. Although Aristotle appears never to have wholly severed his links with the Academy, he nonetheless opened, in 335, a rival institution in the Lyceum, a gymnasium attached to the temple of Apollo Lyceus, situated in a grove just outside Athens. The place had for some time been frequented by other teachers—Plato even mentions it as having been one of Socrates' haunts—and the name of the temple came to be applied to Aristotle's school in particular. But it was probably only after Aristotle's death that the school, under Theophrastus, acquired extensive property. From the fact that his instruction was given in the *peripatos,* or covered walkway, of the gymnasium, the school has derived its name of Peripatetic. Informal as the school may have been under Aristotle, it was very important to him because, by coordinating the work of a number of scholars, he was able for the next 12 years to organize it as a centre for speculation and research in every field of inquiry and to give lectures on a wide range of scientific and philosophical questions. The chief difference between the new school and the Academy was that the scientific interests of the Platonists centred on mathematics whereas the main contributions of the Lyceum lay in biology and history.

Athens: the Lyceum

On the death of Alexander the Great in 323 a brief but vigorous anti-Macedonian agitation broke out in Athens. Aristotle, who had long-standing Macedonian connections and was a friend of Antipater, the Macedonian regent of Athens, felt himself in danger. He therefore left Athens and withdrew to his mother's estates in Chalcis on the island of Euboea. There he died in the following year from a stomach illness at the age of 62 or 63. It was reported that he abandoned Athens in order to save the Athenians from sinning twice against philosophy (referring to Socrates as the earlier victim).

Death

PERSONALITY, CHARACTER, AND PHILOSOPHICAL STANCE

The features of Aristotle, familiar from busts and engravings, appear handsome and refined. An ancient tradition, possibly from an unfriendly source, says, however, that Aristotle had spindleshanks and small eyes and that he spoke with a lisp. In compensation for these physical defects, he was notably well dressed. His cloak and sandals were of the best quality and he sported rings. Presumably he was rich, with large family holdings at Stagira. One use that he made of his money was to collect books. Plato, with a touch of contempt for Aristotle's devotion to reading and perhaps not without some envy of his affluence, called him "the reader." Aristotle was an intellectual but not devoid of passion. A story is told of Plato giving a reading of his *Phaedo,* a purported record of Socrates' last day. The dialogue is moving and solemn. As Plato was reading, however, his audience gradually melted away. In the end, Aristotle alone was left. Probably fictitious, the anecdote was invented to express a truth: Aristotle was, in fact, spellbound by the Socratic doctrine of immortality as expounded by Plato. It not only interested him intellectually but also absorbed him emotionally. His earliest works, dialogues written when he was still a member of the Academy (now lost except for some fragments), were in part concerned with thoughts of the next world and the worthlessness of this one.

The anecdotes related of him reveal him as a kindly, affectionate character, and they show barely any trace of the self-importance that some scholars think they can detect in his works. His will, which has been preserved, exhibits the same kindly traits; he makes references to his happy family life and takes solicitous care of his children, as well as his servants.

This personal happiness is reflected in *On Philosophy,* perhaps the last of his strictly literary works. After writing this work, which he completed in around 348, he devoted his energies to research, teaching, and the writing of more technical treatises. The greatness of *On Philosophy,* which survives only in fragments, is evident in its influence on the thought of later antiquity; perhaps more than any other single work it established philosophy as a profession. In the extant part, Aristotle defines the specific role of the philosopher. Dividing the historical development of civilization into five main stages, Aristotle sees the emergence of philosophy as its culmination. First, men are compelled to devote themselves to the creation of the necessities because without them they could not survive. Next come the arts that refine life and then the discovery of the art of politics, the prerequisite of the good life as Aristotle conceived it. To these necessities and refinements of life is added the knowledge of their proper use in the fourth stage. Only with the emergence of the well-regulated state comes the leisure for intellectual adventure, used at first for the study of the material causes of existing things. Finally comes the shift from natural to divine philosophy, when the mind lifts itself above the material world and grasps the formal and final causes of things, realizing the intelligible aspect of reality and the purpose that informs all change.

Concept of philosophy

This divine philosophy gave its attention to the astral gods. Aristotle had experienced in Athens the long intellectual struggle to discover perfect order in the heavens. He had learned that perfection was not to be confined to the mathematical abstractions, to which Plato had at first directed the attention of his pupils, but had come to recognize that the visible heavens themselves could be accepted as the embodiment of the divine. With the declaration of this intimacy between the deities and the work of their hands in the material universe, Aristotle issued his manifesto, which is an optimistic affirmation of the values of this world; simultaneously he rejected the Platonic doctrine that the soul is imprisoned in the body and in need of struggling free from the bonds of matter. It was by this stroke that Aristotle established his own identity in the history of thought.

WRITINGS

Aristotle's writings fall into two groups: the first consists of works published by Aristotle but now lost; the second of works not published by Aristotle and, in fact, not intended for publication but collected and preserved by others. In the first group are included (1) the writings that Aristotle himself termed "exoteric," or popular—that is, those written in dialogue or other current literary forms

and meant for the general reading public—and (2) those that he termed "hypomnematic," or notes to aid the memory, and collections of materials for further work. Of these, only fragments are extant. Finally, the writings that generally have survived, termed "acroamatic," or treatises (*logoi, methodoi, pragmateiai*), were meant for use in Aristotle's school and were written in a concise and individualistic style. In later antiquity Aristotle's writings filled several hundred rolls; today the surviving 30 works fill some 2,000 printed pages. Three ancient catalogs list a total of more than 170 separate works by Aristotle, a figure corroborated by references and lists of titles in the extant treatises as well as by a number of citations and paraphrases in early commentators. Cicero must have been alluding to Aristotle's popular dialogues when he described in the *Academica* "the suave style of Aristotle . . . . A river of gold." The extant works contain several passages of polished prose, but for the most part their style is clipped.

Lost works published by Aristotle. The lost popular works include poetry and letters as well as essays and dialogues in the Platonic manner. Several problems have confronted scholars in their attempts to reconstitute these lost popular works. The lost dialogues, for example, appear to diverge widely from the doctrines of the surviving treatises. Indeed, they appear to outdo Plato in his own teaching. Thus, what is known of Aristotle's dialogue *Eudemus,* or *On the Soul,* compares the relation of the soul to the body with an unnatural union, like that of the torture that the Tyrrhenian pirates inflicted on their prisoners by binding each of them to a corpse. Inasmuch as Aristotle in his extant treatises criticized his Platonist friends for making soul and body enemies, Alexander of Aphrodisias, an authoritative Aristotelian commentator of the late 2nd century AD, raised the question whether he expressed "two truths," one "exoteric" for public consumption, the other "esoteric" and reserved for his students in the Lyceum. The present consensus of scholars is that Aristotle's popular writings generally derived from the early stage of his intellectual development during his time in Plato's Academy: they represent not his "public" but his juvenile thoughts.

Popular dialogues

Chief among the lost works are: *Eudemus,* in the tradition of Plato's *Phaedo; On Philosophy,* a type of philosophical program containing themes to be developed later in his *Metaphysics;* the *Protrepticus,* or exhortation to the life of philosophy; *Gryllus,* or *On Rhetoric; On Justice,* expressing nascent themes of his *Politics;* and *On Ideas,* which criticizes Plato's theory of Forms.

Extant works. The works that have been preserved derive from manuscripts left by Aristotle on his death; many of them were probably used by him as lecture notes. These are the "esoteric" writings of a concentrated, academic nature intended for the ears of the initiates. From classical antiquity romanticized accounts circulated of the way these manuscripts were preserved; *e.g.,* in Plutarch's *Sulla,* chapter 26; and in Strabo's *Geography* 13:54. According to these versions, Aristotle's and Theophrastus' notes had been bequeathed to an old colleague, Neleus of Scepsis, whose heirs apparently were not interested in the contents but, in order to prevent them from being confiscated for the library of the kings of Pergamum, hid them in a cellar in Scepsis. Long afterward, in the 1st century BC, the descendants sold them to Apellicon of Teos, a philosopher, who brought them back to Athens. When Athens was conquered by Sulla in 86 BC, he appropriated the books and sent them to Rome, where they were purchased by Tyrannion the grammarian. The manuscripts suffered further maltreatment, first at the hands of copyists, then through subjective restoration of worm-eaten passages and systematic ordering irrespective of actual chronology, until Andronicus of Rhodes, the last head of the Lyceum, acquired the copies and edited and published them about 60 BC.

The story is improbable. It is difficult to imagine that the Lyceum would have allowed the manuscripts of its founder to have been so carelessly looked after. And it is now known that the "esoteric" writings were not wholly ignored in the two centuries after Theophrastus' death. It is true, nevertheless, that the Andronicus edition is the first publication of Aristotle's works, even if the story of the edition's appearance was spread by Andronicus to emphasize its novelty. The form, titles, and order of Aristotle's texts that are studied today were given to them by Andronicus almost three centuries after the philosopher's death, and the long history of commentary upon them began at this stage.

These facts have affected the interpretation of Aristotle. The books of Aristotle that are known today were, in effect, never edited by him. Thus, for example, Aristotle is not the author of the work called *Metaphysics;* rather, he wrote a dozen little treatises: on the theory of causes in the history of philosophy, on the chief philosophical problems, on the multiplicity of meanings of certain key philosophical terms, on act and potency, on being and essence, on the philosophy of mathematics, and on God. Those that the editors thought worth collecting were given the title *Metaphysics; i.e.,* the tract that is to be read after the *Physics.* It is not surprising, then, that the *Metaphysics* and the other works of Aristotle sometimes seem to lack unity or any clear progression of thought, that they are sometimes repetitious and at times even contradictory. The texts furthermore suggest that students or subsequent members of the Lyceum even revised Aristotle's expressions. It is probable that Aristotle would never have released the work. Andronicus, assisted by previous editors, imposed a logical and didactic order upon all the writings, undoubtedly influenced by Aristotle's own emphasis on logic as the propaedeutic (preparatory study) of all understanding. By ignoring the chronological order of the treatises and by grouping dissertations from different periods under the same title, the editors fashioned the Aristotelian corpus into a systematic whole. It is quite likely that Aristotle himself had never thought of his writings in this way.

Formation of existing works

Aristotle's treatises reveal the philosopher at work. He defines the problem he is to deal with, assesses the views of his predecessors, formulates his own preliminary opinion, considers whether there is a need to modify it in the light of difficulties and objections, rehearses the arguments for different points of view—always searching, in short, for the most adequate solution or resolution of his problem. The reader, therefore, sees Aristotle at work, not dogmatically propounding a doctrine but often laboriously developing a perspective or an insight that emerges from difficulties, contradictions, and paradoxes. Not surprisingly, few syllogisms appear in Aristotle's treatises; the reader, however, should perceive in them a structure that Aristotle himself terms "dialectical"; *i.e.,* in the manner of a dialogue by an exchange of arguments for and against.

"Dialectical" nature of Aristotle's thought

THEORIES OF THE EVOLUTION OF ARISTOTLE'S THOUGHT

From the conclusions of Alexander of Aphrodisias in the latter part of the 2nd century, a distinction was established between the doctrine expressed in Aristotle's treatises (the technical writings that have come down to the present) and the popular Platonizing dialogues (the "exoteric" works surviving only in fragments). The orthodox view for 17 centuries was that the treatises were the sources for Aristotle's genuine thought; Valentin Rose, a 19th-century German scholar, proposed that all of the lost dialogues are spurious because their doctrine was inconsistent with that of the treatises. The underlying assumption was that a man of such strict and systematic mind as Aristotle would maintain strict constancy and never abandon opinions once formed.

Jaeger's developmental theory and rejoinders. In the first half of the 20th century a developmental theory of Aristotle's thought was submitted by Thomas Case, an English scholar, and elaborated in detail by Werner Jaeger, a German historian of Greek philosophy, in his *Studien zur Entstehungsgeschichte der Metaphysik des Aristoteles* (1912; "Studies in the History of the Origin of Aristotle's Metaphysics") and later in his *Aristoteles: Grundlegung einer Geschichte seiner Entwicklung* (1923; *Aristotle: Fundamentals of the History of His Development*). Employing a historicogenetic methodology, Jaeger announced that the greater part of Aristotle's lost works represented his

thought while he was still at the Academy and under the immediate influence of Plato: the preponderance of such themes as the immortality of the soul, disdain for the material world, the doctrine of the "recollection" of Ideas, the supremacy of wisdom, asceticism, and the existence of God were recognizably Platonic. These dialogues addressed a wide audience and were presented in an elegant literary style that fascinated classical authors. According to Jaeger, Aristotle gradually distanced himself from Plato's position, even during his time at the Academy—rejecting certain Platonic arguments, adopting contrary positions, continually evolving from Platonic Idealism to a marked empiricism.

Theory of Aristotle's early Platonism

In response to this evolutionary theory, critics have noted that the analysis of Aristotle's works generates complex problems, as observed above in the account of the formation of the text of his existing writings. The works edited by Andronicus of Rhodes are compilations of texts from different periods. Thus, the *Metaphysics* covers almost the entire career of Aristotle, as does the *Politics.* Often within the same chapter, even in a single paragraph, one discerns elements from different stages of Aristotle's thought: his early phase at the Academy, his maturity, his period of travel away from Athens, and his Lyceum experience, which purportedly was divided between morning sessions with his best students and afternoon meetings with a wider audience from Athens. Given all this, there are serious obstacles in the way of discussing the chronology of the treatises: it becomes extremely difficult to put a date on a work that was revised and modified in various ways during a considerable portion of Aristotle's intellectual career.

One of Jaeger's main assumptions, moreover, is questionable. He supposes that Aristotle only agreed with Plato during his early years at the Academy or, at the latest, until the close of the first Athenian period (347 BC). This assumption, however, is arbitrary and cannot be corroborated by the evidence.

One could conceivably defend the converse: that Aristotle, in a self-assured youth, could have strongly challenged Plato but, having subsequently become conscious of the more profound significance of his master's philosophical postulates, did not hesitate to integrate with his own thought one or more Platonic theses. Indeed, in one of the logical treatises, *Topics,* considered one of Aristotle's early works because it reflects discussions at the Academy, and in the *Eudemian Ethics,* the first version of Aristotle's course on ethics, there are strongly anti-Platonic views expressed.

Evolution of thought on the soul

Jaeger's theory is most plausible with regard to Aristotle's psychology, as demonstrated by François Nuyens, a Dutch historian of philosophy, in 1939. He held that in his early period, represented by the *Eudemus* and the *Protrepticus,* Aristotle began as a Platonist, describing the soul as a separate substance in an unnatural relationship with the body; next, in an intermediate stage, he described the body as the instrument of the soul, whose function is analogous to that of a pilot steering a ship; finally, in the tract *On the Soul,* he advanced more clearly the concept of a substantial unity of body and soul by making the soul the form, or actuality, of a natural body.

Other authorities on Aristotle have observed that such a linear transition in his thought occurs rarely among his writings. Even in the works on psychology, moreover, Aristotle's concern for metaphysical thought does not end with his intermediate biological phase but continues and extends even into his empirical stage at the Lyceum. Such a simultaneous preoccupation with both scientific and philosophical thought is further manifested in the sequence of books in the *Metaphysics.*

At the other extreme is the hypothesis of the German scholar Josef Zürcher, in *Aristoteles Werke und Geist* (1952; "The Works and Spirit of Aristotle"), who asserted that Aristotle's own thought always remained Platonic and that all of the characteristically Peripatetic philosophy came from his disciple and successor, Theophrastus, who is the true author of about three-quarters of the existing Aristotelian treatises. According to this theory, there never existed a young Aristotle and an elder Aristotle in terms of any development of his thought: there was simply a Platonic Aristotle and an anti-Platonic Theophrastus, an empiricist. This eccentric theory has found no followers. In reaction to it, the Swedish scholar Ingemar Düring suggested in 1966 that Aristotle never really subscribed to the Platonic theory of transcendent Forms, or Ideas, but maintained lifelong and coexistent interests in empirical investigation and metaphysical speculation; for Düring, as for Zürcher, there is no need to postulate any fundamental change or development in Aristotle's thought.

Recent analyses of Aristotle's development and systematizing. Aristotelian scholars have generally concluded that a basis exists for a theory of evolution in his thought but that the determination of the chronology and the degree of change presents a difficult set of problems. It is quite possible to agree with Jaeger that during Aristotle's first years at the Academy he acknowledged Plato's teaching on Ideas, and that he later rejected the theory. It is another matter, however, to suggest that in his later years he renounced such Platonically influenced doctrines as the immortality of the soul or the conception of a religious philosophy concluding in an ultimate being termed God. Increased attention to data of the senses in subsequent phases of his life, moreover, is not a sufficient argument for the emergence of an empiricist Aristotle, who could not but oppose a spiritualist and idealist Plato. It is true that Aristotle later criticized the doctrine of Ideas as inadequate and contradictory. But he continued, nevertheless, to recognize the effectiveness of metaphysical thought in arriving at the concept of a transcendent, nonmaterial, and subsistent intellect as the necessary explanation for the fact that anything exists. The consensus of modern commentators thus suggests that not every aspect of Platonic idealism was rejected by Aristotle as his appreciation of empirical knowledge and of the dynamic aspects of matter grew. Rather, alongside his experimental work in biology and physics was his continued insistence on the crucial differences between perception and thought, between accidental characteristics and the essential natures of things.

The inconsistencies, contrasts, and varying degrees of emphasis on different modes of thought throughout the Aristotelian corpus are not adequately explained either by positing intervening editors and copyists or simply by different stages in Aristotle's thought. He clearly attempted in all of the treatises to relate his own views to the whole history of thought before his time. On many occasions he was concerned, at the same point in the development of his thinking, to state different views seen as alternative possibilities. Often his method was deliberately aporetic; that is to say, he raised difficulties that he knew had to be faced but for which he supplied no immediate or definitive solutions. Left by Plato with a vast body of problems, Aristotle conscientiously pursued the ideal of correcting and complementing the intellectual tradition bequeathed to him. To this end he often followed parallel but distinct paths of investigation. His method was exploratory, and he used it on whatever fertile soil he was free to work. Only relatively late in life was he able to unify his results with any degree of success. The philosophy of Aristotle does not unfold simply by deducing consequences from assumed principles. Rather, it starts from *aporiai,* from puzzles or problems, and it proceeds by piecemeal, tentative, and multiform attempts at solutions. The end result that Aristotle in his optimistic moments hoped to achieve was indeed a fixed body of knowledge, systematically ordered and deductively demonstrated. But his method of inquiry was not deductive, and the finished system remained an aspiration rather than an accomplishment.

Aporetic nature of Aristotle's thought

SYNOPSES OF THE ARISTOTELIAN CORPUS

Works on logic. The term logic was not invented by Aristotle but goes back to Xenocrates of the Academy. Aristotle, however, attributed extensive significance to language (*logos*) and to the rules of discourse; thus he emphasized that language is distinctive of the human species, and he defined man as a rational animal, which in the Greek also means an animal possessing a language or speech or word.

In Aristotle's view, the purpose of language is to express the feelings and experiences of the soul, and consequently

words are signs, or symbols, of thoughts and other mental phenomena.

The logical treatises of Aristotle make up the collection known as the *Organon* ("tool"). This title was adopted by later commentators, who, in accordance with the well-established Peripatetic tradition, regarded logic as an instrument for doing philosophy. In Aristotle's preferred view logic was not included in the classification of the sciences at all, but it was treated as a preliminary to the study of each and every branch of knowledge. Aristotle's own name for logic was "analytics." The term logic, however, is employed in a somewhat restricted sense in Aristotle's own writings; *e.g.,* in the *Topics* I, 14. And there is some evidence that it was beginning to be used as the equivalent of dialectic or analytics almost immediately after Aristotle's death.

Contents of the *Organon*

The *Organon* contains the following treatises: the *Categories, On Interpretation, Prior Analytics, Posterior Analytics, Topics,* and the *Sophistical Refutations.* The arrangement within the collection is meant to be systematic rather than chronological. Indeed, the original chronological order can hardly be determined now with any certainty because Aristotle, or other editors, apparently used later insertions to supplement the original treatises. In a possible sequence of their composition, the *Categories, Topics,* and the *Sophistical Refutations* are listed earlier than *On Interpretation,* and this work, in turn, is earlier than the *Prior Analytics* and the *Posterior Analytics.* The chapters on modal logic in the *Prior Analytics* are probably the last that Aristotle added to the body of the *Organon.* Apart from the *Organon,* the fourth book of the *Metaphysics* could be described as a logical work inasmuch as it is centrally concerned with certain general principles of thought (the principle of noncontradiction, the law of the excluded middle).

In the *Categories,* Aristotle distinguished expressions that exhibit propositional unity from expressions that do not; that is, he distinguished between a simple term and a composite statement that relates a subject to a predicate. This notion of propositional unity can be traced back to Plato, but the treatment of simple expressions was Aristotle's innovation. He considered simple expressions neither true nor false and held that they may signify things in one or another of the following categories: substance, quantity, quality, relation, place, time, position, state, action, and affection. It is by no means clear whether this classification is to be regarded as primarily ontological (concerning the nature of reality) or as primarily verbal—*i.e.,* whether it is about actual things or about words and expressions; the same ambiguity has been characteristic of practically every other scheme of categories suggested since Aristotle's time.

Anti-Platonic evolution of logic

As a part of a theory of reality Aristotle later used the categories to criticize Plato's theory of Forms. For Aristotle, Plato was involved in a confusion between the category of substance and the other categories when, for example, he attributed substantiality, or concrete existence, to qualities such as beauty or wisdom. In chapter 5 of the *Categories,* Aristotle distinguishes within the category of substance between "primary substance" and "secondary substance." Primary substances are particular men, particular horses, particular stones, etc., and secondary substances are the species and genera to which the individuals belong. There Aristotle treated genus and species as substances of a derived kind. In the *Metaphysics,* however, species and forms appear to be substances of a primary kind. Aristotle's view, it must be said, is far from clear—some scholars see the *Metaphysics* as a return to a more Platonic conception of ontology.

On Interpretation begins with a brief but influential discussion of the simple parts of sentences, such as "names" and "verbs"; it then considers complete sentences of various kinds and examines the logical relationships (contrariety, contradiction, implication) holding among them. The work also contains a pioneering account of "modal" sentences ("It is possible that . . ."; "It is necessary that . . .") and a celebrated discussion of "future contingents." (If it is already true that there will be a sea battle tomorrow, then how can the battle be considered a contingent event? For if the truth is already determined, surely the battle is fixed and necessary? Aristotle's answer to this is that certain types of sentences about the future are neither true nor false.)

The *Topics* appears to have been intended as a manual for participants in contests that involved argumentation. For the most part this treatise consists of suggestions about how to look for an argument that will either establish or refute a given thesis; thus it elucidates general logical laws or rules.

The *Sophistical Refutations* exposes forms of reasoning that appear valid on the surface but are in fact fallacious. Examples of fallacious arguments are "begging the question," or circular argument (*e.g.,* a "proof" that the soul continues to exist after death because it is immortal); the "fallacy of the consequent," or arguing from a consequent to its condition (*e.g.,* if a man is a drunkard he becomes destitute; Peter is destitute: therefore Peter is a drunkard); and the "fallacy of the irrelevant conclusion," wherein, instead of proving the fact in dispute, the arguer seeks to gain his point by diverting attention to some extraneous fact.

The main achievement of the *Prior Analytics* is the development of the logical system now known as Aristotelian syllogistic. A syllogism is a form of argument consisting of three propositions (two premises and a conclusion). The stock example of a valid syllogism is the following:

Every Greek is a man.
Every man is mortal.

Every Greek is mortal.

Both premises are either affirmative or negative and contain two terms (the subject and the predicate) together with a sign of "quantity" ("every," "some," "no"). In addition, the propositions are either "assertoric" or "apodeictic" or "problematic"—they express the idea that something is or must or can be the case. "Every man must be rational" is apodeictic and affirmative; it is universal in quantity ("every"); its subject term is "man" and its predicate is "being rational." The *Prior Analytics* examines, with astonishing rigour and sophistication, the various possible forms of syllogistic argument.

In the *Posterior Analytics* Aristotle seeks to apply his logical theory to scientific and epistemological ends. He discusses the proper structure of scientific knowledge, urging that each science must depend on a set of first principles, or axioms, that are necessarily true and directly knowable. The truths, or theorems, that together constitute a science are to be deduced from its axioms, which both necessitate and explain them. Aristotle came to hope that all these scientific deductions could be formulated by way of apodeictic syllogisms. For this reason much of the second book of the *Posterior Analytics* is devoted to the theory of "definition," for Aristotle thought that the most important axioms of any science would be definitions of its proper subject matter. Among the various axioms of geometry, for example, there would be a definition of the triangle—an account of what a triangle really is or of the essence of a triangle.

Works on the philosophy of nature. In his treatise *Physics* Aristotle deals with natural bodies in general, or with all that is corporeal; special kinds of material bodies are discussed in his other physical works, such as *On the Heavens* or the *Meteorology.* The first book of the *Physics* is concerned with the intrinsic, constitutive elements of a natural body, those that he called "matter" and "form"; *i.e.,* the substratum that persists through change and the feature whose acquisition determines the nature of change. The second book treats mainly the different types of cause studied by the physicist, the material and the formal causes just mentioned, and the final and efficient causes, or the goal for the sake of which and the agent by means of which anything comes into being. Books 3 through 7 deal with movement, or motion, and the notions implied in it—such as space, position, and time, their magnitudes and continuity. The subject of Book 8 is the first mover, which, though not itself a natural body, is the cause of all movement in natural bodies; its necessary attributes—such as immovability and eternalness—are also examined.

Whatever the virtues or defects, clarity or obscurity,

of Aristotle's physical treatises, they assume that the distinction between physics and metaphysics (the "first philosophy," or the science of being as being) is valid. Although the conception of a continuous scale of nature from inorganic substances to biological and psychological phenomena is basic in all of his science, explanation does not consist in running uniformly up the hierarchy of beings to God nor in reducing all functioning to some material organ. The sciences of Aristotle are based on a multiple system of classification, not on a simple scheme of mutually exclusive and independently existent genera and species. The very distinction of causes in existing and mutable things permits the differentiation of the subject matters of the natural sciences.

The general principles discussed in the *Physics* are applied to the universe as a whole in *On the Heavens* (where Aristotle argues that the world is spatially finite but temporally eternal) and to the inanimate parts of the universe in *On Generation and Corruption* and the *Meteorology.* The former treatise discusses, in general terms, the four "elements" of the Aristotelian system (earth, air, fire, water) and their interrelations; Aristotle pays particular attention to the question of elemental change, whereby one element can alter and become another. The *Meteorology* deals with what, from a modern point of view, is a miscellany of topics—astronomy (*e.g.,* comets), geography (*e.g.,* rivers), chemistry (*e.g.,* burning), as well as meteorology (*e.g.,* rainbows). In addition to the general principles of physics and the theory of the elements, Aristotle relies on a further postulate: he supposes that "exhalations," some moist and some dry (steam and smoke), are constantly given off by the earth, and he attempts to explain the various phenomena he investigates in terms of the operation of these exhalations.

The principles of the *Physics* are also evident in Aristotle's biological and zoological writings. The largest of these, the *History of Animals* (a better translation of the Greek title would be *Inquiry into Animals*) consists in the main of descriptions of different animal species. Some of these descriptions—notably those of the crustaceans—are remarkable for their detail and accuracy. Some scholars regard the *History* as no more than a repository of raw data, collected for scientific scrutiny but not yet ordered or systematized. Others, however, think that Aristotle is concerned with constructing a biological taxonomy that divides the animal world into genera and species. The truth probably lies somewhere between these two views. There is neither a fully fledged Aristotelian taxonomy nor a fixed system of genera and species, but the material in the *History* is not simply an unorganized heap—the subject matter of the work is intelligently and significantly arranged.

However that may be, the *Parts of Animals* and the *Generation of Animals,* although they too present a quantity of empirical data, are primarily scientific and explanatory in intention. Aristotle is concerned with the nature and the function of the various animal organs and other "parts"; he wants not merely to describe and list them, but to "explain" them, both by reference to similarities across different animal species and also—and more strikingly—by reference to their functions within the animal's bodily system and behaviour. It is here that Aristotle's insistence on teleological explanation is most apparent: "nature," he says, "does nothing in vain," and although he does not, strictly speaking, hold that all features of animate beings have a functional explanation (the colour of eyes, for example, is accidental), such explanations are pervasive and are the mark of good science. The *Generation of Animals* considers specifically the problems of reproduction and growth. What contributions do male and female parents make to the embryo? What characteristics are inherited, and from whom, and how? How do embryos grow and develop, and how in particular do they acquire the different faculties that together constitute their souls? In this, Aristotle's most mature scientific work, the virtues and the vices of his method are most plainly to be seen: he is usually modest, careful, exact; he advances theoretical explanations, but he does not let theory prejudice observation; he attempts to produce a genuinely scientific work. On the other hand, the limitations of his knowledge—and of his means of acquiring new knowledge—are evident; and at least some of his theoretical concepts are crude and inadequate.

The biological works also include two short essays discussing animal locomotion entitled the *Movement of Animals* and the *Progression of Animals.* Here Aristotle attempts—not wholly successfully—to combine a rigorously mechanical account of animal motion, considered in terms of the physiology of the body and the nature of the medium, with a psychological discussion of the mental antecedents (perception, thought, desire) that explain animal behaviour.

Works on psychology. The relation between the active principle and the passive continuum (or between form and matter) that is operational in sentient and intellectual life is examined in *On the Soul.* After exploring the concept and the conditions of life, Aristotle relates the function of matter and form (body and soul) in human life to all of life's biological and psychological phenomena while rejecting Platonic transcendentalist and pre-Socratic materialist theories on the nature of the soul. The soul, as the form of the organic body, consists of an ordered set of faculties; these are, in hierarchical order, the nutritive, the perceptual, and the intellectual faculties. The nutritive faculty is common to all living things and is responsible for growth and nutrition; the perceptual faculty is common to all animals and is responsible for, among other things, sight, hearing, smell, and locomotion; and the intellectual faculty is peculiar to humans. Aristotle gives detailed accounts of the modes of perception (in addition to the five senses and their objects he postulates the existence of a "common sense" that unites their deliverances) and a notoriously difficult account of thought (which distinguishes an "active" from a "passive" intellect). The work also contains a discussion of animal movement and of its preconditions—of imagination and of desire.

In the *Parva Naturalia,* the medieval designation for a collection of short treatises on natural functions, the argument of *On the Soul* is supplemented by a sequence of treatises on sense and the sensible, memory and reminiscence, sleeping and waking, prophecy in sleep, the length and brevity of life, youth and old age, life and death, and respiration.

Works on metaphysics. The study of metaphysics, the function and content of which have generated neither conviction nor consensus of opinion on the scope of its subject matter, is—together with the syllogism and the differentiation of kinds of premises—an innovation of Aristotle. In his *Metaphysics* the doctrines that Aristotle sometimes refers to as "wisdom" and sometimes as "first philosophy" or even as "theology" are developed. Its task is that of describing the most general or abstract features of reality and the principles that have universal validity. In a famous (and misleading) phrase, he describes metaphysics as the study of "being *qua* being." By that he means that metaphysics studies whatever must be true of all existent things just insofar as they exist, that it studies the general conditions which any existing thing must satisfy.

Book 1 of the *Metaphysics* discusses in a preparatory way the problem of causal explanation. Aristotle gives a survey of the forms of explanation used or discussed by his predecessors, and discovers that his own theory of "four causes" represents the truth toward which they were struggling. This survey is one of the most important sources for information about the pre-Socratics and also about certain aspects of Plato's philosophy. Book 2 is a short essay on the principles of science, and Book 3 sets out a long series of metaphysical puzzles, or *aporiai.* The puzzles receive a preliminary discussion: most, but not all, of them are dealt with at greater length in later parts of the *Metaphysics.*

Book 4 explains Aristotle's conception of "first philosophy" as the general study of the conditions of existence, and it contains a defense of the principle of noncontradiction ("not both *P* and not-*P*") and the law of the excluded middle ("either *P* or not-*P*"). Book 5, sometimes called Aristotle's philosophical lexicon, is devoted to ambiguous philosophical terms: Aristotle analyzes and distinguishes

the different usages of some 40 key words. Book 6 returns to the issues of Book 4.

Books 7–9 form a unit. These central books are among the most difficult that Aristotle ever wrote, and they defy summary. The question to which they address themselves is this: What is substance? What are the fundamental constituents of the world, the things that enjoy an independent existence and can be known and defined? Aristotle's discussion is tortuous. It turns on the ideas of matter and form, of substance and essence, of change and generation, of actuality and potentiality. Aristotle's conclusion, it seems, is that substances are, in some sense, forms. They are not abstract Platonic Forms, but concrete, particular forms. They are the things designated by such phrases as "this man," "that horse," or "this oak tree."

Book 10 is a self-contained essay on "oneness"—on unity, continuity, identity, and related concepts. Book 11, which simply summarizes parts of the *Physics* and earlier parts of the *Metaphysics,* is generally regarded as spurious. Book 12 gives Aristotle's "theology": he asks how many causes must be posited to explain the world and arrives eventually at the conception of God, or of the first, or unmoved, mover. Aristotle's God, however, is not a personal God interested in the affairs of this world. Instead he is pure intelligence and as such completely indifferent to the vicissitudes of the world (as is implied in the concept of unmoved mover). In addition, the concept of first mover is not to be understood in a temporal sense. The first mover is not the creator of the world—indeed, Aristotle thought that the world was not created at all but had been in existence for all eternity—but the fountainhead of all motion. In that sense he is the ultimate cause of everything that happens in the world. Finally, Books 13 and 14 contain a long discussion—mostly critical and directed against Plato—of the nature of mathematical objects.

Works on ethics and politics. In emphasizing the crucial differences in the purposes of the theoretical and practical sciences, Aristotle indicates that the practical disciplines, unlike the theoretical, are for the sake of doing or making something and not for the sake of contemplating, defining, or knowing it. Thus, at the start of his *Nicomachean Ethics* he explains how the practical sciences are incapable of the exactness of the theoretical sciences, for their subject matters are not limited to things that are amenable to precise definition, but involve habits and skills, which can be acquired and lost, and associations and institutions, which in their changes affect the accomplishment of political actions and the practicability of moral ends. And however precise biological or psychological definitions may be, man varies as moral agent and as citizen according to environmental determination, educational background, and the influences of family, economic position, social class, means of livelihood, and even the associations of his leisure.

Nature of ethics as a practical science

Relating ethics to politics, Aristotle set out to demonstrate that problems of morality as they affect the individual cannot be separated from each other or from problems of political association. The *Ethics* and *Politics* therefore do not develop separate sciences or independent subject matters but rather supplement each other by treating a common field according to different aspects.

Although he treated moral problems in terms of the potentialities of individual men, the ability to practice and actualize these potentialities is dependent upon political circumstances. Therefore, in the first chapters of Book 1 of the *Ethics,* Aristotle begins by introducing moral considerations into the broad context of political philosophy, and he ends by returning, in the concluding Book 10, from the examination of happiness and the contemplative life to a shrewd statement of the contribution of law to moral questions, which forms the transition from ethics to politics.

Aristotle's approach to ethics is teleological; that is, he discusses ethics not in terms of moral absolutes but in terms of what is conducive to man's good. This approach leads him to examine various kinds of good and to arrive at the identification of the highest good with the attainment of happiness. After careful discussion of the problematic concept of happiness, Aristotle arrives at a definition of happiness as activity of the soul in accordance with virtue.

Aristotle distinguishes moral virtues and intellectual virtues, which are determined, respectively, by the irrational and the rational powers of the soul. Man, however, does not possess these virtues at birth but comes endowed with the capacity, or disposition, for developing them in the course of time. For example, a child begins by following his parents' injunction to tell the truth without initially realizing the moral excellence of his action; yet eventually the habit of veracity becomes an ingrained part of his moral character. Aristotle then differentiates virtue from vice, arriving at the definition of virtue as a "mean," or middle disposition, between the vicious extremes of excess and deficiency; courage, a virtue, for example, is the mean between cowardice and rashness.

Aristotle concludes his discussion by defining the highest happiness open to man. Because happiness is an activity in accordance with virtue, it follows that the highest happiness should be in accordance with man's highest virtue. And that, according to Aristotle, is the activity which distinguishes man from the other animals, namely the activity of reason or activity in accordance with reason. Thus in its ideal form happiness turns out to consist in a life of intellectual contemplation. Aristotle, on the other hand, also concedes that the political life (activity in accordance with moral virtue) can bring happiness, albeit "in a secondary degree."

Happiness as virtuous activity

The *Politics* takes up the problems of human action and association as they bear on the ends of communal life encompassed in living well. But the choice of political ends requires a complex examination of possible criteria capable of application to the vast diversity of men and human conditions. Grounding his argument on the premise that man is "naturally" a political animal, Aristotle develops the theory of the state, distinguishes various kinds of constitution, and considers the best state for the particular circumstances, character, and conditions of the citizens. Aristotle also discusses the nature and causes of political instability and revolution. The last two books of the *Politics*—part of an unfinished description of the ideal state—are largely concerned with education.

Works on art and rhetoric. Aristotle analyzes rhetoric in terms of its end, or final, cause, which is persuasion. Like dialectic it is not a science, and therefore it has no specific subject matter, no single method, and no proper set of principles. It is simply the faculty, or power, of observing in any given case the available means of persuasion. According to Aristotle, there are three modes of persuasion that a speaker may exercise: the persuasive power of his own character, the excitation of desired emotions in the audience, and proof or apparent proof.

In the *Poetics* Aristotle's analysis of poetry provides for careful isolation of the specific character of poetry. In comparing poetry to history, he states that poetry is more philosophic than history and thus of greater intrinsic worth. The difference is attributable not to form—history written in metre is still history—but to the fact that the historian deals with singulars (*i.e.,* with specific events and specific personages). The poet, on the other hand, creates types and situations that, while imitating nature, are, nonetheless, akin to universals; that is, the poet describes what is possible as though it were both likely and necessary. Yet Aristotle also permits the analogy of poetry to oratory as well as the consideration of the moral, political, and educational effects of both. Tragedy, however, which is the only kind of poetry analyzed in the extant portions of Aristotle's work, is defined in terms of its form, not in terms of its purpose, as a kind of imitation rather than as a mode of persuasion or excitation. Thus, in the famous definition of the sixth chapter, it imitates a serious action of great magnitude in a dramatic form and accomplishes the purification (*katharsis*) of the emotions of pity and fear.

Nature of tragic poetry

Using this definition as the basis for the discussion of poetry, Aristotle considered poetic art in terms of the characteristics and interrelations of the six parts, or components, of tragedy: plot, character, and thought (the objects of imitation); diction and melody (the means of imitation); and spectacle (the manner of imitation).

The last four chapters of the *Poetics* return to more general questions of value and to final causes by means of detailed comparisons of tragedy with comparable poetic works and specifically with epic.

ON READING ARISTOTLE

Goethe compared Aristotle's philosophy to a pyramid rising on high in regular form from a broad base on the Earth. Because each part of Aristotle's philosophy contributes to the understanding of other parts, it is generally true of his works—even more than those of most philosophers—that they cannot be read initially with a sure and well-grounded understanding but they must be reread for the sake of perceiving the primary conceptual and methodological relationships. Faced with the mass of materials that constitute the imposing body of his works, the reader might best start with the treatment of those problems that are relevant, interesting, or important to him. Aristotle himself often reiterated the suggestion that the inquirer concentrate first on sense experience as something that is better known to him and attend only afterward to the essential concepts of things in the effort to organize knowledge and constitute the principles of the particular sciences. Indeed, he frequently distinguished the process of inquiry and discovery from that of demonstration and proof.

Interrelationship of all Aristotle's writings

Aristotle's thought can be said to be known in two ways: (1) as remnants of his doctrines constituting the speech of Western culture, in the tradition of Western thought, and in the history of its sciences; and (2) as it is known from the attentive study of his writings. What Aristotle discovered in his intellectual inquiries and what he said can be most readily intelligible to someone of Western culture in the modified form in which it still constitutes part of that thought and conviction. It may be suggested, therefore, that Aristotle be read for the first time in an order the reverse of that in which his works have traditionally been arranged and that his conclusions and analyses be examined before inquiring about his principles. It is true, for example, that what Aristotle said concerning the poetic and rhetorical arts is more complex in manner of analysis and more difficult in systematic construction than what modern writers might say on such subjects; but it, nonetheless, still approximates more nearly to contemporary thought in this area than do his works on physics and metaphysics. And, for the same reason, his moral and political theory led him, in the course of its development, to many distinctions and statements that a modern reader would be disposed to accept or to reject without too lengthy critical discussion.

Once the manner of Aristotle's analysis is more firmly grasped and better appreciated, the reader can proceed to take up his logical works in the *Organon* and his investigations of space, time, and motion in the *Physics.* Through later consultation with the more complex thought in the *Metaphysics* and in *On the Soul,* he can review the earlier tentative conclusions made in the study of ethics and literary theory in the light of the deeper insights, acute distinctions, and strength of argument thus acquired.

When a student approaches Aristotle's conclusions in the light of his principles, allowing the text to illuminate his own experiences, Aristotle can then be appreciated not only for his expression of a philosophy but also as a help in the cultivation of the mind. And this is a task to which Aristotle himself thought all men should devote themselves and to which his philosophy remains a unique contribution. (A.H.Ao./Ed.)

ASSESSMENT AND NATURE OF ARISTOTELIANISM

The extent to which Aristotelian thought has become a component of civilization can hardly be overestimated. To begin, there are certain words that have become indispensable for the articulate communication of thoughts, experiences, and problems. Some words still carry their Greek form, whereas others have become established in their more important meanings as Latin equivalents of Aristotle's own words. The centuries-long impact of Aristotelian schooling lies at the root of the establishment of the following vocabulary: "subject" and "predicate" in grammar and logic; "form" (information, transform) and "matter" as expressing the two correlative aspects of something that has acquired or acquires something else that is possibly essential to it; "energy" as the active power inherent in a thing; "potential" for what is latent but can be released; "substance" and "essence," "quantity" and "quality," "accidental," "relation," "cause" (and the many meanings of "because" corresponding to the four causes), "genus" and "species" (general, special), "individual," "indivisible" (atomic)—these constitute only a small sample of terms that still carry the mark of Aristotle's philosophy.

Characteristic features

Beyond language, features that cumulatively or severally characterize Aristotelianism include, in philosophical methodology, a critical approach to previous, contemporary, or hypothetical doctrines; the raising and discussing of doctrinal difficulties; the use of deductive reasoning proceeding from self-evident principles or discovered general truths; and syllogistic forms of demonstrative or persuasive arguments.

In epistemology, or the theory of knowledge, Aristotelianism includes a concentration on knowledge either accessible by natural means or accountable for by reason; an inductive, analytical empiricism, or stress on experience, in the study of nature—including the study of men, their behaviour and organizations—leading from the perception of contingent individual occurrences to the discovery of permanent, universal patterns; and the primacy of the universal, that which is expressed by common or general terms.

In metaphysics, or the theory of Being, Aristotelianism involves belief in the primacy of the individual in the realm of existence; in the applicability to reality of a certain set of explanatory concepts (*e.g.,* 10 categories; genus–species–individual, matter–form, potentiality–actuality, essential–accidental; the four material elements and their basic qualities; and the four causes—formal, material, efficient, and final); in the soul as the inseparable form of each living body in the vegetable and animal kingdoms; in activity as the essence of things; and in the primacy of speculative over practical activity.

In the philosophy of nature, Aristotelianism denotes an optimistic position concerning nature's aims and its economy; believing in the perfection and in the eternity of the heavenly, geocentric spheres, perceiving them as driven by intelligent movers, as carrying in their circular movements the stars, the Sun, the planets, and the Moon, and as also influencing the sublunary world; and holding that light bodies rise naturally away from the centre of the Earth, while heavy bodies move naturally toward it with a speed related to their weight.

In aesthetics, ethics, and politics, Aristotelian thought holds that poetry is an imitation of what is possible in real life; that tragedy, by imitation of a serious action cast in dramatic form, achieves purification (*katharsis*) through fear and pity; that virtue is a middle between extremes; that man's happiness consists primarily in intellectual activity and secondarily in the exercise of the virtues; and that the state is a self-sufficient society, necessary for men to achieve happiness.

HISTORY OF ARISTOTELIANISM

Continuity of the Aristotelian tradition. Since Aristotle's death there have been, without interruption until the present, schools and individuals who have cultivated the study of his works and fully or partly adopted and expounded his doctrines and methods. They have interpreted or misinterpreted, approved or condemned, and reshaped or utterly transformed them. The languages in which this interest was most forcibly expressed have changed in turn and over time from Greek to Latin; to Syriac, Arabic, and Hebrew; to Italian, French, English, and German. The main centres in which it appeared have been as far apart as Greece, North Africa, and Rome in the ancient world; Persia and Spain, Sicily and the British Isles in the Middle Ages; and Germany and North America in more recent times.

Main centres and traditions

The main strand of the Aristotelian tradition has been the Greek line, which lasted 2,000 years, mainly in the area along the eastern Mediterranean Sea, and branched

off at various stages between the 4th and 15th centuries, giving rise to (or strengthening) other traditions. The Latin branch originated in Rome in the 4th century and acquired a new impulse, probably from Athens, in the early 6th century. From these beginnings it was revived in the 9th century and again in the 12th, at which time a second and even stronger Aristotelian wave emerged from Constantinople, to be followed by a third, via the western Arabic schools, from Spain; and both branches spread to Italy, France, and the British Isles. The final direct contribution from the Greek to the Latin tradition came to Italy, once more from or through Constantinople, in the 15th century.

Shortly after the beginning of Latin Aristotelianism certain Armenian and Syrian members of the Greek schools of Athens and Alexandria in Egypt introduced Aristotelian teachings into their schools. The Armenian tradition was still alive in the 19th century in such places as Madras and Venice; and the Syrian tradition, which never completely disappeared, was still powerful in the 14th century, after having given birth, in the 9th and 10th centuries, to an Arabic tradition. Arabic Aristotelianism was the product of Syrians, Persians, Turks, Jews, and Arabs who wrote and taught in their own countries as well as in Africa and Spain until the 12th century. Much of it and of what the Jews produced in Hebrew in the following two centuries passed into the Latin tradition between 1130 and 1550. Thus, all of the varied heritage that had derived ultimately from the Greek line and had been vastly enriched by other cultures came to be collected, through the Latin branch, by modern Western philosophical movements.

The Greek tradition. *Early development.* For some decades after his death Aristotle's own school, the Peripatos or Lyceum, remained, in a truly Aristotelian spirit, a centre for critical research—not for the dogmatic acceptance of a closed system. Aristotle's immediate successor, Theophrastus, independently elaborated his master's metaphysics and psychology and added to his study of nature (botany and mineralogy) and logic (theory of propositions and hypothetical syllogisms). Various members of the Lyceum coordinated Aristotelian thought with other current schools of philosophy. Thus Aristoxenus joined Aristotelian and Pythagorean doctrines; Critolaus united Aristotle's theory of the influence of the heavens on the world with the Stoic theory of providence; and Clearchus of Soli combined Plato's views on the human soul with Aristotle's.

The Stoics, Skeptics, and Epicureans

Outside the Lyceum, the Stoic school was partly following Aristotle in its interest in formal logic, the theory of meaning, and use of the categories (*e.g.*, substance, quality, relation). It was Aristotelian also in its empiricism, as well as in its concentration on nature, in several aspects of natural science, and in its belief that man is intrinsically a social being. The Skeptics sometimes relied on Aristotelian forms of argument to prove their systematic doubts. Even Epicurus, who may have fought against Aristotle's early theology and psychology and ignored his mature philosophy, was, nonetheless, near him in his doctrine of the will and in his conception of friendship and the pursuit of knowledge as the high aims that give satisfaction and pleasure to man.

Although relatively little was known of Aristotle's "esoteric" works until the 1st century BC, his more popular, literary, and Platonizing writings influenced eclectics such as Panaetius and his pupil Poseidonius; and this influence continued, helped by the Roman philosopher and lawyer Cicero, well into the 4th and 5th centuries AD. Upon it was based the tendency to establish a harmony between the thought of Plato and Aristotle—a feature that recurred through the whole history of Aristotelianism—and perhaps the ascription to Aristotle of the *De mundo* ("On the Universe"), a cosmological treatise of the 1st century BC, which found favour with all of the different traditions until the 16th century.

In the 1st century BC Aristotle's "esoteric" writings were organized into a corpus and critically edited by Andronicus of Rhodes and other scholars. The edition was used by Nicholas of Damascus, a historian and philosopher, in an attempt to expound Aristotle's system. This may be viewed as the beginning of a new era of a scholarly and scholastic Aristotelianism in which Aristotle had to be taken as the basis for the acquisition of true knowledge in a number of fields. Individual works began to be commented and lectured upon; organized philosophical studies began to have as their introduction Aristotle's works on logic, especially the *Categories.* Thus the pattern was set for the next 17 centuries. Almost pure Aristotelianism, based on the "esoteric" works, lived on until the 4th century. Many scholars—the most eminent of them being Alexander of Aphrodisias, who from AD 195 held the Athenian chair of Aristotelian studies created by the Roman emperor Marcus Aurelius—provided the works on logic, ethics, metaphysics, natural philosophy, and psychology with detailed and penetrating commentaries meant for the specialist. The interpretation of Aristotle was for many generations molded by these scholars. Others—the greatest being Themistius, a professor in Constantinople in about AD 350—practically rewrote many of Aristotle's treatises in a more modern language and more readable style.

This new, scholarly Aristotelianism had established itself sufficiently as the philosophical and methodological frame of learning for it to be adopted, at least in part, by most men of culture—including Ptolemy, the greatest astronomer of antiquity, and Galen, the most eminent medical scientist.

The Neoplatonists

Relationship to Neoplatonism. Aristotle's works were adopted by the systematic builders of Neoplatonism in the 3rd century AD. Plotinus, the school's chief representative, followed Aristotle wherever he found a possibility of agreement or development, as he did in Aristotle's theory of the intellect. And Plotinus' pupil Porphyry, the first great harmonizer of Plato and Aristotle, provided the field of logic with a short introduction (*Isagoge*). The *Isagoge,* in fact, is only concerned with a simple and rather mechanical treatment of five concepts that had been much used by Aristotle. These were the concepts of genus, or kind (as animal is the genus, or kind, under which Socrates falls); species, or sort (Socrates is a man); differentia, or distinguishing characteristic (rationality distinguishes men from other members of the genus animal); property (being capable of laughter was said to be a "property" of men inasmuch as all and only men are capable of laughter); and accident, or characteristic in general (as it might be an accident of Socrates to be pale). This introduction soon became an integral part of the *Organon* (the logical works of Aristotle) and thus acquired undeserved Aristotelian authority in all schools for more than 1,500 years. From that time on, Aristotelianism became indissolubly tied up with Neoplatonism.

Neoplatonism dominated the school of Athens, where, apart from logic, Aristotle's writings were destined to be studied mainly as a basis for philosophical disputations—disputations in which the Platonic view was usually victorious. Scholars like Ammonius—a pupil of Proclus, the most accomplished systematizer of Neoplatonism, head of the Athenian school in the mid-5th century, and himself extremely well-versed in Aristotle—found Alexandria a considerably more attractive place for Aristotelian studies, in that it was tolerant of many views. There pagans and Christians coexisted and cooperated, and from there they carried Aristotelian learning to a number of other schools: Simplicius, a pupil of Ammonius who was inclined to Platonism, took it back to Athens and—when Justinian closed that pagan school in 529—to Persia; Sergius, a physician and Nestorian priest, carried it to the Christian schools of Syria; and Stephanus of Alexandria took it to Constantinople. The schools of Alexandria and Athens produced from about AD 475 to 545 the most intensive collection of Aristotelian commentaries, by scholars like Ammonius, philosophers of science like Simplicius, and philosopher-theologians like Philoponus (see also PLATONISM).

Relationship with early Christian theology

Before the 5th century, Christian theology had been affected only marginally and indirectly by Aristotle. The elementary study of Aristotelian logic had proved indispensable for a disciplined training of theologians, and some of the concepts from Aristotle's *Physics* and *Metaphysics* that entered into the elaboration of this logic became equally essential for the rational formulation of points of

dogma. The aforementioned five terms of Porphyry and the 10 categories of Aristotle were used or implied in the mystical theology of Pseudo-Dionysius (an unidentified 5th-century Christian Neoplatonist), which was to become one of the principal components of Christian speculation in the Greek, Oriental, and Latin schools. Descriptions of God and distinctions between the three Persons of the Trinity came to include, in an increasingly technical sense, the Aristotelian terms substance, essence, accident, form and matter, species and nature, quality, quantity, and property; these terms were not always used in a purely Aristotelian sense, however. In this way, as well as through the purely philosophical schools, Aristotelianism entered the first Greek Scholasticism of St. John of Damascus, an 8th-century doctor of the church.

From the Byzantine renaissance to the 15th century. The Byzantine scholarly renaissance in the 9th century included a revival of interest in Aristotle: the old books were rediscovered and reedited (the oldest manuscripts still existing today belong to this time). Photius, patriarch of Constantinople and a leading figure in that renaissance, included in his encyclopaedic works summaries of the elements of Aristotelian logic. More extensive scholarly activity resulted from the reestablishment of the Academy in Constantinople in the 11th and 12th centuries under the successive leadership of such men as Michael Psellus, an encyclopaedic philosopher; his student John Italus; Michael, the archbishop of Ephesus; and Eustratius, the metropolitan of Nicaea. At the Academy teaching and exegetical work went hand in hand; debates on the superiority of Plato or Aristotle and attacks on philosophy by the religious schools did not seriously weaken these activities. There was perhaps not much that was new in the understanding or the development of Aristotle's doctrine; but logic was no longer the only focal point of Aristotelian studies. Indeed, they covered, more widely than had been done in Alexandria, practically the whole corpus, including some work on Aristotle's political theory, on his ethics, and on his biology. In addition, there were philosophical debates similar to those taking place in the Latin schools; they were based on texts of Aristotle and treated such issues as the theory of universals and the logical structure of language.

In the 13th and 14th centuries popularization and systematization—in an encyclopaedic or philosophical form—took the upper hand in the work of Nicephorus Blemmydes, George Pachymeres, and Theodore Metochites. At a time when Greek thought was being strongly influenced by the Latin tradition, especially by the work of Thomas Aquinas, the traditional debate on Plato and Aristotle took new forms. Aristotelianism appeared in the teaching of Barlaam the Calabrian, who sought to champion rationalism in faith; this was combated from a Platonic point of view by Nicephorus Gregoras. In the 15th century, when Greeks were becoming part of the Italian philosophical scene, Aristotelian rationalism was strongly defended by the upholders of Christian theology against such men as George Gemistus Plethon, who proposed a new universal religiosity tinged with an admiration for Plato and paganism. The victory in this intellectual battle went to the moderates like John Bessarion, Plethon's influential pupil, who, though he preferred Plato, admired Aristotle, translated his *Metaphysics,* and collected manuscripts of his works; he converted from the Greek (*i.e.,* the Greek Orthodox) church to the Latin (*i.e.,* what is now called the Roman Catholic) church, in which latter communion he became a cardinal.

The early Latin tradition. The echoes of Aristotle's early writings in Cicero, a few signs of his indirect influence on other writers, and a more considerable contribution to post-Aristotelian logic in Apuleius, a Platonic philosopher who flourished in the 2nd century AD, are indications of the general cultural intercourse in this area between Latins and Greeks. The presence of Plotinus and Porphyry in Rome in the 3rd and early 4th centuries probably started the more serious interest in Aristotle there, of which the first results were, perhaps, Victorinus' adaptations in Latin of Porphyry's *Isagoge* and Aristotle's *Categories.* Logic was still the only part of Aristotle that had entered Latin culture when Themistius' teaching attracted the attention of Roman pagan circles in the 4th century.

Influence on Boethius

Again, only the logical works of Aristotle, together with some extracts from Greek commentaries on them, seem to have reached the hands of Boethius, a Roman scholar and statesman of the early 6th century, when he was attempting to transmit to the Latins as much as he could of Greek learning. He translated these works and elaborated on the commentaries and on some other later texts of logic that are partly based on Aristotle. He acted primarily as a conduit, and some scholars are not prepared to ascribe to him interpretations and plans contained in the Latin works that bear his name. Even the plan of commenting on "as much of Aristotle as would come into his hands" and showing that Aristotle and Plato agreed was the traditional approach going back at least to Porphyry. Nothing remains to show where Boethius himself stood in judging Aristotle and the several parts of his philosophy. The same observations probably hold true with regard to Boethius' various theological treatises, in which the Aristotelian concepts that helped to organize the theology of the Trinity were unmistakably taken over from similar Greek treatises. A disproportionate value, however, was later attached to Boethius' own original contribution in both logic and theology; simply the fact that his name was connected with these texts made people in the Middle Ages ascribe to him the primary responsibility for their contents.

The Syriac, Arabic, and Jewish traditions. The increased sense of linguistic and national identity and the religious movements of the 5th and 6th centuries such as Nestorianism (a heterodox doctrine that so stressed the distinction between the divine and human natures of Jesus Christ as to suggest that they belonged to two persons) and Monophysitism (a heterodox doctrine asserting that there is only one nature in Jesus Christ) led to the foundation of Syriac centres of studies in the Persian and Byzantine empires, especially at Edessa (now Urfa, Tur.) and Antioch. Proba and Sergius of Resaina were among those who contributed, through translations of the basic logical texts and commentaries on them, to the establishment of Aristotelian studies in these centres. At the time of the Arabic invasion of the Byzantine and Sāsānian empires around 640, and for several generations afterward, these centres continued to grow in importance. Most notable was the great school of Kinnesrin, which was represented by such men as Severus Sebokht, who wrote on Aristotle's syllogisms; Jacob, bishop of Edessa, a theologian, grammarian, and translator; and Georgius, bishop of the Arabs, author of a commentary on the *Organon.* Interest remained, however, mainly confined to logic and its application to theology.

The Syrian Christians formed the philosophical and scientific intelligentsia when in the 9th century al-Ma'mūn, the seventh ʿAbbāsid caliph, organized the Arabic centre of learning of the new Islāmic empire in Baghdad. By then the Syrian scholars had acquired and translated most of Aristotle's works. They also then translated them into Arabic, both from the Syriac and directly from the Greek, and added many texts of commentators on Aristotle. In this way Ḥunayn ibn Isḥāq, his son Isḥāq, Abū Bishr Mattā ibn Yūnus, Yaḥyā ibn ʿAdī, and many other Syrians provided the basis for a brilliant philosophical activity in Arabic. The Syrians retained their own independent culture; as late as the 13th century their language was used by the converted Jew Bar Hebraeus, "Son of the Hebrew" (also known as Gregorius or Abū al-Faraj), an encyclopaedist, philosopher, and theologian, who expounded all the works of Aristotle in his *Kĕthabhā dhē-ḥēwath ḥekhmĕthā* (*Book of the Cream of Wisdom*), elaborating many sections on the basis of the Greek and Arabic Aristotelians.

In the 9th century the Arab al-Kindī was the first notable scholar to use the Arabic language in a general introduction of mainly Aristotelian philosophy. In the following century the Turkish Muslim al-Fārābī produced a more specialized study in which he commented upon and expounded the books of logic and attempted to establish the relationship between philosophy and Islām. It was through the writings of Avicenna and Averroës, however,

Contributions of Avicenna and Averroës

that Aristotle's thought became an integral part of lay Arabic culture.

Early in the 11th century, the Arab Avicenna (Ibn Sīnā) made Aristotle's philosophy the foundation of an original system of his own. For this he also found inspiration in a group of Plotinian texts that had been translated into Arabic under the title "Theology of Aristotle." Aristotle became, in Avicenna's hands, a much more systematic and coherent thinker than he really had intended to be; problems and solutions that were, at best, hinted at by Aristotle (*e.g.*, the distinction between essence and existence or the relation between possible and necessary existence) were among the distinctive marks of Avicenna's own work.

For the Spanish Arab Averroës (Ibn Rushd) in the 12th century, Aristotle was "the measure and model offered by nature to show the ultimate perfection of man." He held that philosophy, specifically Aristotelian philosophy, was and taught truth; revelation or revealed religion was a debased philosophy for the simple. Averroës dissected Aristotle's works, analyzing and reconstructing them with a fine scholarly and philosophical sense and an incredible wealth of information derived from previous Greek and Arabic philosophers. He elicited doctrines that are not easily apparent and made them in some cases more compelling than the texts themselves might allow, but he rarely forced his own views onto Aristotle without at least finding some support in the texts themselves. The doctrines concerning the mortality of the individual soul, the eternity of the world, and the existence of a single Mind for the whole human race to the exclusion of individual minds were key doctrines for Averroës; they had some basis—but not much—in the thought of Aristotle.

Relation between philosophy and Judaism

Until the 13th century, Jewish Aristotelianism developed within the Arabic culture of North Africa, Mesopotamia, and Spain. This work was carried out in the Arabic language and distinguished itself for its almost constant concern with the relation between philosophy and Judaism. Many Aristotelian concepts were considered and discussed by Isaac ben Solomon Israeli, a 10th-century Neoplatonist, in his *Kitāb al-ḥudūd* ("Book of Definitions") and *Kitāb al-usṭuqusāt* ("Book of the Elements"). Form and matter were the basis of the metaphysical structure of the Neoplatonic system of Solomon ibn Gabirol, an 11th-century poet and philosopher known as Avicebron. A fully conscious plan of inserting Aristotle—or at least the Aristotle of al-Fārābī and Avicenna—into the intellectual and spiritual life of Judaism was carried out by Abraham ibn Daud of Toledo in the mid-12th century. Moses Maimonides of Córdoba found a way of reconciling the claims of empirical knowledge with those of revelation, which places him into clear contrast with his contemporary Spaniard Averroës, and in so doing he provided a Jewish anticipation of Thomas Aquinas' Christian compromise. His proofs for the existence of God and his acceptance of a theory of creation from eternity were typical of his approach. From the 13th century onward philosophical works, particularly those of Averroës on Aristotle, were being translated into Hebrew; a vast Hebrew literature of "super-commentaries" (those on the works of Aristotle as commented on by Averroës) appeared, and independent works were also produced, notably by Levi ben Gershom (Gersonides), who was faithful to both Maimonides and Averroës. Soon after, however, the more orthodox tradition based upon the Bible and the Talmud prevailed. Aristotelian works by Jews and Hebrew versions of Averroës, translated into Latin, contributed their share to the Italian philosophical movement of the 16th century (see also JUDAISM: *Jewish philosophy*).

The later Latin tradition. *The discovery of Aristotle's works in the Latin West.* Before 1115 only the very short *Categories* and *On Interpretation* were known in Latin, and these two works circulated, from about 800, in a version by Boethius. By 1278 practically the whole of the Aristotelian corpus existed in translations from the Greek, and much of it had a wide circulation. Apart from three other works of logic in translations done by Boethius, which reappeared in about 1115, this wholesale discovery was the result of cultural contacts with Constantinople and a few other Greek centres and the personal initiative of a few scholars. Most notable and first of these was James of Venice, who was in Constantinople and translated the *Posterior Analytics, Physics, On the Soul, Metaphysics,* and several minor texts before or around 1150; other scholars translated anew or for the first time works on ethics, natural philosophy, and logic before 1200. With higher standards of linguistic scholarship, Robert Grosseteste, about 1240, revised and completed the translation of the *Nicomachean Ethics* and translated *On the Heavens* for the first time from the Greek.

Latin translations

The Flemish translator William of Moerbeke, active between about 1255 and 1278, completed the Latin Aristotelian corpus; he was the first to translate the *Politics* and *Poetics* and to give a full and reliable translation of the books on animals; he also translated anew some books of natural philosophy, and he revised several of the older translations. About half of the works were also translated from the Arabic, mainly in Toledo by Gerard of Cremona and Michael Scot, between 1165 and 1230. With two or three exceptions, these translations came after those from the Greek; all had a much more limited circulation and influence. A considerable contribution to the knowledge of Aristotle came from the translations of the ancient commentaries; nearly all of these were made from the Greek.

The view that Aristotle came to be known in Latin by way of the Arabic scholars must be understood as true only in the sense that a number of Aristotelian doctrines—partly transformed in the process—spread in Latin circles from the works of such men as al-Fārābī, Avicenna, and Albumazar before the texts of Aristotle were accessible or had been properly interpreted. Further, there is little truth in a view that in the Latin world in the Middle Ages Aristotle was seen in a Neoplatonic light because Plotinian and Proclan texts translated from the Arabic—namely the *Theologia Aristotelis* ("Theology of Aristotle") and the *Liber de causis* ("Book of Causes")—were ascribed to him.

From the 9th through the mid-13th century. The study of Porphyry's *Isagoge,* of Aristotle's *Categories* and *On Interpretation,* and of theological texts containing Aristotelian elements formed the basis, from the 9th century onward, of logical methodology (dialectic) in a wide number of fields. When applied to problems concerning the Trinity or the Eucharist, or in general to problems concerning individuality and universality of concepts and things, dialectic was perceived as a powerful instrument for clarifying faith or—on the opposite side—for endangering it. For Abelard, the first great Aristotelian of the Middle Ages, dialectic was an essential method for analysis and the discovery of truth. As part of his study, he produced an illuminating account of the linguistic, mental, and objective aspects of universals on the basis of Aristotelian doctrines. Soon thereafter, new developments of Aristotle's theory of language and logic took place, partly as a result of the recently acquired knowledge of his *Sophistical Refutations.*

At the same time, in the later 12th century and during the beginning of the 13th century, Aristotle's physics, cosmology, and metaphysics began to attract attention through the Latin texts both of Arabic works on science and philosophy and of Aristotle's own works, and did so mainly among scientists of the famous medical school at Salerno and among the English philosophers. Around 1190 Alfred of Sareshel used the new texts in his treatise *De motu cordis* ("On the Movement of the Heart"). Between 1210 and 1235 Robert Grosseteste commented on Aristotle's *Physics* and drew on various aspects of Aristotle's natural philosophy for his own scientific and philosophical treatises, and around 1245 Roger Bacon commented on the *Physics* and part of the *Metaphysics.* It would be wrong, however, to try to find in this scholarship the origin of modern experimental science, which is rather to be found in the study of ancient and more recent mechanics, medicine, and technology or in original inventiveness.

Reactions to Aristotelianism at the University of Paris

The introduction of the new Aristotle met with difficulties in Paris. The impact of non-Christian Aristotelian and Arabic philosophy engendered fears, doubts, and suspicions. Although the masters at Paris were free to teach Aristotle's logic, which was value free, and although no obstacle was put in the way of lecturing on any of Aristotle's

works at the universities of Oxford and Toulouse, in the first part of the 13th century the ecclesiastical authorities at Paris imposed a ban on lectures relating to the physics, the metaphysics, and the psychology of Aristotle and his commentators. While this ban succeeded in slowing down some activities it also quickened reactions and aroused strong curiosity; the very demand for some kind of censorship of the works led to more intimate study of them. Certainly by the 1240s the prohibition against teaching Aristotle had become a dead letter at Paris, as can be seen from the fact that Roger Bacon was then commenting on the "dangerous" *Physics* and *Metaphysics.* Shortly thereafter, before 1255, all of Aristotle's philosophical treatises then known had become a required part of the Parisian Master of Arts curriculum, and, around the same time, Albertus Magnus—committed though he was, as a Dominican friar, to safeguarding the purity of faith and dogma—made Aristotle's works an indissoluble part of philosophical and scientific literature in the Latin world. Albertus Magnus announced it as his intention to make all of Aristotle's natural philosophy "intelligible to the Latins." His vast encyclopaedia of secular knowledge and wisdom consisted of an analytical exposition of Aristotle's thought combined with all the information and interpretations that Albertus had gathered from other, mainly Arabic, sources or that he had gained as the product of his own extensive research and speculation. Faced with the danger of being accused of following Aristotle against church dogma, he asserted: "I expound, I do not endorse, Aristotle."

The approach of Albertus' pupil, Thomas Aquinas, to Aristotle was that of a scholar. He wrote numerous detailed commentaries on a variety of Aristotle's works, including the *Physics, Metaphysics, Ethics,* and *Politics;* he analyzed the structure of every section of most works; he tried to discover their organization and to follow the arguments; and he was careful to obtain the best texts and to get from them the genuine meaning. Above all, Thomas Aquinas drew heavily on Aristotle's thought in composing his own masterwork, the *Summa theologiae.* He respected Aristotle's authoritativeness and credited him with reasonableness, even when that was not explicitly justified. Sometimes he drew inferences that went beyond Aristotle's own conclusions, and he allowed himself considerable freedom whenever Aristotle had left loose ends in his attempts to solve difficulties. At these points he often went his own way, without ascribing the new steps to Aristotle but without feeling that he was going against him. Compromises followed; for example, he stepped beyond Aristotle when he argued that the individual soul, although remaining essentially and indissolubly the form of the individual body, is separable from it and immortal. Aristotle's account was stretched almost to the breaking point but it was not transformed. Beyond that point Thomas Aquinas was not a Christian Aristotle but a man of faith and dogma; he divorced himself from Aristotle when necessary and approached closer to St. Augustine, to the Neoplatonists, or to Avicenna.

From the late 13th century through the 15th century. The suspicion that reading Aristotle might lead to heresy became stronger when the closer study of his texts and of Averroës' interpretations enhanced the admiration for The Philosopher and increased the following of The Commentator, as these two thinkers were known respectively. Siger de Brabant was the most redoubtable of many Averroistic Aristotelians. What came to be called Averroism was in fact a tendency to accept genuine or consistent Aristotelian tenets, particularly those concerning the eternity of the world, the unity of the intellect, and the ability of humans to achieve happiness on earth. Ecclesiastical condemnations of propositions considered false or dangerous and threats against the holders of doctrines implied by these propositions gave a more definite status to the Averroists, although many propositions condemned at Paris and Oxford in 1270 and 1277 had nothing to do with Aristotle and little with Averroës. The effect of the condemnations soon became visible: it took the form of a separation between the teaching of "philosophy" in the faculty of arts and the teaching of "truth" in the faculty of theology. This separation became rigid, with the ambiguous result that two "truths"—truth of coherence in philosophical contexts and revealed truth—were thought to coexist.

At the turn of the century, however, Dante's powerful poetical vision could still merge the Averroists' Aristotle, who claimed that natural truths were self-sufficient, and Thomas Aquinas' Aristotle, who endorsed many of the truths of faith. For Dante, as for Averroës, Aristotle was the embodiment of total human knowledge—"the master of them that know." A remarkable index of Dante's commitment to Aristotelianism is the fact that he placed Siger de Brabant, by that time condemned for his Aristotelian heresy, in Paradise. In Aristotle's *Nicomachean Ethics* Dante found moral guidance (he even said that this work "showed man his true happiness"), and in Aristotle's scientific books he found the key to understanding the workings of nature. In some aspects of Averroës' theory of a universal human Intellect combined with the Stoic-Aristotelian principle that all men are by nature citizens of one city, he found the basis of the Empire, seeing it as the one polity (*civilitas*) for the whole human race.

Views of Scotus and Ockham; Averroism

The 14th century was no less Aristotelian than the 13th. Some scholars have indeed claimed that Aristotelianism collapsed, but such an assertion does not take into account the non-Aristotelian components of previous philosophies and the permanent acceptance of Aristotelian doctrines in the new ones. Form, matter, causality, and the idea of a universe in which events occurred with regularity but were not necessitated provided the Aristotelian frame of the system of Duns Scotus. The nominalism (or "terminism") of William of Ockham, an English Franciscan, his rejection of "useless entities," his metaphysics of a world of individual self-contained things, and his conceptualism gave neat, though extreme, expression to Aristotle's theory of language, the economy of nature, and the primacy of individuals in existence and of universals in intellectual knowledge. He followed Aristotle closely in his views on the scientific coordination of notions. He was more faithful to Aristotle than either Thomas Aquinas or Averroës when he said that Aristotle did not give a clear lead on the question of the immortality of the soul. The various schools of Scholastic philosophy—Thomism, Scotism, Ockhamism—that asserted themselves in the 14th century and that lived on had a common Aristotelian basis, but they had different ways of interpreting it (see also PHILOSOPHICAL SCHOOLS AND DOCTRINES, MAJOR WESTERN: *Scholasticism;* CHRISTIANITY: *Christian philosophy*).

Averroistic Aristotelianism flourished in this century in connection with, or independently of, the other trends. The Italian medical faculties at Bologna and Padua were lively centres of logical and philosophical studies; for example, Peter of Abano, a professor of medicine at Padua who had been trained at Paris, pushed Aristotle's cosmology to the brink of determinism in human affairs and used his logic to suggest that Christ's death was only apparent. Political science, which had been a field for lofty speculations or restrained exercises in the analysis and exposition of texts, became important for those who practiced politics and those who wanted to satisfy, under the aegis of Aristotle's doctrine, the potentialities of human beings for happiness. John of Paris wanted France to be self-sufficient, self-controlling, and without interference from the pope; John of Jandun, a successor of Siger de Brabant, upheld Aristotle's *Politics* in all its worldliness; and Marsilius of Padua, John of Jandun's friend in Paris, followed Aristotle in his insistence that government had no supernatural origins but arose naturally from the needs of the governed and that priests should be considered in the same way as members of a guild in a city, without special privileges.

Perhaps with less attachment to the details of Aristotle's doctrines and with a keen critical sense, the Mertonians, a group of logician-philosophers based in Merton College, Oxford (*e.g.,* Thomas Bradwardine, William of Heytesbury), and encyclopaedists, scientists, and philosophers in France (*e.g.,* Jean Buridan and Nicholas Oresme) made laborious efforts to express science wholly in terms of mathematics, to quantify changes in quality, and to determine the nature of continuity in movement and the acceleration and speed of falling bodies. Their starting points were the

Physics and the other texts of Aristotle. In a similar (almost mathematical) spirit, many of the same men carried logic even further than Ockham had done into the fields of logical calculus, paradoxes, and sophisms. Thus one may say that Aristotle was not abandoned but expanded.

Modern developments. *From the Renaissance to the 18th century.* In the 15th century Italy became the focal point at which various forms of Aristotelianism converged. Certain links between Italian universities and religious schools and the University of Paris had already flourished for a long time. In the late 14th century Paolo Nicoletti (Paulus Venetus) returned from Oxford to Padua after having absorbed the new logic and physics of the Mertonians and the radical nominalism of Ockham and after having increased his acquaintance with the French Averroistic trend; works by the Englishmen and by Paolo were textbooks in Italian universities for many generations. At the end of the century a number of Spanish and Italian Jews were passing on, in Latin, still more texts of Averroës on Aristotle, as well as the Jews' own recent contributions to Aristotelian learning.

ristote-
anism
1 15th-
entury
aly

A more spectacular contribution of books, linguistic and didactic competence, and stimulating debates came with an influx of Greek scholars into the Western sphere. They were attracted by the humanists' craving for classical learning, the theological discussions between Orthodox and Roman Catholic leaders, and the relative freedom offered by the Republic of Venice and by Florence to those who were taking refuge from Turkish domination. Many manuscripts were taken to Italy, and many were transcribed in Italy by the Greeks, who also taught the Greek language to the Italian scholars. An editorial masterpiece by Aldus Manutius, an early printer, publisher, and editor, at the end of the 15th century made accessible to many almost the complete Greek corpus of Aristotle's works. A great number of Greek and Latin scholars—such as Bessarion, John Argyropoulos, Leonardo Bruni, and Lorenzo Valla—produced new translations of those texts; others translated many works on Aristotle previously unknown in Latin.

As soon as printing had been established (that is to say, by the late 15th century), editorial activity was directed to the production of many complete as well as partial editions of the Latin versions of Aristotle and Averroës in both their older and newer versions from the Greek, the Arabic, and the Hebrew. At the universities of Padua and Bologna and at Ferrara and Venice, Averroists such as Agostino Nifo and Nicoletto Vernia and independent interpreters such as Pietro Pomponazzi were dominating the philosophical scene. For Pomponazzi, Aristotle, whether right or wrong, had to be studied directly by way of his own works and not by way of his interpreters; yet he did not think that Aristotle had a monopoly on knowledge, and for this reason his mistakes concerning facts needed to be exposed.

There were others who followed Aristotle in his vast scientific achievements or searched his works for a clearer formulation of scientific methods. It was this scientific spirit that kept alive the interest in Aristotle's methodology and in his philosophy of nature down to the time, in the 17th century, when William Harvey, the English physician who discovered the circulation of the blood, was lecturing on Aristotle's books on animals and Galileo was writing on science and logic.

nfluence
on
nethod-
ology,
osmology,
neta-
physics,
nd politics

In a less apparent form, Aristotelianism, still strongly entrenched in most European schools, continued to have its effect on the most modern philosophers. The methodology of Francis Bacon, English philosopher, scientist, and statesman, grew out of it, and his basic metaphysical concepts were borrowed from Aristotle, although he was critical of the distorted version of Aristotelianism in the academic circles of his day. The Polish astronomer Copernicus was still attached to the perfection of circular movements. Gottfried Wilhelm Leibniz, the German Rationalist and mathematician, not only admired Aristotle's logic but also built his own metaphysics of individuals ("monads") around the theory of matter and form. Like Aristotle, political theorists such as Jean Bodin in France carried on their inquiries into the nature of the state by studying existing organizations and their natural backgrounds.

In the literary field, Aristotle's *Poetics,* practically unknown until 1500, was now read and analyzed in both the Greek and Latin versions; its doctrines were compared and partly made to harmonize with the then-prevailing views of the ancient Roman poet Horace, and Aristotle's view that art imitates nature prevailed for many over the conflicting theory that stressed the creativity of the poet. The doctrine of the unities of action, place, and time—though actually a later development resulting from forced interpretations of Aristotle—ruled over the work of many writers of tragedies (*e.g.,* Gian Giorgio Trissino in Italy, Jean Racine and Pierre Corneille in France, and, to a certain extent, Goethe in Germany). Many critics (including the English critics from Sir Philip Sidney to Matthew Arnold) accepted those rules, although few English poets—the great exception was John Milton—welcomed them. A lesser influence was exercised by Aristotle's *Rhetoric* outside the field of systematic theory.

Scholasticism and Aristotle's detractors

Scholasticism in these centuries belonged to the history of Aristotelianism. All over western and central Europe and also in Spanish America the continuance of Scholasticism ensured that higher education remained generally within an Aristotelian framework. Remarkable work was produced by Scholastics in the fields of commentaries and of detailed interpretation; Pedro de Fonseca, the "Portuguese Aristotle," in the 16th century and Sylvester Maurus, author of short but pithy commentaries on all of Aristotle's works, in Rome in the 17th are noteworthy examples. Insofar as the different Scholasticisms were living and interesting philosophical movements, however, they had more to do with newer philosophies than with Aristotle.

Martin Luther's rebellion against Rome, on the other hand, involved a rebellion against Scholastic philosophy and its distorted Aristotelian structure, although not against Aristotle. In fact, when Luther's follower Philipp Melanchthon undertook to reorganize the curriculum for higher education, a more genuine, humanistic Aristotle emerged as the great master of philosophy, independent of theology. Once again, as in the early 13th century in Paris, Aristotle took pride of place, particularly in the realms of logic and ethics, and to a lesser extent in metaphysics and natural philosophy.

The anti-Aristotelianism of the 16th to 18th century touched only a small part of the real Aristotle. Partly it was a reaction against Scholasticism, as though this had faithfully represented Aristotle's own philosophy. Thus, Aristotle was wrongly accused of extreme formalism, irresponsible use of syllogisms consisting of empty or irrelevant concepts, a multiplication of pseudo-real entities, and the application of "scientific" methods to facts that could be vouched for only by faith. For other critics the whole of Aristotle's canon stood condemned because of his unsatisfactory account of local movement and the consequences it had in the areas of mechanics, dynamics, cosmology, and astronomy. His downfall in the 17th century was the result, above all, of his failure to create, in the 4th century BC, a language that allowed him to describe the forms of things and events (*i.e.,* their knowable aspects) in mathematical formulas and of his failure to lay sufficient stress, in his philosophy of experience, on the need for experiments.

Modern critical editions, translations, and studies

The 19th and 20th centuries. The anti-Aristotelian movement was countered mainly by historical and philological scholarship. As Friedrich Adolf Trendelenburg, a German philosopher, saw it, Aristotle's personality and works must be known as exactly as possible because he provides the indispensable historical basis of any serious philosophy. Such a type of study had declined after the great achievements of the 16th century. After the work done between the first new learned edition of the collected Greek texts of Aristotle by J.G. Buhle (1791–93) and a vast collection of all documentary material in the Aristoteles-Archiv at Berlin (which began in 1965), there is little, if anything, that remains to be discovered concerning the original and deteriorated forms of Aristotle's traditional corpus. A monumental edition sponsored by the Prussian Academy from 1831 to 1870 became the ba-

sis for almost innumerable critical editions of individual works. A rich crop of fragments, which were identified and edited in the last centuries, brought to light previously almost unknown aspects of Aristotle's early activity. And in 1890 a papyrus was discovered in Egypt that contained most of the otherwise lost *Constitution of Athens.* European and American academies have sponsored the editing of ancient and medieval commentaries and translations in Greek, Latin, Arabic, and Hebrew. Historical, philological, and philosophical exegesis has explored in great detail the contents and background of most of Aristotle's writings. Translations of all the works into English, German, and French and of many of them into most of the other European languages as well as into Hebrew, Arabic, and Japanese have made Aristotle widely accessible. Historians of ideas have investigated Aristotle's relationship to Plato and to the Greece of his day, his influence in following ages, and his own philosophical development.

Philosophical Aristotelianism has been mainly confined to the German schools established by Trendelenburg and Franz Brentano. Trendelenburg was concerned to effect a revaluation of Aristotle's metaphysics in the face of German idealism; he had a measure of influence in the United States on such thinkers as Felix Adler, George Sylvester Morris, and John Dewey. Aristotle's theories of being and knowledge formed the point of departure for Brentano's "descriptive psychology" and his doctrine of human experience, and they also contributed to Edmund Husserl's phenomenology. Outside Germany, J.-G.-F.-L. Ravaisson-Mollien, a spiritualist philosopher, and Sir David Ross, editor and translator of Aristotle's works, acknowledged a debt to Aristotle, respectively, for their metaphysics and ethics; and the reestablishment of Thomas Aquinas, by Pope Leo XIII in 1879, as the great doctor of the church increased the interest in Aristotle and in his influence on the history of Christian thought. Contemporary philosophy in the Anglo-Saxon world is often associated with a keen interest in Aristotle (nor is he entirely neglected in other philosophical traditions), and the name of the Aristotelian Society (London) reflects the view that good philosophy must be practiced in the spirit of Aristotle.

(L.M.-P./Ed.)

MAJOR WORKS

LOGIC: These six works are known collectively as the *Organon: Katēgoriai* (*Categories*); *Peri hermēneias* (Latin trans., *De Interpretatione;* Eng. trans., *On Interpretation*); *Analytika protera* (*Prior Analytics*); *Analytika hystera* (*Posterior Analytics*); *Topika* (*Topics*); and *Peri sophistikōn elegchōn* (*Sophistical Refutations*).

NATURAL PHILOSOPHY AND NATURAL SCIENCE: *Physikē* (*Physics*); *Peri ouranou* (*On the Heavens*); *Peri geneseōs kai phthoras* (*On Generation and Corruption; On Coming to Be and Passing Away*); *Meteōrologika* (*Meteorology*); *Peri kosmou* (spurious; Latin trans., *De mundo;* Eng. trans., *On the Universe*); *Peri ta zōa historiai* (*History of Animals*); *Peri zōōn moriōn* (*Parts of Animals*); *Peri zōōn kinēseōs* (*Movement of Animals*); *Peri poreias zōōn* (*Progression of Animals*); *Peri zōōn geneseōs* (*Generation of Animals*); and the works collectively known as the *Parva Naturalia: Peri aisthēseōs* (*On the Senses and Their Objects; On Sense and Sensible Objects*); *Peri mnēmēs kai anamnēseōs* (*On Memory and Recollection*); *Peri hypnou kai egrēgorseōs* (*On Sleep and Waking*); *Peri enypniōn* (*On Dreams*); *Peri tēs kath hypnon mantikēs* (*On Divination in Sleep; On Prophecy in Sleep*); *Peri makrobiotētos kai brachybiotētos* (*On Length and Shortness of Life*); *Peri neotētos kai gērōs* (*On Youth and Old Age*); *Peri zōēs kai thanatou* (*On Life and Death*); *Peri anapnoēs* (*On Respiration*); and *Peri pneumatos* (spurious; *On Breath*).

PSYCHOLOGY: *Peri psychēs* (Latin trans., *De anima;* Eng. trans., *On the Soul*).

METAPHYSICS: *Ta meta ta physika* (*Metaphysics*).

ETHICS AND POLITICS: *Ēthika Nikomacheia* (*Nichomachean Ethics*); *Ēthika Eudēmeia* (*Eudemian Ethics*); *Ēthika megala* (spurious; Latin and Eng. trans., *Magna moralia*); *Peri aretōn kai kakiōn* (spurious; *On Virtues and Vices*); *Politika* (*Politics*); *Oikonomika* (spurious; *Economics*); and *Athēnaiōn politeia* (incomplete; *Constitution of Athens*).

AESTHETICS AND LITERATURE: *Technē rhētorikē* (*Rhetoric*); *Rhētorikē pros Alexandron* (spurious; *Rhetoric to Alexander*); and *Peri poiētikēs* (incomplete; *Poetics*).

OTHER WORKS: These remain in the corpus but are believed by scholars to be falsely attributed to Aristotle: *Peri chrōmatōn* (*On Colours*); *Peri akoustōn* (*On Things Heard*); *Physiognōmonika* (*Physiognomonics*); *Peri phytōn* (*On Plants*); *Peri thaumasiōn akousmatōn* (*On Marvellous Things Heard*); *Mēchanika* (*Mechanics*); *Problēmata* (*Problems*); *Peri atomōn grammōn* (*On Indivisible Lines*); *Anemōn theseis kai prosēgoriai* (*The Situations and Names of Winds*); and *Peri Melissou, peri Xenophanous, peri Gorgiou* (*On Melissus, Xenophanes, Gorgias*).

TEXTS: The standard edition of the Greek text is the Berlin Academy edition, *Aristotelis Opera,* ed. by Immanuel Bekker, 5 vol. (1831–70, reissued 5 vol. in 4, 1960–61); and the standard edition of the fragments is *Aristotelis qui Ferebantur Librorum Fragmenta,* ed. by Valentin Rose (1870, reissued 1967). For most works these texts have been superseded by more recent editions, notably by the volumes of the Teubner series, the Oxford Classical Text series, the Loeb Classical Library series (with English translations), and the Budé series (with French translations). The medieval Latin translations of Aristotle are being printed in *Aristoteles Latinus,* ed. by L. Minio-Paluello (1939–); see also *Aristotelis opera cum Averrois commentariis,* 9 vol. in 11 (1562–74, reissued 1962). In addition there is much useful information of a textual nature in the early Greek commentaries, the most important of which have been published in *Commentaris in Aristotelem Graeca,* 23 vol. in 46 (1882–1909). An invaluable aid to the study of Aristotle is Hermann Bonitz, *Index Aristotelicus* (1870, reprinted 1955).

RECOMMENDED EDITIONS: Numerous English translations of the major treatises are available. The standard complete edition is Jonathan Barnes (ed.), *The Complete Works of Aristotle: The Revised Oxford Translation,* 2 vol. (1984). Of the many editions of and commentaries on individual works, the following may be mentioned: J.L. Ackrill (trans.), *Categories, and De Interpretatione* (1963, reprinted 1978); W.D. Ross (ed.), *Prior and Posterior Analytics* (1949, reprinted 1957); Jonathan Barnes (trans.), *Aristotle's Posterior Analytics* (1976); W.D. Ross (ed.), *Physics* (1950, reprinted 1977); W. Charlton (trans.), *Aristotle's Physics: Books 1 & 2* (1970); Edward Hussey (trans.), *Aristotle's Physics, Books III and IV* (1983); Harold H. Joachim (ed.), *Aristotle on Coming-to-Be and Passing-Away* (*De Generatione et Corruptione*) (1922, reprinted 1982); C.J.F. Williams (trans.), *Aristotle's De Generatione et Corruptione* (1982); R.D. Hicks (trans.), *De Anima* (1907, reprinted 1976); W.D. Ross (ed.), *Parva Naturalia* (1955, reprinted 1970); G.R.T. Ross (trans.), *De Sensu and De Memoria* (1906, reprinted 1973); Richard Sorabji, *Aristotle on Memory* (1972); D.M. Balme (trans.), *Aristotle's De Partibus Animalium I; and, De Generatione Animalium I* (1972); Martha Craven Nussbaum (ed. and trans.), *Aristotle's De Motu Animalium* (1978); W.D. Ross (ed.), *Metaphysics,* 2nd ed. (1928); Christopher Kirwan (trans.), *Aristotle's Metaphysics* (1971), Books 4–6; Myles Burnyeat (ed.), *Notes on Book Zeta of Aristotle's Metaphysics* (1979), and *Notes on Books Eta and Theta of Aristotle's Metaphysics* (1984); Julia Annas (trans.), *Aristotle's Metaphysics* (1976), Books 13–14; J.A. Stewart, *Notes on the Nichomachean Ethics of Aristotle* (1892, reprinted 1973); Michael Woods (trans.), *Aristotle's Eudemian Ethics: Books I, II, and VIII* (1982); W.L. Newman (ed.), *The Politics of Aristotle,* 4 vol. (1887–1902, reprinted 1973); Richard Robinson (trans.), *Politics, Books III and IV* (1962); Edward Meredith Cope (ed.), *The Rhetoric of Aristotle,* 3 vol. (1877, reprinted 1973); D.W. Lucas (ed.), *Poetics* (1968, reprinted 1980); P.J. Rhodes (trans.), *The Athenian Constitution* (1984); and Ingemar Düring (ed.), *Protrepticus: An Attempt at Reconstruction* (1961).

BIBLIOGRAPHY

Aristotle. *General works:* There are several good introductions to Aristotle's thought: JONATHAN BARNES, *Aristotle* (1982); J.L. ACKRILL, *Aristotle the Philosopher* (1981); D.J. ALLAN, *The Philosophy of Aristotle,* 2nd ed. (1970, reissued 1978); G.E.R. LLOYD, *Aristotle: The Growth and Structure of His Thought* (1968); W.D. ROSS, *Aristotle,* 5th ed. (1949, reprinted 1977); and FRANZ BRENTANO, *Aristotle and His World View* (1978; originally published in German, 1911). For a comprehensive survey see INGEMAR DÜRING, *Aristoteles: Darstellung und Interpretation seiner Denkens* (1966). Two of the most influential books on Aristotle written in the 20th century are WERNER W. JAEGER, *Aristotle: Fundamentals of the History of His Development,* 2nd ed. (1948, reissued 1962; originally published in German, 1923), which advances a theory of the development of Aristotle's thought; and HAROLD CHERNISS, *Aristotle's Criticism of Plato and the Academy* (1944, reissued 1962), which discusses, in a uniformly critical spirit, Aristotle's knowledge and assessment of Plato's work.

Most of the scholarly work done on Aristotle appears in articles rather than in books. There is a useful anthology: JONATHAN BARNES, MALCOLM SCHOFIELD, and RICHARD SORABJI (eds.), *Articles on Aristotle,* 4 vol. (1975–79). The proceedings of the triennial *Symposium Aristotelicum* contain some of the most up-to-date work.

Life: For all aspects of Aristotle's life, see INGEMAR DÜRING,

Aristotle in the Ancient Biographical Tradition (1957); for his writings, see PAUL MORAUX, *Les Listes anciennes des ouvrages d'Aristote* (1951); for the history of the Lyceum, see JOHN PATRICK LYNCH, *Aristotle's School: A Study of a Greek Educational Institution* (1972); and PAUL MORAUX, *Der Aristotelismus bei den Griechen: Von Andronikos bis Alexander von Aphrodisias,* 2 vol. (1973–84).

Thought: (*Logic*): On Aristotle's formal syllogistic the classic study is JAN ŁUKASIEWICZ, *Aristotle's Syllogistic from the Standpoint of Modern Formal Logic,* 2nd ed. enlarged (1957, reprinted 1967); and the standard work is GÜNTHER PATZIG, *Aristotle's Theory of the Syllogism: A Logico-Philological Study of Book "A" of the "Prior Analytics"* (1969; originally published in German, 2nd ed. 1963). A less formal account can be found in ERNEST KAPP, *Greek Foundations of Traditional Logic* (1942, reissued 1967). See also JONATHAN LEAR, *Aristotle and Logical Theory* (1980); and, for the *Topics,* the introduction to JACQUES BRUNSCHWIG (trans.), *Topiques* (1967). On the development of Aristotle's ideas in logic, see FRIEDRICH SOLMSEN, *Die Entwicklung der aristotelischen Logik und Rhetorik* (1929, reprinted 1975). For Aristotle's modal logic, see STORRS MCCALL, *Aristotle's Modal Syllogisms* (1963); and for less formal treatments of his ideas about modality, see JAAKKO HINTIKKA, *Time & Necessity: Studies in Aristotle's Theory of Modality* (1973); and SARAH WATERLOW, *Passage and Possibility: A Study of Aristotle's Modal Concepts* (1982). On the connection between Aristotle's logic and his scientific methodology, see J.M. LE BLOND, *Logique et méthode chez Aristote: étude sur la recherche des principes dans la physique aristotélicinne,* 2nd ed. (1970).

(*Theory of science*): The standard introduction to the *Physics* is AUGUSTE MANSION, *Introduction à la physique aristotélicienne,* 2nd rev. ed. (1946); see also FRIEDRICH SOLMSEN, *Aristotle's System of the Physical World: A Comparison with His Predecessors* (1960, reprinted 1970). Among the most stimulating recent studies are WOLFGANG WIELAND, *Die aristotelische Physik;* 2nd rev. ed. (1970); RICHARD SORABJI, *Necessity, Cause, and Blame: Perspectives on Aristotle's Theory* (1980), and *Time, Creation, and the Continuum: Theories in Antiquity and the Early Middle Ages* (1983); and SARAH WATERLOW, *Nature, Change, and Agency in Aristotle's "Physics"* (1982).

(*Biology*): It is still worth consulting D'ARCY WENTWORTH THOMPSON, *On Aristotle as a Biologist* (1913); the best recent study is PIERRE PELLEGRIN, *La Classification des animaux chez Aristote: statut de la biologie et unité de l'aristotélisme* (1982).

(*Psychology*): FRANZ BRENTANO, *The Psychology of Aristotle: In Particular His Doctrine of the Active Intellect* (1977; originally published in German, 1867), remains one of the most valuable works in this area. The standard study of the development of Aristotle's views on the soul is FRANÇOIS NUYENS, *L'Évolution de la psychologie d'Aristote* (1948, reissued 1973). Among more recent works are EDWIN HARTMAN, *Substance, Body, and Soul: Aristotelian Investigations* (1977); and DAVID CHARLES, *Aristotle's Philosophy of Action* (1984).

(*Metaphysics*): There are two large and comprehensive volumes: JOSEPH OWENS, *The Doctrine of Being in the Aristotelian Metaphysics: A Study in the Greek Background of Medieval Thought,* 3rd ed. rev. (1978); and PIERRE AUBENQUE, *Le Problème de l'être chez Aristote: essai sur la problèmatique aristotélicienne,* 4th ed. (1977). There is a helpful brief introduction in G.E.M. ANSCOMBE and P.T. GEACH, *Three Philosophers* (1961, reprinted 1963). On special aspects of the metaphysics, see FRANZ BRENTANO, *On the Several Senses of Being in Aristotle* (1975, reprinted 1981; originally published in German, 1862); R.M. DANCY, *Sense and Contradiction: A Study in Aristotle* (1975); SUZANNE MANSION, *Le Jugement d'existence chez Aristote,* 2nd ed. rev. (1976); and A.C. LLOYD, *Form and Universal in Aristotle* (1981).

(*Ethics*): W.F.R. HARDIE, *Aristotle's Ethical Theory,* 2nd ed. (1980), provides a helpful companion. Some of the best recent work is collected in AMÉLIE OKSENBERG RORTY (ed.), *Essays on Aristotle's "Ethics"* (1980). See also STEPHEN R.L. CLARK, *Aristotle's Man: Speculations upon Aristotelian Anthropology* (1975), reprinted 1983); JAMES J. WALSH, *Aristotle's Conception of Moral Weakness* (1963); JOHN M. COOPER, *Reason and Human Good in Aristotle* (1975); ANTHONY KENNY, *The Aristotelian Ethics: A Study of the Relationship Between the Eudemian and Nicomachean Ethics of Aristotle* (1978), and *Aristotle's Theory of the Will* (1979); and TROELS ENGBERG-PEDERSEN, *Aristotle's Theory of Moral Insight* (1983, reprinted 1985).

(*Politics*): The standard discussion is ERNEST BARKER, *The Political Thought of Plato and Aristotle* (1906, reissued 1959); see also R.G. MULGAN, *Aristotle's Political Theory: An Introduction for Students of Political Theory* (1977). On Aristotle's historical interests, see GEORGE HUXLEY, *On Aristotle and Greek Society: An Essay* (1979).

(*Rhetoric*): WILLIAM M.A. GRIMALDI, *Studies in the Philosophy of Aristotle's Rhetoric* (1972). On the psychological aspects of rhetoric, see W.W. FORTENBAUGH, *Aristotle on Emotion: A Contribution to Philosophical Psychology, Rhetoric, Poetics, Politics, and Ethics* (1975).

(*Poetics*): JOHN JONES, *On Aristotle and Greek Tragedy* (1962, reissued 1980); and RICHARD JANKO, *Aristotle on Comedy: Towards a Reconstruction of "Poetics" II* (1984).

Aristotelianism. *Aristotelianism as covered in general histories:* Extensive treatment of Aristotelianism is included in the fundamental history of philosophy by FRIEDRICH UEBERWEG, *A History of Philosophy, from Thales to the Present Time,* 2 vol. (1872–74, reprinted 1972; originally published in German, 4th ed., 3 vol., 1871–73), with a vast bibliography. Useful histories of philosophy, general or partial, are FREDERICK C. COPLESTON, *A History of Philosophy,* 9 vol. (1946–74); MEYRICK H. CARRÉ, *Phases of Thought in England* (1949, reprinted 1972), which is particularly good on Aristotelianism; JOHN HERMAN RANDALL, *The Career of Philosophy,* 2 vol. (1962–65, reissued 1970), imaginative and stimulating; and ÉTIENNE GILSON, *History of Christian Philosophy in the Middle Ages* (1955, reissued 1980), a personal interpretation, with documentation and bibliography.

Aristotelianism in various periods or cultures: INGEMAR DÜRING, "Von Aristoteles bis Leibniz: Einige Hauptlinien in der Geschichte des Aristotelismus," *Antike und Abendland,* 4:118–154 (1954), mostly on Greek and medieval Aristotelianism; LORENZO MINIO-PALUELLO, *Opuscula: The Latin Aristotle* (1972), a collection of articles and essays concerning the Latin transmission of Aristotle's works; and RICHARD MCKEON, "Aristotelianism in Western Christianity," in JOHN THOMAS MCNEILL, MATTHEW SPINKA, and HAROLD R. WILLOUGHBY (eds.), *Environmental Factors in Christian History,* pp. 206–231 (1939, reissued 1970). On Boethius, see HENRY CHADWICK, *Boethius: The Consolation of Music, Logic, Theology, and Philosophy* (1981); and MARGARET GIBSON (ed.), *Boethius: His Life, Thought, and Influence* (1981). (*On Greek Aristotelianism*): EDUARD ZELLER, *Die Philosophie der Griechen,* vol. 2, *Sokrates, Plato, Aristoteles* (1846), and vol. 3, parts 1–2, *Die nacharistotelische Philosophie* (1852), parts of which have been translated from various editions: *Aristotle and the Earlier Peripatetics,* trans. by B.F.C. COSTELLOE and J.H. MUIRHEAD (1897); and *A History of Eclecticism in Greek Philosophy,* trans. by S.F. ALLEYNE (1883), fundamental for the first eight centuries; PAUL MORAUX, *D'Aristote à Bessarion: trois exposés sur l'histoire et la transmission de l'aristotélisme grec* (1970); "Rückblick: Der Peripatos in vorchristlicher Zeit," in FRITZ R. WEHRLI (ed.), *Die Schule des Aristoteles,* vol. 10, pp. 93–128 (1959); KLAUS OEHLER, "Aristotle in Byzantium," *Greek, Roman, and Byzantine Studies,* 5(2):133–146 (Summer 1964); and BASILE TATAKIS, *La Philosophie byzantine,* 2nd ed. (1959), an extensive survey, with a rich bibliography. (*On Latin Aristotelianism*): FERNAND VAN STEENBERGHEN, *Aristotle in the West: The Origins of Latin Aristotelianism,* 2nd ed. (1970; originally published in French, 1946), a scholarly survey of contemporary studies; RICHARD J. LEMAY, *Abu Ma'shar and Latin Aristotelianism in the Twelfth Century: The Recovery of Aristotle's "Natural Philosophy" Through Arabic Astrology* (1962), important contributions; D.A. CALLUS, "Introduction of Aristotelian Learning to Oxford," *Proceedings of the British Academy,* 29:229–281 (1943), original, fundamental research; PAUL MORAUX *et al., Aristote et Saint Thomas d'Aquin* (1957), which includes some of the most reliable studies on the subject; M.-D. CHENU, *La Théologie comme science au XIII^e siècle,* 3rd ed. rev. (1957, reissued 1969), on the interplay of Aristotelian methodology and dogma; and HASTINGS RASHDALL, *The Universities of Europe in the Middle Ages,* new ed., 3 vol. (1936, reissued 1969), basic for Aristotelianism in the schools. (*On Syriac, Arabic, and Jewish Aristotelianism*): ANTON BAUMSTARK, *Geschichte der syrischen Literatur mit Ausschluss der christlich-palästinensischen Texte* (1922, reprinted 1968), with exhaustive factual information and a bibliography; ANTON BAUMSTARK (ed.), *Aristoteles bei den Syrern vom 5. bis 8. Jahrhundert: Syrische Texte* (1900, reprinted 1975), specialized research and texts; T.J. DE BOER, *The History of Philosophy in Islam* (1903, reprinted 1983; originally published in German, 1901); CARL BROCKELMANN, *Geschichte der arabischen Litteratur,* 2 vol. (1898–1902), exhaustive factual information and bibliography; F.E. PETERS, *Aristoteles Arabus: The Oriental Translations and Commentaries of the Aristotelian Corpus* (1968), from Syriac and Arabic; R. WALZER, "Arisṭūṭālīs," in *The Encyclopaedia of Islam,* new ed., vol. 1, pp. 630–633, and related articles; ISAAC HUSIK, *A History of Mediaeval Jewish Philosophy* (1916, reissued 1974); GEORGES VAJDA, *Introduction à la pensée juive du Moyen Age* (1947), limited in scope, with a good bibliography; HARRY A. WOLFSON, "Revised Plan for the Publication of a *Corpus Commentariorum Averrois in Aristotelem,*" *Speculum,* 38(1):88–104 (January 1963), complete lists of Arabic, Latin, and Hebrew texts of Averroës' commentaries, and *Crescas' Critique of Aristotle: Problems of Aristotle's "Physics" in Jewish and Arabic Philosophy* (1929, reprinted 1971); and "Aristot-

le," in *Encyclopaedia Judaica,* vol. 3, col. 445–449 (1971), and related articles. (*On Renaissance and later Aristotelianism*): PAUL OSKAR KRISTELLER, *Renaissance Philosophy and the Mediaeval Tradition* (1966), a brilliant survey, with bibliography, and *Studies in Renaissance Thought and Letters* (1956, reprinted 1969), many relevant essays; BRUNO NARDI, *Saggi sull'Aristotelismo padovano dal secolo XIV al XVI* (1958), one of several fundamental works by this author; PETER PETERSEN, *Geschichte der aristotelischen Philosophie in protestantischen Deutschland* (1921, reprinted 1964), and *Die Philosophie Friedrich Adolf Trendelenburgs: ein Beitrag zur Geschichte des Aristoteles im 19. Jahrhundert* (1913); and CHARLES B. SCHMITT, *Aristotle and the Renaissance* (1983).

Aristotelianism in various areas or disciplines: (*On logic*): WILLIAM KNEALE and MARTHA KNEALE, *The Development of Logic* (1962, reprinted 1984), an objective assessment of the Aristotelian and non-Aristotelian elements in the history of logic; and I.M. BOCHEŃSKI, *A History of Formal Logic,* 2nd ed. (1970; originally published in German, 1956), technical, with much bibliography. (*On science*): GEORGE SARTON, *Introduction to the History of Science,* 3 vol. in 5 (1927–48, reprinted 1975), fundamental, with extensive information and bibliography; RENÉ TATON (ed.), *A General History of the Sciences,* 4 vol. (1963–66; originally published in French, 1957–64); ALASTAIR C. CROMBIE, *Robert Grosseteste and the Origins of Experimental Science, 1100–1700* (1953, reissued 1971), which upholds the view of Aristotelian impact on experimental method; ANNELIESE MAIER, *Studien zur Naturphilosophie der Spätscholastik,* 5 vol. (1949–58), fundamental research; and ALEXANDRE KOYRÉ, *Galileo Studies* (1978; originally published in French, 1939), indispensable for a proper evaluation of anti-Aristotelianism. (*On politics*): GEORGE H. SABINE, *A History of Political Theory,* 4th ed. rev. by THOMAS LANDON THORSON (1973); ALEXANDER PASSERIN D'ENTRÈVES, *The Medieval Contribution to Political Thought: Thomas Aquinas, Marsilius of Padua, Richard Hooker* (1939, reprinted 1959); GEORGES DE LAGARDE, *La Naissance de l'esprit laïque au déclin du Moyen Age,* 3rd ed., 5 vol. (1956–70), fundamental for the 14th century; and HORST DREITZEL, *Protestantischer Aristotelismus und absoluter Staat: Die "Politica" des Henning Arnisaeus (ca. 1575–1636)* (1970), excellent, with an extensive bibliography on German Aristotelianism. (*Poetics and rhetoric*): BERNARD WEINBERG, *A History of Literary Criticism in the Italian Renaissance,* 2 vol. (1961, reprinted 1974), containing good surveys concerning Aristotle; LANE COOPER, *The Poetics of Aristotle: Its Meaning and Influence* (1923, reissued 1972); MARVIN T. HERRICK, *The Fusion of Horatian and Aristotelian Literary Criticism, 1531–1555* (1946), and *The Poetics of Aristotle in England* (1930, reprinted 1976), indispensable complements to Cooper's book; and CHARLES S. BALDWIN, *Renaissance Literary Theory and Practice: Classicism in the Rhetoric and Poetic of Italy, France, and England, 1400–1600* (1939, reissued 1959), useful for both poetics and rhetoric.

(A.H.Ao./L.M.-P./Ed.)

Arithmetic

Arithmetic (a term derived from the Greek word *arithmos,* "number") refers generally to the elementary aspects of the theory of numbers, arts of mensuration (measurement), and numerical computation (that is, the processes of addition, subtraction, multiplication, division, raising to powers, and extraction of roots). As is often the case, however, such meaning, has not been uniform in mathematical usage. An eminent German mathematician, Carl Friedrich Gauss, in *Disquisitiones Arithmeticae* (1801), and certain contemporary mathematicians have used the term to include some aspects of number theory that are considerably more advanced. The reader interested in the latter is referred to NUMBER THEORY.

This article is divided into the following sections:

FUNDAMENTAL DEFINITIONS AND LAWS

In a collection (or set) of 1, 2, 3, or, generally, n objects (or elements), the act of determining the number of objects present is called counting. For an empty set, no object is present, and the count yields the number 0. The numbers n thus obtained are called natural numbers; there are some mathematicians who include 0 among the natural numbers, whereas there are others who do not include it.

If n_1 and n_2 are the numbers of objects in the two sets S_1 and S_2 respectively, then those objects in S_1 and S_2 may possibly be paired in such a way that each pair contains one object from S_1 and one from S_2 and no object from either set remains unpaired. The two numbers n_1, n_2 would then be said to be equal; that is, $n_1 = n_2$. The two sets, in such a case, are said to be equivalent. The concept of equivalent sets is basic to the foundations of modern mathematics and has been introduced into primary education, notably as part of the "new math" that has been alternately acclaimed and decried since it appeared in the 1960s. The theory of sets is discussed in the article SET THEORY.

Equivalent sets

Taking the two sets S_1 and S_2 together to form a new set S_3, it is clear that S_3 contains $n_1 + n_2 = n_3$ objects. The number n_3 is called the sum of n_1, n_2; and each of the latter is called a summand. The operation of forming the sum is called addition, the symbol + being read as "plus."

From the definition of counting it is evident that two fundamental laws hold for the operation of addition (see Box, laws 1 and 2), in which the order of the summands can be changed and the order of the operation of addition can be changed, when applied to three summands, without affecting the sum. These are called the laws of

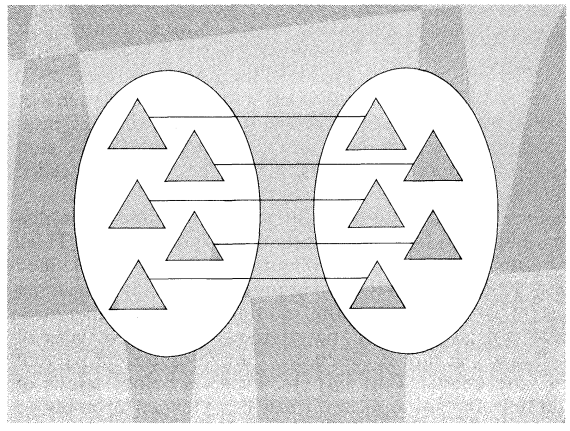


Figure 1: Page from a first grade workbook typical of "new math" might state: "Draw connecting lines from triangles in the first set to triangles in the second set. Are the two sets equivalent in number?"

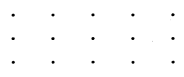
commutativity and associativity for addition. To dispel any notion that these laws are trivially self-evident, it is notable that the first of these, for example, does not hold for the addition of ordinary (finite) rotations. If a closed book that is thin in comparison with its page size is lying flat on a table in an orientation ready to be read, and if a denotes a rotation of the book through 90° about a horizontal axis fixed at the upper edge of the book, and b denotes a similar subsequent rotation about a vertical axis, then adding the rotations $a + b$, in this order, will readily be found to lead to a final position of the book in which the latter is vertical, upside-down, and with the "thin" dimension facing the reader. Performing the sum $b + a$, on the other hand, will cause the book's thin dimension to rest on the table, the front cover facing toward (or away from) the reader. Hence, in this case, $a + b$ is clearly not equal to $b + a$.

ymbols > nd <

If there exists a natural number k such that $a = b + k$, it is said that a is greater than b (written $a > b$), and that b is less than a (written $b < a$). If a and b are any two natural numbers, then it is the case that either $a = b$ or $a > b$ or $a < b$.

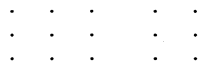
From the above laws it is evident that a repeated sum such as $5 + 5 + 5$ is independent of the way in which the summands are grouped and is written 3×5. Thus, a second binary operation called multiplication is defined. The number 5 is called the multiplicand, the number 3, which denotes the number of summands, is called the multiplier, and the result 3×5 is called the product. The symbol $\times$ of this operation is read "times." If such letters as a and b are used to denote the numbers, the product $a \times b$ is often written $a \cdot b$ or simply ab.

If three rows of five dots each are written, as illustrated below,



It is clear that the total number of dots in the array is 3×5, or 15. This same number of dots can evidently be written in five rows of three dots each, whence $5 \times 3 = 15$. The argument is general, leading to the law (see 3) that the order of the multiplicands does not affect the product, called the commutative law of multiplication. Here again, it is notable that this law does not apply to all mathematical entities. Indeed, much of the mathematical formulation of modern physics, for example, depends crucially on the fact that some entities do not commute.

By the use of a three-dimensional array of dots, it becomes evident that the order of multiplication when applied to three numbers does not affect the product (see 4). Such a law is called the associative law of multiplication. If the 15 dots written above have separated into two sets, namely,



then the first set consists of three columns of three dots each, or 3×3 dots; the second set consists of two columns of three dots each, or 2×3 dots; the sum $(3 \times 3) + (2 \times 3)$ consists of $3 + 2 = 5$ columns of three dots each, or $(3 + 2) \times 3$ dots. In general, one may prove that the multiplication of a sum by a number is the same as the sum of two appropriate products. Such a law (see 5) is called the distributive law.

Subtraction has not been introduced for the simple reason that it can be defined as the inverse of addition. Thus, the difference $a - b$ of two numbers a and b is defined as a solution x of the equation $b + x = a$. If a number system is restricted to the natural numbers, differences need not always exist, but if they do, the five laws of arithmetic, as already discussed, can be used to prove that they are unique. Furthermore, the laws of operations of addition and multiplication can be extended to apply to differences. The natural numbers (including zero) can be extended to include the solution of $1 + x = 0$, that is, the number -1, as well as all products of the form $-1 \times n$, in which n is a natural number. The extended collection of numbers is called the integers, of which the positive integers are the same as the natural numbers (excluding zero). The numbers that are newly introduced in this way are called negative integers.

(1) The commutative law of addition: $a + b = b + a$

(2) The associative law of addition: $a + (b + c) = (a + b) + c$

(3) The commutative law of multiplication: $ab = ba$

(4) The associative law of multiplication: $a(bc) = (ab)c$

(5) The distributive law: $(a + b)c = ac + bc$

(6) $a^m a^n = a^{m+n}$

(7) $(a^m)^n = a^{mn}$

(8) $a^m b^m = (ab)^m$

(9) $a^m \div a^n = a^{m-n} \quad m > n$

Just as a repeated sum $a + a + \cdots + a$ of k summands is written ka, so a repeated product $a \times a \times \cdots \times a$ of k factors is written a^k. The number k is called the exponent, and a the base of the power a^k.

The fundamental laws (see 6–8) of exponents follow easily from the definitions, and other laws (see 9) are immediate consequences of the fundamental ones. The first five laws are often called the five fundamental laws of arithmetic.

THEORY OF DIVISORS

At this point an interesting development occurs, for, so long as only additions and/or multiplications are performed with integers, the resulting numbers are invariably themselves integers; that is, numbers of the same kind as their antecedents. This characteristic changes drastically, however, as soon as division is introduced. Performing division leads to results, called quotients or fractions, which surprisingly include numbers of a new kind, namely, rationals that are not integers. These, though arising from the combination of integers, patently constitute a distinct extension of the natural-number and integer concepts as defined above. By means of the application of the division operation, the domain of the natural numbers becomes extended and enriched immeasurably beyond the integers (see below *Rational numbers*).

Extension of classes of numbers

The preceding illustrates, simply but clearly, one of the proclivities that are often associated with mathematical thought: relatively simple concepts (such as integers), initially based on very concrete operations (for example, counting), are found to be capable of assuming novel meanings, and potential uses, extending far beyond the limits of the concept as originally defined. A similar extension of basic concepts, with even more powerful results, will be found with the introduction of irrationals (see below *Irrational numbers*).

A second example of this is presented by the following: Under the primitive definition, with k equal to either zero or a fraction, a^k would, at first sight, appear to be utterly devoid of meaning. Clarification is needed before writing a repeated product of either zero factors or a fractional number of factors. Yet, limiting attention to the case $k = 0$ here (k = fraction will be considered below), a little reflection shows that a^0 can, in fact, assume a perfectly precise meaning, coupled with an additional and quite extraordinary property. Since the result of dividing any (nonzero) number by itself is unity, it follows that $a^m \div a^m = a^{m-m} = a^0 = 1$. Not only can the definition of a^k be extended to include the case $k = 0$, but the ensuing result also possesses the noteworthy property that it is independent of the particular (nonzero) value of the base a. A similar argument may be given to show that a^k is a meaningful expression even when k is negative,

namely, $a^{-k} = 1/a^k$. The original concept of exponent is thus broadened to a great extent.

Fundamental theory. If three positive integers a, b, and c are in the relation $ab = c$, it is said that a and b are divisors or factors of c, or that a divides c (written $a \mid c$), and b divides c. The number c is said to be a multiple of a and a multiple of b.

The number 1 is called the unit, and it is clear that 1 is a divisor of every positive integer. If c can be expressed as a product ab in which a and b are positive integers each greater than 1, then c is called composite. A positive integer neither 1 nor composite is called a prime. Thus, 2, 3, 5, 7, 11, 13, 17, 19, $\cdots$ are prime numbers. Euclid proved that the number of prime numbers is infinite (*Elements,* book IX, proposition 20).

The fundamental theorem of arithmetic was proved by Gauss in his *Disquisitiones Arithmeticae.* It states that every composite number can be expressed as a product of prime numbers and that, save for the order in which the factors are written, this representation is unique. This theorem follows rather directly from a theorem of Euclid (*Elements,* book VII, proposition 30) to the effect that if a prime divides a product, it divides one of its factors, and the fundamental theorem is therefore sometimes credited to Euclid.

Gauss's contribution

For every finite set $a_1, a_2, \cdots, a^k$ of positive integers, there exists a largest integer d that divides each of these numbers, called their greatest common divisor (GCD). If $d = 1$, the numbers are said to be relatively prime. There also exists a smallest positive integer m that is a multiple of each of the numbers. This is called their least common multiple (LCM).

If $p_1, p_2, \cdots, p_h$ are the distinct primes that divide all of the numbers $a_1, a_2, \cdots, a_k$, and if e_i is the smallest exponent to which p_i occurs in any of them, then (see 10) the product of the powers formed with typical number p_i as base and e_i as exponent is the GCD of $a_1, a_2, \cdots, a_k$. If $p_1, p_2, \cdots, p_i$ are the distinct primes that divide any one or more of the numbers $a_1, a_2, \cdots, a_k$, and if n_i is the largest exponent to which p_i occurs in any of them, then (see 11) the product of the powers formed with typical number p_i as base and n_i as exponent is the LCM of $a_1, a_2, \cdots, a_k$. An example (see 12) is easily constructed. When only two numbers are involved, the GCD and the LCM combine to give the same product as the product of the original numbers.

If a and b are two positive integers, $a > b$, by means of the division algorithm two integers q and r can be determined such that (see 13) a is the sum of q numbers of magnitude b plus a number r that is less than b.

The number q is called the partial quotient (the quotient if $r = 0$), and r is called the remainder. The GCD of a and b is equal to the GCD of b and r. If the division algorithm is applied successively, a remainder 0 must ultimately appear. The last positive remainder is the GCD of a and b. Thus, if $a = 544$, $b = 119$, a simple calculation (see 14) shows that the GCD of 544 and 119 is 17. This process is known as the Euclidean algorithm. By means of it, the GCD can be obtained without first factoring the numbers a and b into prime factors.

(10) $$d = p_1^{e_1} p_2^{e_2} \cdots p_h^{e_h}$$

(11) $$m = p_1^{n_1} p_2^{n_2} \cdots p_t^{n_t}$$

(12) $$a_1 = 2^3 3^1 5^3 7^4 \qquad a_2 = 2^1 3^3 7^3 (11)^2$$
$$d = 2^1 3^1 7^3 \qquad m = 2^3 3^3 5^3 7^4 (11)^2$$

(13) $$a = bq + r \qquad 0 \leq r < b$$

(14) $$544 = 4(119) + 68 \qquad 119 = 1(68) + 51$$
$$68 = 1(51) + 17 \qquad 51 = 3(17)$$

(15) $$\frac{a}{d} = \frac{ka}{kd}$$

(16) $$\frac{a}{d} = \frac{ka}{kd} = \frac{ka}{m}, \qquad \frac{b}{e} = \frac{lb}{le} = \frac{lb}{m}$$

(17) $$\frac{a}{d} \pm \frac{b}{e} = \frac{ka}{m} \pm \frac{lb}{m} = \frac{ka \pm lb}{m}$$

Rational numbers. From a less abstract point of view, the notion of division, or fraction, may also be considered to arise as follows: if the time duration of a given process is required to be known to an accuracy of better than 1 hour, the number of minutes may be specified; or, if the hour is to be retained as the fundamental unit, each minute may be represented by $\frac{1}{60}$, or by $^1/_{60}$.

In general, the fractional unit $1/d$ is defined by the property $d \times (1/d) = 1$. The number $a \times (1/d)$ is written a/d and is called a common fraction. It may be considered as the quotient of a divided by d. The number d is called the denominator (it determines the fractional unit or denomination), and a is called the numerator (it enumerates the number of fractional units that are taken). The numerator and denominator together are called the terms of the fraction. A positive fraction a/d is said to be proper if $a < d$; otherwise it is improper.

Common fraction

The numerator and denominator of a fraction are not unique, since for every positive integer k, the numerator and denominator of a fraction can each simultaneously be multiplied by the integer k without altering the fractional value (see 15). Thus, every fraction can be written as the quotient of two relatively prime integers. In this form it is said to be in lowest terms.

The integers and fractions constitute what are called the rational numbers. The five fundamental laws stated earlier with regard to the positive integers can be generalized to apply to all rational numbers.

Adding and subtracting fractions. From the definition of fraction it follows that the sum (or difference) of two fractions having the same denominator is another fraction with this denominator, the numerator of which is the sum (or difference) of the numerators of the given fractions. Two fractions having different denominators may be added or subtracted by first reducing them to fractions with the same denominator. Thus, to add a/d and b/e the LCM m of d and e, often called the least common denominator of the fractions, is determined. It follows that there exist numbers k and l such that $m = kd = le$, and both fractions can be written (see 16) with common denominator m, so that their sum or difference (see 17) is obtained by the simple operation of adding or subtracting the numerators only.

Multiplying and dividing fractions. The product of two fractions (see 18) is a fraction the numerator of which is the product of the numerators of the factors, and the denominator of which is the product of the denominators of the factors. The quotient of two fractions (see 19) is equal to the product of the dividend by the divisor inverted; that is, the divisor with its terms interchanged.

Theory of rationals. A method of introducing the positive rational numbers that is free from intuition (that is, with all logical steps included) was given by a German mathematician, Ernst Steinitz, in 1910. In considering the set of all number pairs $(a, d), (b, e), \cdots$ in which $a, b, d, e, \cdots$ are positive integers, the equals relation $(a, d) = (b, e)$ is defined to mean that $ae = bd$, and the two operations + and × are defined so that the sum of a pair (see 20) is a pair and the product of a pair (see 21) is a pair. It can be proved that, if these sums and products are properly specified, the fundamental laws (see 1–5) hold for these pairs, and that the pairs of the type $(a, 1)$ are abstractly identical with the positive integers a. Moreover, $d \times (a, d) = a$, so that the pair (a, d) is abstractly identical with the fraction a/d.

Irrational numbers. It was known to the Pythagoreans that, given a straight line segment a and a unit segment u, it is not always possible to find a fractional unit such that both a and u are multiples of it. Thus, the hypotenuse of an isosceles right triangle the sides of which are taken as the unit u must by the Pythagorean theorem have a

(18) $\frac{a}{d}\times\frac{b}{e}=\frac{ab}{de}$

(19) $\frac{a}{d}\div\frac{b}{e}=\frac{a}{d}\times\frac{e}{b}=\frac{ae}{db}$

(20) $(a, d)+(b, e)=(ae+bd, de)$

(21) $(a, d)\times(b, e)=(ab, de)$

(22) $a<\alpha<b \qquad a-b<\epsilon$

(23) $\sqrt[3]{(\sqrt{2}+\sqrt[5]{3})/\sqrt{5}}$

(24) $\sqrt[3]{\frac{125}{18}}=\sqrt[3]{\frac{125(12)}{18(12)}}=\sqrt[3]{\frac{5^3(12)}{6^3}}=\frac{5}{6}\sqrt[3]{12}$

(25) $\sqrt[n]{a}(\sqrt[n]{b})=\sqrt[n]{ab}$

(26) $\sqrt[n]{\sqrt[m]{a}}=\sqrt[mn]{a}$

(27) $(\sqrt[m]{a})^k=\sqrt[m]{a^k}$

(28) $\sqrt[mn]{a^{kn}}=\sqrt[m]{a^k}$

(29) $\sqrt[n]{\frac{a}{b}}=\frac{\sqrt[n]{a}}{\sqrt[n]{b}}$

(30) $a^0=1 \qquad a^{-n}=\frac{1}{a^n} \qquad a^{p/q}=\sqrt[q]{a^p} \qquad a\neq 0$

(31) $\sqrt[n]{a}(\sqrt[n]{b})=a^{1/n}b^{1/n}=(ab)^{1/n}=\sqrt[n]{ab}$

(32) $\sqrt[m]{a^k}=a^{k/m}=a^{kn/mn}=\sqrt[mn]{a^{kn}}$

length the square of which is 2. But there exists no rational number the square of which is 2.

Eudoxus' contribution

Eudoxus of Cnidus, a contemporary of Plato, established the technique necessary to extend numbers beyond the rationals. His contribution, one of the most important in the history of mathematics, was included in Euclid's *Elements,* Book V and elsewhere, and then lay dormant until the modern period of growth in mathematical analysis in Germany in the 19th century.

It is customary to assume on an intuitive basis that, corresponding to every line segment and every unit length, there exists a number (called a positive real number) that represents the length of the line segment. Not all such numbers are rational, but every one can be approximated arbitrarily closely by a rational number. That is, if α is real and ϵ is any positive rational number no matter how small, it is possible to find two positive rational numbers a and b within ϵ distance from each other such that (see 22) α is between them. In problems in mensuration, irrational numbers are usually replaced by suitable rational approximations.

A rigorous development of the irrational numbers is beyond the scope of arithmetic. They are most satisfactorily introduced by means of Dedekind cuts, as introduced by the German mathematician Richard Dedekind, or sequences of rationals, as introduced by Eudoxus and developed by the German mathematician Georg Cantor. These methods are discussed in the theory of functions of a real variable (see ANALYSIS: *Real analysis*).

The employment of irrational numbers greatly increases the scope and usefulness of arithmetic. For instance, if n is any whole number and a is any positive real number, there exists a unique positive real number $\sqrt[n]{a}$ called the nth root of a, whose nth power is a. The root symbol $\sqrt{\ }$ is a conventionalized r for radix, or root. The term evolution is sometimes applied to the process of finding a rational approximation to an nth root.

The symbol $\sqrt{\ }$

Surds. A number (see 23) that is obtainable from rational numbers by a finite number of additions, multiplications, divisions, and root extractions is called a surd. An irrational number of the form $\sqrt[n]{a}$ in which a is rational is called a pure surd of index n. For $n=2$, it is called a quadratic surd or square root and is written simply $\sqrt{a}$. A surd that is not pure is called mixed.

A pure surd of index n may be reduced to a rational multiple of the nth root of a positive integer no prime factor of which occurs to an exponent as high as n. Such a surd is in lowest terms. Thus consider $\sqrt[3]{125/18}$. Since $18=2(3^2)$, it becomes a perfect cube upon multiplication by $2^2(3)=12$. Then (see 24) the cube root of 125/18 can be expressed as a rational multiple of the cube root of 12.

Two surds that can be written as rational multiples of the same surd are called similar. Their sum is similar to each of them, its rational coefficient being the sum of the rational coefficients of the summands.

Properties of surds. The elementary properties of surds are readily provable (see 25–29) and guide one in calculating products of roots (see 25), roots of roots (see 26), powers of roots (see 27), roots of powers (see 28), and roots of fractions (see 29).

The theory of surds can be much simplified by the introduction of negative and fractional exponents, in which (see 30) the zeroth power of a number is unity, the negative power of a number is the reciprocal of the number, and a fractional power is expressed as a surd.

It can readily be shown that the fundamental laws of exponents, together with the properties of the rational fractions, are equivalent to the elementary properties (see 25–28) of surds. Exemplary verification (see 31, 32) is not difficult.

Complex numbers. Early mathematicians were led to consider purely formal situations involving square roots of negative numbers.

Thus, Heron of Alexandria (*c.* AD 100) obtained the quantity $\sqrt{-63}$, and Girolamo Cardan (1545) wrote $40=(5+\sqrt{-15})(5-\sqrt{-15})$. These numbers were considered to be quite meaningless, and the term imaginary was applied to them. They have since become indispensable in several branches of modern mathematics and have applications in mechanics and electricity.

In 1832 Gauss proposed the term complex for numbers of the form $a+bi$ where a and b are real and i is defined as $\sqrt{-1}$. The modern development of complex numbers began with the discovery of a geometric interpretation for them. This was indistinctly set forth by John Wallis (1685) and in completely satisfactory form by Caspar Wessel (1799).

Wessel's work received no attention, and the geometric interpretation was rediscovered by Jean Robert Argand (1806) and again by Gauss (1831). It is frequently called the Argand diagram (Figure 2).

Argand diagram

Just as the real numbers represent points on a line, the complex numbers can be put into correspondence with the points on a plane. The multiples of i are called purely imaginary numbers and are plotted as points on the imaginary axis perpendicular to the axis of reals at the O-point. Then the number $a+bi$ corresponds to the point P where

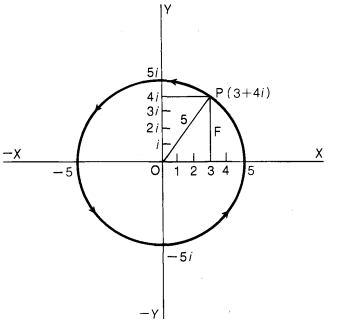


Figure 2: The complex number.

OP is the diagonal of the rectangle whose sides are a and bi. If $a + bi$ is multiplied by -1, its corresponding line or vector *OP* is rotated through 180°. Similarly, if $a + bi$ is multiplied by i, its vector is rotated through 90°. Thus, two successive multiplications by i produce the same effect as one multiplication by -1, so that in this sense $i^2 = -1$.

The first satisfactory introduction of the complex numbers was given by Sir William Rowan Hamilton in 1835, although Gauss afterward stated that the same idea had occurred to him in 1831. (C.C.MacD.)

NUMBER SYSTEMS AND NOTATION

Just as in language, where infinitely many words can be composed of but a small variety of different characters, so also in arithmetic are numbers, infinitely many of them, composed of but a small variety of numerals. The devising of a scheme whereby an infinitude of things can be represented by means of but a small number of symbols must be ranked among the most important achievements of the human intellect, for without it an advanced development of either language or arithmetic is unimaginable. The key ideas are two: the positional principle (to be described below) and the symbol zero. "The idea is so simple," wrote the 18th–19th-century mathematician Pierre-Simon, marquis de Laplace, speaking of the positional principle, "that this very simplicity is the reason for our not being sufficiently aware how much admiration it deserves." Similarly, the invention, probably by the Hindus, of the digit zero has been described as one of the greatest importance in the history of mathematics. Hindu literature gives evidence that the zero may have been known before the birth of Christ, but no inscription has been found with such a symbol before the 9th century.

Argand diagram

Applied to the construction of numbers, the positional principle operates thus: the sequence of digits $\cdots srqp$ is defined to signify a number the magnitude of which is a sum of products involving powers of a number a (see 33), in which a is called the base or radix; that is, the position of each of the coefficients $p, q, r, s, \cdots$ is associated, in reverse order to the representation $srqp$, with the zeroth, first, second, third, $\cdots$ powers of the base a. The number of distinct numerals required in this notation is readily seen to be a.

$$(33) \qquad p \times a^0 + q \times a^1 + r \times a^2 + s \times a^3 + \cdots$$

$$(34) \qquad p \times 1 + q \times 10 + r \times 100 + s \times 1000 + \cdots$$

Each numeral, when part of a number, can therefore be said to have two values: an intrinsic value, which is simply that signified by the isolated symbol itself; and a local value, which is that possessed by virtue of its position, or location, within the sequence of digits used to express the given number.

Systems are called binary, ternary, quaternary, quinary, senary, septenary, octenary (or octal), nonary, denary (or decimal), undenary, duodenary, hexadecimal, vigesimal, and sexagesimal, corresponding to values of $a = 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 16, 20$, and 60, respectively. The pair system, in which the counting goes "one, two, two and one, two twos, two and two and one," etc., is found among the ethnologically oldest tribes of Australia, in many Papuan languages of Torres Strait and the adjacent coast of New Guinea, among some African Pygmies, and in various South American tribes. Other tribes of Tierra del Fuego and the South American continent use number systems with bases three and four. The quinary scale, or number system with base five, is very old but in pure form seems to be used at present only by speakers of Saraveca, a South American Arawakan language; elsewhere it is combined with the decimal or the vigesimal system, where the base is 20. Similarly, the pure base six scale seems to occur only sparsely in Northwest Africa and is otherwise combined with the duodecimal, or base 12, system.

In the course of history the decimal system finally overshadowed all others, and it is now found in all nations of high culture on the entire globe, except those of Mexico and Central America, where the number 20 was used in astronomy and thus became firmly entrenched. Nevertheless, there are still many vestiges of other systems in nations of high culture, chiefly in commercial and domestic units, where change always meets the resistance of tradition. Thus, 12 occurs as the number of inches in a foot, pence in a shilling, months in a year, ounces in a pound (troy or apothecaries'), and twice 12 hours in a day; and both dozen and gross measure by 12s. In English the base 20 occurs chiefly in the score ("Four score and seven years ago . . . "); in French it survives in the word *quatre-vingts* ("four twenties") for 80; other traces are found in pre-English Celtic, Gaelic, Danish, and Welsh.

The number zero

The Babylonians developed (2000–3000 BC) a positional system with base 60—a sexagesimal system. With such a large base it would be awkward to have unrelated names for the digits 0, 1, $\cdots$, 59, so a simple grouping system to base 10 was used for these numbers. For example, the number 258,458 is written in Figure 3 as it would have appeared in Babylon. The base 60 still occurs in measurement of time and angles.

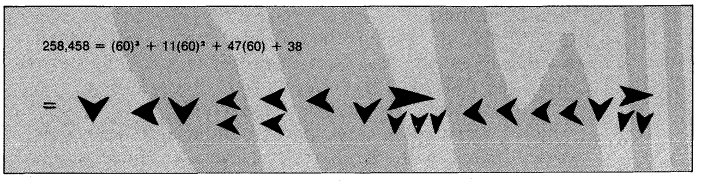


Figure 3: The number 258,458 expressed in the sexagesimal system of the Babylonians and in cuneiform.

In addition to being somewhat cumbersome because of the large base chosen, the Babylonian system suffered until very late from the lack of a zero symbol; the resulting ambiguities may well have bothered the Babylonians as much as later translators.

In the course of early Spanish expeditions into Yucatán it was discovered that the Mayans, at an early but still undated time, had a well-developed positional system, complete with zero. It seems to have been used primarily for the calendar rather than for commercial or other computation; this is reflected in the fact that although the base is 20, the third digit from the end does not signify multiples of 20^2 but of 18×20, thus giving their year a simple number of days. The digits 0, 1, $\cdots$, 19 are, as in the Babylonian, formed by a simple grouping system, in this case to base 5; the groups were written vertically, as in Figure 4.

The earliest numerals of which there is definite record were simple straight marks for the small numbers, with some special form for 10. These symbols appear in Egypt

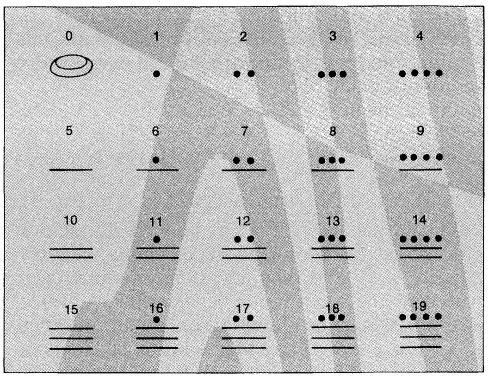


Figure 4: The Mayan vigesimal number system (see text).

as early as *c.* 3400 BC and in Mesopotamia as early as *c.* 3000 BC. These dates long precede the first known inscriptions containing numerals in India (*c.* 300 BC), in China (3rd century BC), and in Crete (*c.* 1200 BC). Some ancient symbols for 1 and 10 are given in Figure 5. This special

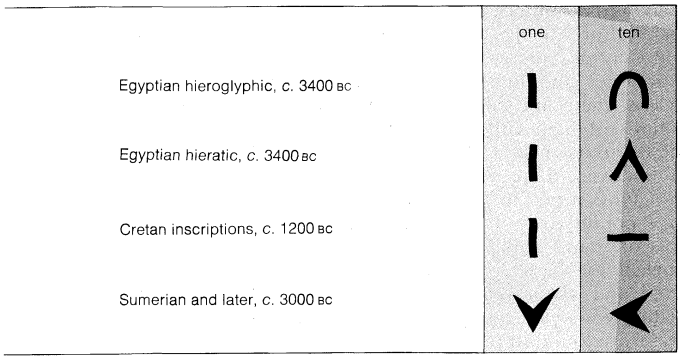


Figure 5: Some ancient symbols for 1 and 10.

position occupied by 10 stems from the number of human fingers, of course, and is still evident in modern usage not only in the logical structure of the decimal system but in the English names for the numbers. Thus, eleven comes from Old English *endleofan*, literally meaning "[ten and] one left [over]," and twelve from *twelf*, meaning "two left"; the endings -teen and -ty both refer to 10, of course, and hundred comes originally from a pre-Greek term meaning "ten times [ten]."

In the decimal system the above number $\cdots srqp$ has a value expressed as a sum of products involving powers of 10 (see 34). The ten numerals 1, 2, $\cdots$ 8, 9, 0 are the only distinct symbols needed here. (These symbols, which originated in large part in India and were widely disseminated by the Arabs, are in several instances stylized forms of the initial characters of the words employed to denote the appropriate numbers. They are called Hindu-Arabic numerals.) This notation is capable of being expanded to include the representation of fractional parts of numbers by means of placing a point (decimal point when $a = 10$) immediately to the right of the coefficient associated with a^0. Thus, $0.uvw \cdots$ represents the sum of products, each term of which includes a negative power of a (see 35).

With the advent of electronic high-speed computers, the choice $a = 2$ and, hence, binary arithmetic have become of great practical importance, though, as has been mentioned above, the binary (or pair) system has been used in tribal, nontechnical cultures. The fundamental reason for the new importance is the basic mode of operation of computing machines, in which an electric current pulse, or the direction of magnetization of an element, has one of two values. The numerals used in the binary system are 1 and 0. As an example, the magnitude of the number $N_2 = 100111$ to the base 2 (indicated by the subscript on N) is $1\times2^0+1\times2^1+1\times2^2+0\times2^3+0\times2^4+1\times2^5$, which in decimal notation becomes $N_{10}=1+2+4+0+0+32=39$. It is apparent that the binary system's use of only two distinct symbols is partly offset by the requirement of many more digits (6 in N_2, as compared with only 2 in N_{10} in the above example). Octal and hexadecimal modes also find frequent application in electronic computing devices.

Binary numbers

In octal notation the above $N_{10} = 39$ is readily seen to become $N_8 = 47$ because $7\times8^0+4\times8^1=7\times1+32=39$ (in decimal notation). (D.E.S./W.J.LeV.)

ARITHMETIC CALCULATION WITH DECIMALS

Addition and subtraction. Numbers in decimal notation can easily be added by adding the coefficients of corresponding powers of 10 and then adjusting, by the process known as carrying, the coefficients that exceed 9. A typical example (see 36) would be the addition of the numbers 47.65, 5.473, and 649.8. One first adds coefficients of equal powers of 10; *e.g.*, one adds 5 and 7 to obtain 12 as a coefficient of 10^{-2}. After adjustment of the coefficients (see 37), this sum becomes 702.923.

The process of addition is commonly carried out as follows (see 38): Beginning at the right, the sum of the coefficients of 10^{-3}, which is 3, is written below the answer line; the sum of the coefficients of 10^{-2} is 12, the digit 2 being placed to the left of the 3 in the sum; however, the 10 becomes a 1 in the preceding position and may be placed at the top of the column of coefficients of 10^{-1}, etc. In practice the carrying is usually performed mentally.

Subtraction is performed by reversing the above procedure. Thus, to subtract 170.8 from 563.142 (see 39), for example, the smaller number is placed under the larger, with the respective decimal points aligned vertically. Again, starting from the right, since $0.8 = 0.800$, 0 subtracted from 2 leaves 2 (which is written below the line), and 0 from 4 leaves 4; but since 8 exceeds 1, a unit is borrowed from the unit's place (which changes the 3 to 2), and thus the 1 in the tenth's place becomes 11; then 8 subtracted from 11 leaves 3, 0 from 2 leaves 2, and so forth.

Another way of performing subtractions consists in finding the number, digit by digit, that, when added to the second, yields the first as sum.

Multiplication. Multiplication of decimal numbers is based upon the distributive law. Thus, the multiplication (see 40) of 25.78 by 7 is accomplished as a sum of four products, each involving 7 as a factor.

By a process known as short multiplication, this may be carried out as follows (see 41): Since $7\times8=56$, the 6 is written below the answer line, starting at the right, and the 5 is placed above the 7 in the multiplicand; then $7\times7=49$, to which is added the 5 that was carried,

(35) $N=0+u\times a^{-1}+v\times a^{-2}+w\times a^{-3}+\cdots$

(36)
$$\begin{array}{r} 47.65 \\ 5.473 \\ \underline{649.8} \\ \end{array}$$

Sum calculated as follows:

$$\begin{array}{r} 4(10)+\ 7+\ 6(10^{-1})+\ 5(10^{-2}) \\ 5+\ 4(10^{-1})+\ 7(10^{-2})+3(10^{-3}) \\ 6(10^2)+4(10)+\ 9+\ 8(10^{-1}) \\ \hline 6(10^2)+8(10)+21+18(10^{-1})+12(10^{-2})+3(10^{-3}) \end{array}$$

(37) $7(10^2)+0(10)+2+9(10^{-1})+$
$+2(10^{-2})+3(10^{-3})=702.923$

(38)
$$\begin{array}{r} \underline{121\ 1} \\ 47.65 \\ 5.473 \\ \underline{649.8} \\ 702.923 \end{array}$$

(39)
$$\begin{array}{r} 563.142 \\ \underline{-\ 170.8} \\ 392.342 \end{array}$$

(40) $25.78\times7=20\times7+5\times7+.7\times7+.08\times7$
$=140+35+4.9+.56=180.46$

(41)
$$\begin{array}{r} \underline{45\ 5} \\ 25.78 \\ \underline{\quad\ 7} \\ 180.46 \end{array}$$

(42)
$$\begin{array}{r} 25.78 \\ \underline{\quad 45.7} \\ 18\ 046 \\ 128\ 90 \\ \underline{1\ 031\ 2} \\ 1\ 178.146 \end{array}$$

making 54; the 4 is written in the answer, the 5 being carried as before. The numbers written above the top line are usually carried mentally.

If the multiplier has more than one digit, long multiplication is used (see 42). This process also is based upon the distributive law. The first partial product is 18.046, which was obtained by multiplying 25.78 by .7 by short multiplication. The second partial product is $25.78 \times 5 = 128.90$; the third is $25.78 \times 40 = 1{,}031.2$. The product is the sum of these partial products. It is customary to ignore the decimal point in these partial products but to indent each partial product one more place than the partial product above it. In the product as many digits are pointed off from the right as the sum of the number of digits to the right of the decimal point in the multiplicand and the number of digits to the right of the decimal point in the multiplier.

(43)
$$\begin{array}{r} 4\,615 \\ 7\overline{)53\,149} \\ 7\,592 \end{array}$$

(44) $$\frac{53{,}149}{7} = 7{,}592\frac{5}{7}$$

(45)
$$\begin{array}{l} 621)\ 83\,742\ (134 \\ \quad\underline{62\,1\ \ } \\ \quad 21\,64 \\ \quad\underline{18\,63\ } \\ \quad\ 3\,012 \\ \quad\underline{2\,484} \\ \quad\ \ 528 \end{array}$$

(46) $83{,}742 = 621 \times 134 + 528$

(47)
$$\begin{array}{l} 621)\ 83\,742\ (134 \\ \quad 21\,64 \\ \quad\ 3\,012 \\ \quad\ \ 528 \end{array}$$

(48) $$\frac{83\ 742}{621} = 134\frac{528}{621}$$

(49)
$$\begin{array}{l} 621)\ 83\,742\ (134.8502 \\ \quad 21\,64 \\ \quad\ 3\,012 \\ \quad\ \ 528\,0 \\ \quad\ \ \ 31\,20 \\ \quad\ \ \ \ 1\,500 \\ \quad\ \ \ \ \ 258 \end{array}$$

(50) $$\frac{83742.0000}{621} = 134.8502$$

(51) $$\frac{83742}{621} = \frac{837.42}{6.21} = 134.8502$$

(52)
$$\begin{array}{l} \ \ 15\ 78.42\ (39.72 \\ \quad\ \ 9 \\ \quad\overline{\ \ } \\ 69\ |\ 6\ 78 \\ \quad\ |\ \underline{6\ 21} \\ 787\ |\ 57\,42 \\ \quad\ |\ \underline{55\,09} \\ 7942\ |\ 2\,33\,00 \\ \quad\ |\ \underline{1\,58\,84} \\ \quad\ \ \ 74\,16 \end{array}$$

Division. If a and b are two whole numbers, $a > b$, there exist two whole numbers q and r such that (see 13) a is written as a sum of two terms, one involving q and one involving r. The process of finding these numbers can be carried out by the processes called long and short division.

If $b < 10$, say $b = 7$, short division may be used (see 43). Then, if $a = 53{,}149$, starting at the left, since $5 < 7$ and considering the first two digits, 53, the largest multiple of 7 that is less than or equal to 53 is $49 = 7 \times 7$. The 7 is written below the 3, and the remainder, $53 - 49 = 4$, above the 3. This remainder with the next digit in the dividend is 41; the largest multiple of $7 \leq 41$ is $35 = 7 \times 5$. The 5 is written below the 1, and the remainder, $41 - 35 = 6$, above the 1. The last remainder is 5. Usually the remainders are carried mentally.

It is notable that this process involves guessing the largest multiple of the divisor that is less than a certain number, but only a few guesses are ever required and no practical difficulty is encountered.

Because $53{,}149 = 7{,}592 \times 7 + 5$, one may write (see 44) the ratio of 53,149 and 7 so that 7,592 is the partial quotient and 5 is the remainder.

If the divisor exceeds 10, long division is preferable (see 45). To divide 83,742 by 621, the 1 is written as the first digit of the quotient and 621×1 is subtracted from 837 since the largest multiple of $621 \leq 837$ is 621×1. Actually what is done is to subtract 621×100 from 83,742, leaving a remainder of 21,642. Since only the 2,164 will be used in the next step, the final 2 of the dividend need not be brought down yet. The largest multiple of $621 \leq 2{,}164$ is $621 \times 3 = 1{,}863$, and the remainder is 301. Now the final 2 must be brought down to make 3,012. The largest multiple of $621 \leq 3{,}012$ is $621 \times 4 = 2{,}484$, and the remainder is 528. Binary numbers

Thus, the dividend (see 46) is the sum of a product and a remainder.

The preceding method may be considerably abbreviated by mentally subtracting the partial product as it is formed (see 47). Again, the resultant quotient may be expressed as the sum of an integer and a proper fraction (see 48). If it is desired to express the result as a decimal fraction, carried to a given number of decimal places, the above process is merely continued (see 49). Any two decimal numbers may be divided in this manner. The number of digits to the right of the decimal point in the quotient is equal to the number of such digits in the dividend, diminished by the number of such digits in the divisor, with care being taken to add to the dividend all the 0's that are brought down. Thus, in the present example (49) the proper fraction (see 50) is replaced by a decimal fraction. An alternative and exceptionally clear and simple method for determining the position of the decimal point in the quotient is (see 51) to divide both numerator and denominator by an appropriate power of 10; *i.e.*, expressing the divisor as a number lying between 0 and 10, which renders it obvious that the quotient lies between 100 and 1,000.

Divisibility rules. In both theoretical and practical arithmetic, it is often important to factor a natural number; *i.e.*, to decompose it into numbers that, when multiplied together, will yield the given number as product.

The following tests for divisibility are therefore given. A composite number is divisible

- by 2 if it is even (*i.e.*, if it ends in 0, 2, 4, 6, or 8);
- by 3 if the sum of its digits is divisible by 3;
- by 4 if the number formed by its last two digits is divisible by 4;
- by 5 if it ends in 0 or 5;
- by 6 if it is even and the sum of its digits is divisible by 3;
- by 8 if the number formed by its last 3 digits is divisible by 8
- by 9 if the sum of its digits is divisible by 9;
- by 10 if it ends with 0;
- by 11 if the difference between the sum of its digits in the odd places and that of the digits in the even places is either 0 or divisible by 11.

Similar rules are easily devised with respect to other divisors by means of appropriate combinations of the rules given above. For example, if a number is to be divisible by 132, say, then it must satisfy the test for divisibility by factors of 132—that is, by 3, by 4, and by 11, respectively; *i.e.*, $3 \times 4 \times 11 = 132$.

(53) $(a+b)^3 - a^3 = (3a^2 + 3ab + b^2)b$

(54)

279 463.000 (65.38—
216
10 800 | 63 463
900
25
11 725 | 58 625
1 267 500 | 4 838 000
5 850
9
1 273 359 | 3 820 077
127 922 700 | 1 017 923 000
156 720
64
128 079 484 | 1 024 635 872

Evolution. An algorithm for the determination of the square root of a decimal number, such as 1,578.42, is carried out as follows (see 52): Starting at the decimal point, the number is separated into periods of two digits each. The leftmost period is 15, and the largest square ≤ 15 is $9 = 3^2$. The 3 is written at the right and the remainder $15 - 9 = 6$ is brought down. The next period 78 is brought down beside the 6, giving 678. The 3 is doubled and the result, 6, written at the left. By trial it is found that 9 is the largest digit, such that $69 \times 9 = 621 \leq 678$. The 9 is written after the 3 in the answer, and the difference, 57, is brought down, followed by the next period, 42. The partial answer 39 is doubled and the result, 78, written at the left under the 69. By trial it is found that $787 \times 7 = 5{,}509 \leq 5{,}742$ while $788 \times 8 > 5{,}742$. The next digit in the answer is therefore 7. The process is continued until the desired degree of accuracy is attained. In the present example (see 52) the last remainder, 7,416, exceeds half of the last trial divisor, 7,942, so that 3 is a better approximation than 2 for the last digit. In fact, $39.73^2 = 1{,}578.4729$, but $39.72^2 = 1{,}577.6784$.

The above process is based upon the relation $(a+b)^2 = a^2 + (2a+b)b$. In each step the part of the square root already obtained is a; the part remaining to be found is b. In the example (see 52), $(a+b)^2 = 1{,}578.42$, $a = 30$. Then $2ab + b^2 = 678.42$. In order to determine the largest integer n in b, one must find the largest integer n (namely 9), such that $(2a+n)n \leq 678.42$. The trial divisor is $2a + n = 69$. In the next step $a = 39$, $2ab + b^2 = 57.42$, etc.

Cube and higher roots. The cube root of a number may be calculated by a similar algorithm based upon the relation (see 53) that expresses the difference of two cubes as a product involving a quadratic as a factor. Thus, it is possible to find the approximate cube root of 279,463 by proceeding as in the example (see 54). This last remainder is too large but much closer than the remainder resulting from using 7 as the last digit.

A fourth root is obtained as the square root of the square root. Fifth and higher roots can be obtained by an algorithm similar to those just given, based upon the expansion of $(a+b)^n$ by the binomial theorem. The method is cumbersome and seldom used, for the method of logarithms is easy and rapid. The fifth root of a can also be quickly found by solving the equation $x^5 - a = 0$ by Newton's or Horner's method. (C.C.MacD.)

LOGARITHMS

Basic principles. Logarithms were invented in the early 17th century to speed up calculations, and they were basic in numerical work for more than 300 years. The perfection of the desk calculating machine in the late 19th century and the electronic computer in the 20th has made them obsolete for large-scale computation.

The operation and nature of logarithms can be seen from a table characteristic of logarithms (see 55) that identifies with the number $\frac{1}{2}$ the logarithm -1, with the number 1 the logarithm 0, with the number 2 the logarithm 1, with the number 4 the logarithm 2, and so forth. Here, the constant 2 raised to any logarithmic power gives the corresponding number. A table of this type can be used for multiplication (and division)—the logarithms of the numbers to be multiplied are selected from the table and added, and the number corresponding to the answer is the result of the multiplication. For example, two numbers are taken from the first row and multiplied together; say, $2 \times 8 = 16$. The corresponding values in the second row are 1, 3, and 4; hence, when 2 is taken to the 4th power, 16 is yielded.

The reason for this is that numbers in the first row are the number 2 to the power of the corresponding number in the second. Thus $\frac{1}{8} = 2^{-3}$, $1 = 2^0$, $8 = 2^3$. In the identity $32 = 2^5$ the number 2 is called the base. The exponent 5 is the logarithm of 32 to the base 2 and is written $5 = \log_2 32$. More generally, the two equations have the same meaning; the first serves to define the second. By this definition y is the logarithm of x to the base b if and only

(22)

$\frac{1}{8}$	$\frac{1}{4}$	$\frac{1}{2}$	1	2	4	8	16	32	64
−3	−2	−1	0	1	2	3	4	5	6

if $x = b^y$. The number x is also called the antilogarithm of y to the base b.

It is evidently easy to find logarithms of numbers that are simple powers of the base, and there are quite efficient methods for calculating logarithms of all numbers to as many decimal places as desired.

Multiplication of any numbers m and n can be accomplished by adding their logarithms; *i.e.*, the logarithm of the product is the sum of the logarithms (see 56). Division of numbers can also be accomplished by subtracting the logarithms: the logarithm of the quotient is the difference of the logarithms (see 57). This is not all; powers and roots can also be found with logarithms. For example, the cube of 4 is 64 (*i.e.*, $4^3 = 64$), and from the table the logarithms

(23) $\log_b mn = \log_b m + \log_b n$

(24) $\log_b (m/n) = \log_b m - \log_b n$

(25) $\log_b (n^p) = p \log_b n$

(26) $\log_b \sqrt[q]{n} = \frac{1}{q} \log_b n$

(27) $\log_b \sqrt{n} = \frac{1}{2} \log_b n$

(28) $\begin{cases} b^x \cdot b^y = b^{x+y} \\ b^x / b^y = b^{x-y} \\ (b^x)^p = b^{px} \end{cases}$

(29) $\begin{cases} m = b^x, & \log_b m = x \\ n = b^y, & \log_b n = y \end{cases}$

(30) $mn = b^x b^y = b^{x+y}$

(31) $\sqrt[q]{b^x} = b^{x/q}$

$$(65)\quad \begin{cases} 2 \cdot 41 \times 10 \\ \log_{10} 24 \cdot 1 = \log_{10}(2 \cdot 41 \times 10) \\ \qquad\qquad = \log_{10} 2 \cdot 41 + \log_{10} 10 \end{cases}$$

$$(66)\quad y = \log_b x$$

$$(67)\quad e = 1 + \frac{1}{1!} + \frac{1}{2!} + \frac{1}{3!} + \cdots + \frac{1}{n!} + \cdots$$

$$(68)\quad \frac{d}{dx} \log_b x = \frac{1}{x} \log_b e$$

$$(69)\quad \frac{d}{dx} \log_e x = \frac{1}{x}$$

$$(70)\quad \log_e(1 + x) = x - \frac{x^2}{2} + \frac{x^3}{3} + \cdots$$

$$(71)\quad \log_b x = \log_a x \log_b a$$

$$(72)\quad \log_{10} x = \log_e x \log_{10} e$$

of 4 and 64 are 2 and 6. Because $6 = 3 \cdot 2$, the logarithm of 4^3 can be found by multiplying the logarithm of 4 by 3. Study of the table will verify that the logarithm of a power can be found by multiplying the logarithm of the number by the index p of the power (see 58). Because logarithms transform multiplication into addition, division into subtraction, and the taking of powers into multiplication, it might be guessed that they would transform the taking of square roots into division. This is the case, for example, in computing the square root of 16, its logarithm (which is 4) is divided by 2. The result is 2, which is the logarithm of 4, as expected. In general, the logarithm of a root is the logarithm of the number divided by the index q of the root (see 59). Because $n^{1/2} = \sqrt[2]{n}$, square roots are given in the logarithmic form as one-half the logarithm of the number for which the square root is taken to be calculated (see 60). Logarithms work this way because they are exponents; and exponents are added for multiplying, subtracted for dividing, multiplied to take a power, and divided to take a root. These ideas may be expressed as laws of exponents (see 61).

The exponents in these equations can be thought of as logarithms. For example, a variable may be expressed as a power of a base (see 62), and another variable may be expressed as a power of the same base. Their product (see 63) is similarly expressible in terms of a sum. Therefore, $\log_b mn = \log_b m + \log_b n$, the relationship for the addition of logarithms (see 64).

Use of logarithms for multiplication

Common logarithms. The most convenient tables for numerical calculations are those in which the logarithms are to the base 10. These are called common logarithms. They have the advantage that a table of logarithms of numbers between one and 10 can be used to find the logarithms of all other numbers. For example, from tables, $\log_{10} 2.41$ is 0.38202 (this means that $10^{0.38202}$ is 2.41; fractional powers of 10 can be found and tabulated; see below). The logarithm of 24.1 can then be found because 24.1 is expressible as a number between two and three times 10 (see 65). Because $\log_{10} 10 = 1$, the logarithm of 24.1 is 1.38202. Thus each common logarithm has two parts, an integer and a decimal less than one. The integral part is called the characteristic and is determined by the position of the decimal point in the number. Thus the log of 241.0 is 2.38202 and its characteristic is 2; for 2.41 the characteristic is 0, for 0.241 it is -1, and for 0.00241 it is -3. The decimal part of a common logarithm is called the mantissa and is found from a table of logarithms by disregarding the position of the decimal point in the number.

When the characteristic is negative it cannot be written with the minus sign in front of the mantissa without causing confusion. For example, the log of 0.00241 has a mantissa of 0.38202 and characteristic of -3. It cannot be written -3.38202 because it is, in fact, $-3 + 0.38202$ (not $-[3 + 0.38202]$). It is customary to write it in the form $\bar{3}.8202$.

Natural logarithms. The logarithmic function, in algebra, is defined by the relationship that identifies y with the logarithm of x (see 66). The problem of finding its derivative involves the problem of finding the limit of an expression of the form $(1 + t)^{1/t}$, as t approaches infinity. This limit is an irrational number e given by the infinite series that sums the reciprocals of factorials (see 67), in which a factorial is a product of consecutive integers beginning with 1 and ending with the integer whose factorial is being taken. The number e has a value approximated by 2.71828, given here to five decimal places. This result leads to the equation that expresses the derivative of the logarithm to an arbitrary base (see 68). If logarithms to the base e are used, the equation simplifies to the form basic to calculus, expressing the derivative of the logarithm of x as the reciprocal of x (see 69).

Logarithms to the base e are sometimes called natural logarithms. They are less useful for computation than common logarithms but are used in calculation, frequently by means of the series for the logarithm of $1 + x$ (see 70): natural logarithms can be computed to any required degree of accuracy by evaluating a sufficient number of terms of the series. Common logarithms can then be obtained by a change of base using the formula that converts from one logarithmic base to another through the operation of multiplication of logarithms (see 71); this follows from the law of exponents. Thus a common logarithm of x is found by the equation that identifies it with the natural logarithm times a constant (see 72). (Ed.)

History of logarithms. The invention of logarithms was foreshadowed by the comparison of arithmetic and geometric series. In the simple table used above, the top line is a geometric series and the bottom line is an arithmetic series. The first table based on this concept was published in 1620 in Prague by Joost Bürgi. The comparison between the two series was not based on any explicit use of the exponential notation; this was a later development.

John Napier, the Scottish mathematician, published his discovery of logarithms in 1614. His purpose was to assist in the multiplication of quantities that were then called sines. The whole sine was the value of the side of a right angled triangle with a large hypotenuse, say 10^7 units long. His definition was given in terms of relative rates.

> The logarithme, therefore, of any sine is a number very neerely expressing the line which increased equally in the meene time whiles the line of the whole sine decreased proportionally into that sine, both motions being equal timed and the beginning equally shift.

In modern terminology, L is the logarithm and X the

$$(73)\quad \frac{dh}{dt} = a, \qquad \frac{dX}{dt} = -bX$$

$$(74)\quad \begin{cases} L = 0 \quad t = 0 \\ i.e.,\ L(r) = 0 \quad \dfrac{dX}{dt} = \dfrac{-aX}{r} \end{cases}$$

$$(75)\quad \begin{cases} L(x) = r(\log_e r - \log_e x) \\ L(x) = r \log_e \dfrac{r}{x} \end{cases}$$

$$(76)\quad L\left[\frac{XY}{Z}\right] = L(X) + L(Y) - L(Z)$$

$$(77)\quad \begin{cases} \log\left[\dfrac{\sin x}{x}\right] \\ \log\left[\dfrac{\tan x}{x}\right] = \log\left[\dfrac{\sin x}{x}\right] - \log \cos x \end{cases}$$

sine. (In modern notation X would be $r \sin \varphi$.) Thus, with modern techniques derivatives can be used (see 73). At $t = 0$, $X = r$ ($\varphi = 90°$), and, since the motion is "equally swift" at the beginning, $a = br$. Furthermore, the function L is known to have certain values at specific points (see 74). Thus in terms of present logarithms to the base e the function L can be expressed in terms of natural logarithms (see 75). Napier's value of r was 10^7. This expression does not have the expected property for $L(XY)$, but the relationship involving products and division (see 76) does apply.

he
ıodern
laperian
ɔgarithm

In cooperation with the English mathematician Henry Briggs, Napier did adjust his logarithms into the form in which it is usually found. For the modern Naperian logarithm (*i.e.*, the logarithm to the base e) the comparison would be between points moving on two straight lines, marked in units of length, the L point moving uniformly from minus infinity to plus infinity, the X point moving on a half line from zero to infinity at a speed proportional to the distance from zero. Furthermore, L is zero when X is one and their speed is equal at this point. The essence of Napier's discovery is that this constitutes a generalization of the relation between the arithmetic and geometric series; *i.e.*, multiplication and raising to a power of the values of the X point correspond to addition and multiplication of the values of the L point. In modern terminology, $\frac{dL}{dt} = k_0$, $\frac{dX}{dt} = k_0 X$ correspond to $\frac{dX}{dL} = X$, $X = e^L$, $L = \log_e(X)$.

From the point of view of the user it is better to limit the L and X motion by the requirement that $L = 1$ at $X = 10$ in addition to the condition that $X = 1$ at $L = 0$. This produces the common logarithms to the base 10.

These are sometimes known as Briggsian logarithms: the natural logarithms are known as Naperian logarithms.

The calculation of logarithms. The treatment above differs from that of Napier in that the word "exactly" is applicable rather than "very neerely." This obscures the ingenious procedure used for calculating the early logarithms, in which powers of numbers such as 1.00001 were used so that multiplication was minimized and replaced by addition. Thus, $X = (1.00001)^n$, $L = n/10^5$ corresponds approximately to $L = \log_e X$, or the natural logarithm. To obtain the Briggs or base 10 table, the calculation would be continued until X exceeded 10 and then the L scale adjusted so that at $X = 10$, $L = 1$.

In addition to the discrete series procedure, Napier and Briggs suggested the calculation of logarithms by extracting roots of 10; *i.e.*, $\log \sqrt{10} = 0.5$, $\log 10^{1/4} = 0.25$. This permits the n computation of the previous paragraph to be shortened, for the Briggs logarithm can be adjusted for by taking $L = 0.25$ for $X = 10^{1/4}$. Power series were not used in the initial construction of the tables. The power series for $\log(1 + x)$ and e^x were only available in the 18th century and rigorously established in the early 19th century.

Use of
ɔower
ɜeries

Logarithm tables. Napier died in 1617. Briggs published a table of logarithms to 14 places of numbers from 1 to 20,000 and from 90,000 to 100,000 in 1624. Adriaan Vlacq published a 10-place table for values from 1 to 100,000 in 1628, adding the 70,000 values. Both Briggs and Vlacq engaged in setting up log trigonometric tables. Such early tables were either to $^1/_{100}$ of a degree or to a minute of arc. In the 18th century tables were published for 10-second intervals, which were convenient for seven-place tables. In general, finer intervals are required for logarithmic functions in which the logarithm is taken of smaller numbers; for example, in the calculation of the functions $\log \sin x$ and $\log \tan x$. The related functions modified by division by x in the argument of the logarithm (see 77) are easily calculated by series for small values of x.

The availability of logarithms greatly influenced the form of plane and spherical trigonometry. Convenient formulas are ones in which the operations that depend on logarithms are done all at once. The recourse to the tables then consists of only two steps. One is obtaining logarithms, the other obtaining antilogs. The procedures of trigonometry were recast to produce such formulas. (F.J.M.)

BIBLIOGRAPHY. MUNRO LEAF, *Arithmetic Can Be Fun* (1949); and ISAAC ASIMOV, *Realm of Numbers* (1959), are introductory presentations at an elementary level. HAROLD D. LARSEN, *Arithmetic for Colleges* (1958); SIDNEY G. HACKER, WILFRED E. BARNES, and CALVIN T. LONG, *Fundamental Concepts of Arithmetic* (1963); and CARL B. ALLENDOERFER, *Mathematics for Parents* (1965), are written from the point of view of education. G.H. HARDY and E.M. WRIGHT, *An Introduction to the Theory of Numbers,* 4th ed. (1960); HAROLD DAVENPORT, *The Higher Arithmetic* (1952); and WILLIAM J. LEVEQUE, *Elementary Theory of Numbers* (1962), discuss fundamental concepts associated with the theory of numbers. Works emphasizing the history of arithmetic as a part of number theory include: OYSTEIN ORE, *Number Theory and Its History* (1948); LEONARD E. DICKSON, *History of the Theory of Numbers,* 3 vol. (1919–23, reprinted 1952); TOBIAS DANTZIG, *Number: The Language of Science,* 4th ed. rev. (1959); and DAVID E. SMITH and AUGUSTUS DE MORGAN, *Rara Arithmetica,* 4th ed. (1970). The following are classic and advanced studies: CARL FARBER, *Arithmetik* (1911); CARL F. GAUSS, *Disquisitiones Arithmeticae,* 2nd ed. (1870; Eng. trans. 1966); GOTTLOB FREGE, *Die Grundlagen der Arithmetik* (1884; Eng. trans., *Foundations of Arithmetic,* 1968); FELIX KLEIN, *Elementarmathematik vom höheren Standpunkte,* 3rd ed., vol. 1 (1924; Eng. trans., *Elementary Mathematics from an Advanced Standpoint,* vol. 1, *Arithmetic, Algebra, Analysis,* 1932, reprinted 1968); JEAN-PIERRE SERRE, *Cours d'Arithmetique* (1970). DONALD GREENSPAN, *Arithmetic Applied Mathematics* (1980), is a demonstration of sophisticated arithmetic techniques; ROBERT L. HERSHEY, *How to Think with Numbers* (1982), is an analysis of consumer applications of arithmetic; A. IBN I. AL-UQLIDISI, *Arithmetic of al-Uqlidisi,* trans. by A.S. SAIDAN (1978), is the story of Hindu-Arabic arithmetic; WILLIAM J. HEMMER, *Arithmetic by Example* (1979), is a developed presentation on the elementary level; PETER HILTON and JEAN PEDERSEN, *Fear No More: An Adult Approach to Mathematics* (1983), shows sophisticated applications of elementary material; SANDRA PREIS and GEORGE COCKS, *Arithmetic,* 2nd ed. (1980), is a programmed text that can be successfully used for self-instruction; STEPHEN P. RICHARDS, *A Number for Your Thoughts: Facts and Speculations About Numbers from Euclid to the Latest Computers* (1982), is a lucid explanation of number theory for a wide range of readers.

(C.C.MacD./D.E.S./W.J.LeV.)

Art Conservation and Restoration

Conservation, including maintenance and preservation (protection from damage or deterioration), and restoration have become an increasingly important aspect of the work not only of museums but also of civic authorities and all those concerned with works of art, whether artists, collectors, or gallerygoers. Technical advances of the 20th century have made possible safer methods of cleaning and repairing objects. Art restoration has become an important tool of research, and it enables the viewer to appreciate the original intention of the artist.

This article is divided into the following sections:

ARCHITECTURE

The conservation and restoration of older architecture is an increasing modern preoccupation. The earliest buildings that have survived generally tend to be those that received religious veneration. When these structures were no longer venerated, they disappeared like other buildings. Even the famed ancient Egyptian Sphinx at Giza lay for centuries under the sand, and it was not until the early 19th century that the Forum of ancient Rome was uncovered and explored.

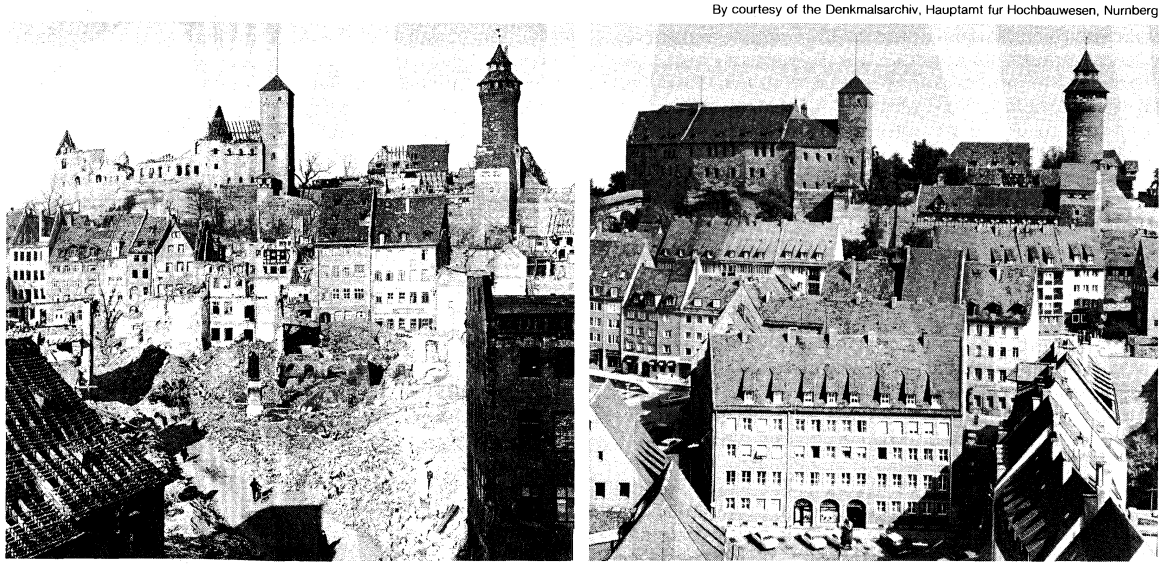
Medieval builders treated the work of their forebears with a healthy lack of awe. Every new Gothic chapel or chantry and virtually every stage in the development of a single Gothic cathedral followed the style of its own day. With the Renaissance in Europe grew a new respect for classical antiquity and a new interest in its architectural forms. By the end of the 18th century a knowledge of archaeology had become an accepted accomplishment of the educated man. Architectural design itself became a matter of "correctness." Old buildings everywhere began to be "restored" to the style of periods especially favoured. The French architect and writer E.-E. Viollet-le-Duc brilliantly restored the Sainte-Chapelle (1840–67) and the cathedral of Notre-Dame de Paris (1845–64). The ancient walls of Carcassonne in France and of Windsor Castle in England were not only repaired but also largely rebuilt.

With the spread of the Industrial Revolution and the increasing reliance on mechanical processes, the labour of hands became more costly, and the value of craftsmanship gained a new significance. Old buildings, which often exhibited the personal touches of master craftsmen, began to command a new respect, and the English art critic John Ruskin (1819–1900) was even able to assert that "the greatest glory of a building is its age." In 1877 the pioneers of the conservation movement, led by the English artist and writer William Morris (1834–96), founded the Society for the Protection of Ancient Buildings (SPAB). Nicknamed Anti-Scrape, the society vehemently opposed the indiscriminate refacing of old stonework and the "conjectural restorations" still so fashionable, such as the new west front of St. Albans Cathedral in England (1880–83). The movement gathered force, and in the 20th century groups throughout the world now devote their efforts to architectural conservation.

An added local impetus has been given by national pride; in countries like Poland, postwar reconstruction became the symbol of national resurgence. Almost every civilized country is increasingly conscious of its heritage of ancient buildings, while cultural bodies such as the United Nations Educational, Scientific and Cultural Organization have lent to the conservation movement a powerful international impetus.

Effects of economic and social change. The development of architecture can be read as a sensitive index of social change. The economic climate and social preoccupations of each age have combined to generate its own architecture and its own towns. Almost every decade of new building displays its own peculiar characteristics and modifies by constant adaptation the buildings and towns of yesterday. But the urban environment is society's investment in its future, and the cycle of renewal is continuous,

By courtesy of the Denkmalsarchiv, Hauptamt fur Hochbauwesen, Nurnberg



Albrecht Dürerplatz, Nürnberg, (left) after bombing in 1945 and (right) after reconstruction.

if often slow. Thus, the problems of building maintenance and renewal are complicated by long-term economic and social change.

Today the most marked trends are still those that brought about the conservation movement itself. First is the accelerated pace of physical growth. Old buildings have become not only relatively rarer but often virtually irreplaceable in terms of labour and craftsmanship and sometimes of materials. In many cases, old buildings give to a locality much of its special character and identity, as, for example, in those English country villages where thatched roofs still predominate. Another and rarer asset is a sheer and intrinsic merit of architectural form. And alongside all these is the tangible evidence that any old building provides for its community of a kind of social and environmental continuity—a reassuring reference point in a constantly changing world.

he factor
land
lue

Under the increasing pressure of population, the value of urban land climbs steeply, with some curious effects on old buildings. Increased demand brings increased values and, at first, better prospects of repair and maintenance. But as values rise higher, the older building must also justify itself in terms of economic efficiency. All over the world, the town houses of the 19th century and earlier serve with varied grace in the 20th century as centres of modern industry and commerce. Their fabric is subjected to new strains, and their room shapes and capacity may become incompatible with new and changed demands. There comes a point at which the old building on a valuable town site can compete no longer with redevelopment. Then it is quickly overtaken, and financial subsidy is powerless. The old building in a deteriorating neighbourhood is at the same time likely to be in no better a situation. Its maintenance may become no longer worthwhile, condemning it to early death by neglect. The destructive effects of both over- and undervalue are clearly displayed side by side in a fine Georgian city like Dublin or in once-distinguished neighbourhoods like Bloomsbury in London or the Marais in Paris. The most successful neighbourhood conservation occurs where values have been held in pace with the architectural capacity of a community, as at Bath in England, or in the Georgetown section of Washington, D.C.

he
ctor of
creased
obility

Another social change is the rapid increase in mobility. The automobile brought better roads and an incentive to use them. Old city centres, after centuries of essentially domestic life, began to be abandoned in favour of ring upon ring of suburbs. This peripheral accretion of cities is allied with their central decay as communities. As a universal result, the twice-daily thrombosis of the highways urges on a constant process of road widening, in which many an intervening historic area has been completely eroded away.

Role of the law. In all conservation of architecture, the first effective step is to decide and define what buildings or sites are worthy of protection. For most countries this has involved a systematic process of inventory and survey. In Great Britain, for example, the Royal Commission on Historical Monuments (RCHM) was set up in 1908, and the Civic Amenities Act of 1967 enabled local planning authorities to define special areas for "conservation and enhancement." In France, the Commission des Secteurs Sauvegardés was set up in 1962 under André Malraux, minister for cultural affairs, to pursue an active program for public protection of historic areas. In the United States, the Historic American Buildings Survey was designed to assemble a national archive of historic American architecture.

riteria
or con-
ervation

Criteria for conservation are rarely well defined. Architectural merit clearly must rank highly—especially in the case of any building that authentically exemplifies its period. Historical associations, such as the birthplace of a famous person, are less easily rated. One pernicious effect of all selection is the way in which it is the most outstanding example of any period, rather than the truly typical, that in the end remains to represent it. Another is that defects as well as merits may be kept warm under the same blanket. This is particularly so in the larger groups of buildings that are coming to be recognized as worthy of conservation.

Once defined, a building's next defense is in specific legal powers for its protection. These may be of varied degree and effectiveness. The most obvious form of legislation is the restriction against demolition. A higher degree of legal sophistication occurs in powers for the annexation of property and its maintenance by the state. Covenanted rights and restrictions are a variant of this principle. Next in the scale of effectiveness comes positive encouragement to owners by means of grants, bringing a public share and interest in the work of repair. In this way, actual legal rights over private property may be confined to a minimum while finance is encouraged from private pockets. Probably the most effective ultimate defense is selective protection, exercised as a regular part of everyday town- and country-planning control.

Negative legislation itself varies in degree. In Italy it is possible to insist upon the return even of certain pictures or chattels illegally dispersed from a building where these are adjudged to be of sufficient national importance. But negative powers are inherently weak. They convey no control over the philistine or intransigent owner and, at best, can only slow down neglect and demolition, whether deliberate or otherwise.

The national acquisition of buildings for conservation in Britain has been carried out chiefly under the Ancient Monuments Consolidation and Amendment Act of 1913, by which suitable unoccupied properties can be "taken into guardianship." A much more rigorous application of the principle is sometimes possible in the United States, whereby the owners of whole groups of buildings held to be of sufficient distinction can in fact be legally dispossessed. These erstwhile owners may then be allowed to remain in residence on condition of the repair and rehabilitation of their buildings to a specified standard. In this way, whole areas of buildings, such as Society Hill in Philadelphia, have been taken over, concentrated redevelopment by high-rise apartments being permitted in selected inner locations, while old buildings with frontage are restored in period styles.

Williamsburg, Virginia

The most exhaustive of all restoration projects is in the United States, at Williamsburg, Virginia. This 170-acre town, the colonial capital of Virginia from 1699 to 1780, has attracted the most expensive restoration program ever undertaken. Commenced in 1926, the project is dedicated to the purpose "that the future may learn from the past." Careful and scholarly restoration has been completed on more than 500 buildings. Environmental management is of a high order. Tourist automobile traffic is excluded from the restored area in season, when a free bus service is provided. The emphasis is frankly educational. The enterprise not only owns its buildings but also staffs them, its employees wearing correct period costume.

One of the most dramatic rescue operations has been in Egypt, where the ancient temples (*c.* 1250 BC) of Abu Simbel were threatened with destruction by the rising waters of the Aswān High Dam. They were sawed into giant blocks and successfully reassembled 200 feet (60 metres) above the original site. This act of preservation was the result of intensive international negotiation and expertise.

Another variant on public ownership may be found in acquisition by a private body, such as the National Trust in Great Britain. Founded in 1895, this property-owning body opens to the public several hundred of its properties. The trust receives no direct government subsidy and relies upon careful economic management, although certain legal preferences operate in its favour. In the United States the National Trust for Historic Preservation operates in a similar way.

Among bodies devoted to grant aid, the Historic Buildings and Monuments Commission for England (as successor to the Historic Buildings Council) disburses grants within a modest annual budget, largely to help building owners penalized by heavy estate duties. These grants are administered to encourage owners to take a pride in their own buildings. The commission is also responsible for the management of more than 400 monuments in the nation's care.

A pioneer training program in architectural conservation has been established by the Faculty of Architecture of

Rome University. Of six months' duration, the course provides specialist training in conservation for architects of all nationalities. In many countries, comparable courses are now available to meet the need for suitably qualified and experienced architects.

Techniques of building conservation. The first requisite in conserving any building is a sensitive assessment of its history and merits. Every building has its own biography. The Parthenon in Athens, originally built (447 to 432 BC) as a temple, subsequently served as a Christian church, a mosque, and a powder magazine before it became one of the world's greatest attractions for the tourist and art lover. A knowledge of the whole life of a building brings an essential understanding of its features and its problems.

Next, the conservator needs a thorough, measured survey. Generally, this is prepared by hand, with tape and rod and level. Modern measuring techniques, including photogrammetry and stereophotogrammetry, are also used and are quick and remarkably accurate.

Assessing the structure's soundness

Third, the architect or surveyor analyzes the structural stability of the subject and its living pattern of movement. No structure is permanently still. Subsoil expands and shrinks, thrust moves against thrust, and materials move with heat and wind. Forceful exercises, like English bell ringing, have an even greater effect on a building's stability. Clay soil is the worst: the building protects the ground underneath but not around; and, with every downpour, a wall on saturated clay may vary the lean of the building. Many ancient buildings had piled foundations—at Winchester, the cathedral was supported on oak piles, which rotted over the centuries. In order to underpin the structure, a diver worked for months in the waterlogged soil. Framed structures can move a great deal. The skeleton of a timber-framed medieval house can be extremely crooked without losing strength, if it is well triangulated and its joints are sound. A wall is theoretically safe until it leans far enough to develop tension on one side, yet even then it may be stiffened by structural cross-walls. Generally, the old, evenly spread load will be stable, and any new point load or thrust will be suspect. The surveyors may check the observations over a period; *e.g.*, by measurement with plumb lines or by simple "tell-tales" (marking devices) set across a crack, or now by electronic measuring devices of remarkable accuracy.

The surveyor lastly tests all services, especially electrical wiring, with its risk of fire; gas lines, with their perils of seepage and explosion; and plumbing, with its danger of leaks. These services are frequently redesigned and simplified as well as improved. Lightning conductors and fire-fighting equipment are an important part of the protection of any ancient building.

The conservator must analyze the good points and bad points of the building, in the context of its current and future use, and define remedies in terms of their relative urgency. He can then prepare a balanced and phased conservation plan, related to the available budget.

Remedying building defects

The first remedial task is to stabilize and consolidate the structure. Ideally, this is best done by restraining, or tying, the point of active thrust and then by replacing, splinting, or in some way giving fresh heart to any failing or defective member. Adding heavy weights such as buttresses can do more harm than good. A load can frequently be spread more widely or more evenly. A structure can, in effect, be corseted by inserting (for example, around a tower) a continuous beam or ring of concrete. This can be done even in delicate masonry and, as in underpinning, by removing alternate sections of a wall, threading in reinforcement, and casting successive sets of concrete stitches, which unite into one strengthening beam. Sometimes a metal rod or tie-bar may be inserted along a direct line of thrust or weakness, linking structural elements in need of support.

After structural movement, the next serious adversary in building conservation is damp. Not only of itself but also allied with almost every other trouble, damp accelerates decay. Weather may be penetrating through whole surfaces, such as porous brickwork, or finding its way through cracks or defects in the roofing. Especially vulnerable are gutters or any part of the rainwater-collecting system. Wet weakens walling, rots timbers, and spoils finishes. The remedy may involve renewing roof finishes. It may entail inserting a continuous moisture barrier, perhaps in a modern material such as stout polyethylene. In this case, special care is needed to avoid future damage by concentrating more trouble at any possible defect. Techniques of waterproofing wet walls include the insertion of high-capillary tubes, designed to draw the moisture to themselves and to expel it, and also the injection of silicone or latex and similar water-repellent solutions into the heart of the walling. Simple methods are best. The traditional ditch, or dry area, drained if necessary, disposes of the water before it reaches the wall. Double or cavity walls, with air between them, are another defense against damp.

Again, dampness compounds decay, and the first attention should be to protective features such as copings. Both in stonework and in brickwork, much harm can be caused by damp, especially when allied with an overly hard mortar jointing. This traps moisture along the lines of the joints, bringing any harmful salts to the surface, where they crystallize and damage the facing. Mortar jointing should always be softer than the brick or stone of a wall.

Much decay is the result of poor construction. Defects are almost always accelerated by the simple contravention of good building practice. In walling, a typical cause of structural instability is a double-skin construction with rough rubble between in which, by uneven loading, one skin has been caused to bulge and to release loose material in the core of the wall. Once on the move, this rapidly gains momentum as a live wedge, forcing apart its two faces. The conservator will insert temporary support, then remedy any uneven loading and rebuild the affected area. In some cases, after loose material is washed out, the unseen cavities can be grouted up, either by gravity or at high pressure, thus strengthening a wall without disturbing the facing stonework.

The roof is a building's first defense. It must be impervious and collect water clear of a building. Roof finishes are commonly either of unit materials such as tiles, slates, or stone, or of boarding covered in sheet metal, such as lead. The failure of unit materials is usually caused by decay of fixings. Iron nails are especially destructive and are best replaced by nonferrous materials, such as copper. The battens that carry the tiles or slates have a longer lifespan but also need periodic renewal. Leadwork failure is usually the result of sheer age. This material has a very long life but, if used in sheets of excessive size, has a tendency to buckle and creep as a result of expansion—especially in sunshine. Leadwork can readily be recast or can be repaired by lead burning a new patch to the original lead. Soldering is less reliable and tends to crack away.

The chief enemies of timber are the natural predators of the forest—fungi and wood-boring insects. The most voracious fungus that attacks building timbers is dry rot (*Merulius lacrymans*). This can spread along infected wood to sound timber, carrying its own moisture supply. It extracts cellulose, which forms the chief part of plant cells, and leaves behind a tindery and useless shell. Stagnant air and warmth accelerate its spread. Eradication must be thorough, or the trouble will rapidly reestablish itself. Modern fungicides are highly effective.

Wood-boring insects include the furniture and deathwatch beetles. From eggs laid in cracks, the larvae tunnel into timber and damage it before emerging as beetles to lay more eggs. The deathwatch beetle inhabits mostly the outer sapwood of oak, when wet or softened by rot. The furniture beetle lives mostly in deal, especially when sappy or damp. Both can be eradicated with modern pesticides.

"Daily care"

Regular maintenance is the key to building conservation; William Morris called this practice "daily care." A building's life can be long, human tenancy relatively short. Yet the cumulative effect of neglect can be desperately damaging. Conversely, a sensitive awareness of a building's needs, with regular attention to them, will extend its life and promote its long enjoyment. The successful conservator identifies himself with a building's life, its structure and demands, with the special needs of an occupant, and with the skills of today's craftsmen. In this spirit, he can hand on to the future the best of the past.

(D.W.I.)

PAINTINGS

The conservator of paintings aims above all at "true conservation," the preservation of the objects in conditions that, as far as possible, will arrest material decay and delay as long as possible the moment when restoration is needed. The correct choice of conditions of display and storage is, therefore, of the first importance. Ideally, each type of painting requires its own special conditions for maximum safety, depending on the original technique and materials used to compose it. Broadly speaking, most paintings can be divided into (1) easel paintings, on either canvas or a solid support, usually wood; (2) wall, or mural, paintings; and (3) painting on paper and ivory.

Easel paintings. More or less portable paintings on canvas or panel are called easel paintings. Basically, they consist of the support (the canvas or panel); the ground, ordinarily a white or tinted pigment or inert substance mixed with either glue or oil; the paint layer itself, which may be complex in structure; and, finally, the surface coating, usually a varnish, to protect the paint and modify its appearance aesthetically. These four layers have many variants but must be constantly borne in mind when considering the problems of conservation.

fects of
midity

Paintings on wood. Wood-panel supports were used almost universally in European art before about 1450, when canvas began to gain ground. Wood has the disadvantage of swelling and shrinking across the grain with variations in the relative humidity of the atmosphere. In northern temperate climates, variations in humidity can be considerable. In England, for example, the seasonal variation in a museum that is centrally heated in the winter can be from 25 percent in midwinter to 90 percent in summer. Although paint has a certain elasticity, it cannot usually take up much movement and generally cracks in a network referred to as craquelure. In continental landmasses, such as the United States, the average relative humidity in dry zones may be consistently low, so that European paintings with wooden supports air-seasoned to a higher humidity may suffer considerably. In both Europe and the United States, the effect of an unsuitable environment of low or changing relative humidity and the restraining effect of the paint layer often produces a permanent bowing of the panel, which is convex at the front surface. To counteract both the shrinkage and the bowing (especially the latter), restorers in the past placed wooden strips called battens or more complex structures across the back of the panel as constraints. This solution, however, often led to severe distortion of the front surface and cracking of the whole panel in lines along the wood grain. Extensive damage to the paint sometimes occurs, and drastic restoration is needed. In terms of preservation, the ideal solution is a form of air conditioning in which the relative humidity is maintained as nearly constant as possible at what is generally agreed to be the most reasonable level; *i.e.,* about 55 percent.

se of
condary
ipports

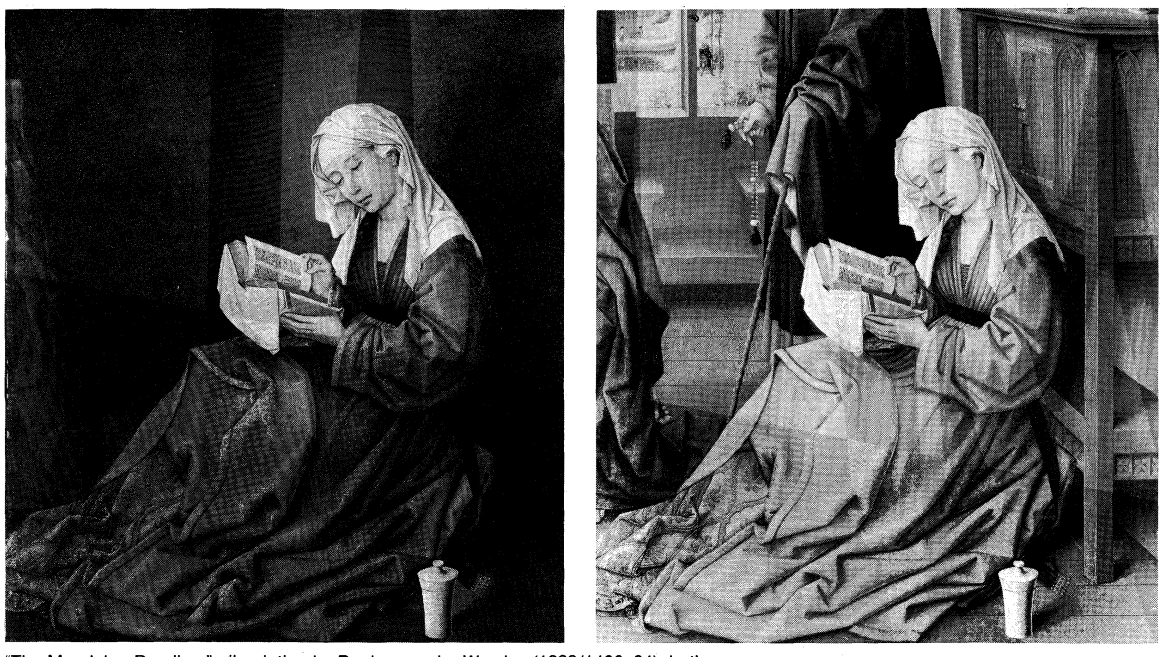
When warping and cracking have already occurred or when the latter seems likely as a result of the mistaken application of secondary supports, such as cross-battens, expert restoration treatment is required. In principle, this consists of removing the cross-battens and applying a reinforcement to the back that imposes a uniform but gentler constraint over the whole surface. It is normal in the 20th century to accept as inevitable some permanent convex curvature. The adhesives used and the composition of the new secondary support take many forms. One consists in backing the panel with strips of a very light, open-textured wood (balsa), using as a cement a mixture of beeswax, a natural resin, such as dammar, and an inert filler. This thermoplastic cement, which is applied as a hot, creamy liquid, solidifies without contraction. The epoxy resins, which also harden without contraction, have been used as well and have the additional advantage of not requiring heat. The material and cement used are chosen according to the nature of the original panel. Some restorers reduce the strength of the original panel, before applying the secondary support, by reducing its thickness. This practice is not universally approved. Occasionally, when the panel is badly worm-eaten or severely cracked, it has to be removed from the paint and ground altogether in the process known as transfer. This is accomplished by pasting a substantial support of paper and, possibly, canvas to the front surface and then gently gouging away the wood on the back. An entirely new, inert support of balsa wood or compressed board is then cemented to the back and the facing removed.

Paintings on canvas. A canvas support expands and contracts with variations in relative humidity, but the effect is not as drastic as with wood. Canvas, however, will deteriorate with age and acid conditions. In many cases, parts of the paint and ground will lift from the surface, a condition described as flaking, blistering, or scaling. In the case of paintings on canvas, the process of transfer is almost never performed. Instead, the canvas is reinforced at the back by attaching a new canvas to the old. This lining process (almost always referred to as relining) can be done in two ways. The traditional method consists of ironing the new canvas to the old, using as adhesive a warm, fluid mixture of animal glue and a farinaceous paste, sometimes with the addition of a small proportion of plasticizer. This method is still used, especially in Italy. It has the advantage that the heat and moisture help to flatten raised paint ("cupping") and local deformations and tears in the canvas. Another method, introduced after the mid-19th century, uses the thermoplastic wax–resin mixture mentioned above, omitting the filler. Originally done with heated irons as is the glue–paste method, it gained in popularity by the introduction, around 1950, of the so-called vacuum hot-table. The canvases are coated with molten adhesive (at about 160° F [70° C]), which is allowed to solidify, and then joined together on an electrically heated platen. They are then covered with a membrane enabling the space between to be evacuated with a pump through holes in the table. Adhesion occurs on cooling. Though the wax protects the canvas from deterioration, the vacuum pressure sometimes makes the texture of the canvas more apparent. Also, wax penetrating the canvas occasionally darkens thin or porous paint layers. To overcome these defects, "heat-seal" adhesives were introduced in the late 1960s. Formulations containing synthetic resins, including polyvinyl acetate, and, increasingly, an ethylene–vinyl acetate copolymer, are applied in solution or dispersion to the surfaces and, after drying, are adhered on the hot table. More recently, cold-setting polymer dispersions in water have been introduced using a "cold-table," from which the water is removed through spaced perforations with a powerful downdraft of air. Pressure-sensitive adhesives also are being evaluated.

Paintings have occasionally in the past been transferred from wood to canvas by a variant of the treatments described above. The reverse of this—*i.e.,* attaching a painting on canvas to a stable rigid support (a process known as *marouflage*)—is still sometimes done for various reasons.

The ground (*i.e.,* the inert paint layer covering the support below the painting itself) can ordinarily be regarded for conservation purposes as part of the painting layers. Occasionally, when the ground is composed of glue and an inert substance such as whiting or gypsum, the glue may deteriorate and the ground lose its adhesion to either the support or the paint layers. In extreme cases, with wood-panel supports, a complete transfer is required, in which not only the support but also the ground must be removed. The restorer has a brief opportunity of seeing the painting or at least its lowest layers in reverse before applying a new ground and support.

The paint layers themselves are subject to a number of maladies as a result of natural decay, faulty original technique, unsuitable conditions, ill treatment, and improper earlier restorations. It must be remembered that, whereas housepaint usually has to be renewed every few years, the paint of easel paintings is required to survive indefinitely and may be already 600 years old. The most prevalent defect, as mentioned above, is flaking, or scaling, where, in local areas of paint, small particles become partly or wholly detached from the support. If the loss is not total, the paint can be secured, according to circumstances, with a dilute gelatin adhesive, the wax–resin adhesive, or a synthetic polymer. The paint is usually pressed firmly into place with an electrically heated spatula. For easel paint-



"The Magdalen Reading," oil painting by Rogier van der Weyden (1399/1400–64). In the National Gallery, London. Cleaning and restoration in 1956 (right) revealed that the painting was a fragment from an altarpiece, not a complete painting as it had appeared (left).
By courtesy of the trustees of the National Gallery, London

Problems in the original paint

ings the binder for the coloured pigments is usually egg yolk, oil, or, occasionally, glue. The first type of method, called egg-tempera painting, was universal before the mid-15th century, when oil painting began to be used increasingly. The condition of egg-tempera paintings where damage has not been caused by deterioration of support or ground is usually good. The condition of oil paintings is often less satisfactory. Sometimes the original technique of the artist is at fault, and this becomes increasingly so from the 18th century onward. Too much oil may have been used, leading to ineradicable wrinkling, or superimposed layers may have dried at different rates, producing a wide craquelure as a result of unequal shrinkage. An enhanced version of the latter occurred increasingly, as the 19th century progressed, by the use of a brown pigment called bitumen. Bituminous paints never dried completely, producing a surface effect resembling crocodile skin. These defects cannot be cured and can be visually ameliorated only by judicious retouching.

The most notable defect arising from poor conservation is the fading or changing of the pigments by excessive light. Although this is more evident with thin-layer paintings, such as watercolours, it is also visible in oil paintings. The palette of the earlier painters was, in general, stable to light; however, some of the pigments used, notably the lakes, which consisted of vegetable dyestuffs mordanted onto translucent inert materials, often faded easily. A transparent green, copper resinate, much used from the 15th to the 18th century, became a deep chocolate brown after prolonged exposure to light. After the discovery of synthetic dyestuffs in 1856, a further series of pigments was created, some of which were later discovered to fade rapidly. Unfortunately, it is impossible to restore the original colour, and in this case conservation, in its true sense of arresting decay, is important; *i.e.,* to limit the light to the lowest possible level consistent with adequate viewing—in practice about 15 lumens per square foot (15 footcandles; 150 lux).

Almost every painting of any degree of antiquity will have losses and damages, and a painting of earlier than the 19th century in perfect condition will usually be an object of special interest. Before a more conscientious approach to restoration became general in the mid-20th century, areas that had a number of small losses were often—indeed, generally—entirely repainted. It was considered normal in any case to repaint not only losses or gravely damaged areas but also a wide area of surrounding original paint, often with materials that have visibly darkened or faded with time. Large areas with significant detail missing were repainted inventively in what was supposed to be the style of the original artist. It is customary nowadays to repaint only the actual missing areas, matching carefully the artist's technique and paint texture. In some cases, as with skies in which areas have been worn away by injudicious cleaning, very little repainting is done at all.

Inpainting

Some restorers adopt various methods of inpainting in which the surrounding original paint is not imitated. The inpainting is done in a colour or with a texture that is intended to eliminate the shock of seeing a completely lost area without actually deceiving the observer. The aim in inpainting is always to use pigments and mediums that do not change with time. Egg tempera has been preferred, though it is difficult to use, and various stable, modern resins are employed in place of or in addition to tempera. Exact imitation of the original entails close study of the painter's technique, especially the multilayer methods, since the successive layers, being partly translucent, contribute to the final visual effect. Minute details of texture, brushstrokes, and craquelure must also be simulated.

This work of inpainting ordinarily has to be done after the top layer, or the varnish, has been removed. Because all varnishes before about 1930 and many since have undergone changes in colour and transparency, they partially obscure the appearance of the original paint and, therefore, must be taken off.

Varnishing

While the use of varnish was partly to protect the paint from accidental damage and abrasion, its main purpose was to improve the appearance. Oil paint, over the course of time, changes chemically by oxidation and polymerization and becomes harder and more brittle. A remarkable range of injurious cleaning agents, including sand and caustic alkali, is known to have been used for cleaning the surface of the paint in the past. As a consequence, some of the hardened oil medium is lost, and the painting becomes matte, or without lustre, and lifeless. A varnish, visually at least, revives it, and if a varnish itself becomes matte a further coat of varnish revives the former varnish coat, and so on. Unfortunately, the varnishes used consisted of hard resins, such as copal, or, more often, soft resins, such as mastic and dammar. These become yellow, brittle, and slightly opaque and also less soluble in harmless solvents. Occasionally, as in London's National Gallery in the mid-19th century, lead driers were added to the varnish to quench the bloom, a blue haze that covered the varnish as

a consequence of the prevalent coal smoke combined with a high and variable humidity. This varnish, known as the "Gallery Varnish," yellowed even more rapidly than the resins alone. The removal of these disfiguring natural-resin varnishes from paintings is an operation that must be carried out with great skill to avoid damaging the original paint. The work of restorers in removing varnish (usually described as cleaning) has been the subject of occasional vehement public criticism from 1850 onward.

When the varnish is in good condition but covered with grime the restorer may, after close inspection, clean the surface with aqueous solutions of nonionic detergents or mild solvents. This should never be attempted by an amateur. Varnish is almost invariably removed by means of a solvent mixture of which the active ingredient is often one of the lower alcohols. The solvent is sparingly applied with a cotton-wool swab. Choice of solvent mixture and mode of application has always depended on the skill and experience of the restorer, but modern scientific theory has clarified the procedures. For revarnishing, the natural resins such as dammar, although excellent in application and appearance, have mostly been abandoned in favour of synthetics. They are chosen for chemical stability with regard to light and the atmosphere so that they can eventually be removed by safe solvents and will not rapidly discolour nor physically deteriorate. Acrylic copolymers and polycyclohexanones have been the most commonly used since the 1960s. Research continues, however, in order to find the "ideal" varnish, combining ease of application, chemical stability, and an acceptable aesthetic quality.

Wall paintings. From the point of view of conservation, the different types of wall painting have a number of features in common, though the techniques of restoration required for each inevitably differ in detail. Among the wall painting techniques is buon fresco, or true fresco, in which pigments mixed with water are painted onto a freshly prepared layer of damp lime plaster. Fresco secco is a method, often used in conjunction with buon fresco, in which a mixture of pigment and egg tempera is painted onto the dry plaster or is used as a retouching or enhancement of a dried buon fresco painting. Wall paintings are also executed with pigments mixed in oil applied either to a prepared dry plaster wall or on canvas, which is then fixed to the wall.

As far as pure conservation is concerned, there are two outstanding factors. The first, which applies to all methods of wall painting and especially to aqueous, or water-based, mediums, is the exclusion of damp. This can attack the painting from several sources. One source is damp rising through the walls of a building; this first affects the bottom of the wall painting and then spreads upward. This is prevented by inserting a metallic or resinous damp course. New damp courses in old buildings are often prohibitively expensive, in which case a possible amelioration is to dig out exterior soil to a depth of at least six inches below the interior floor. The second source of damp is from the outside wall. It is important at least to avoid treating the painting with a water-impermeable material, such as wax or silicates, so that the damp can penetrate freely without meeting a barrier at the inner surface. The third source is condensation on the inner surface, which is particularly prevalent in churches that are heated only on weekends. More continuous and uniform heat is the solution, provided that the air is not dried out so rapidly that efflorescence, the formation of a powdery surface, occurs. The fourth and most easily remedied source, though often neglected, is from leaking roofs and clogged drainpipes.

Deterioration of lime

The second important hazard is more insidious. It affects solely those murals painted on lime mortar, which inevitably, by the action of air, becomes calcium carbonate. Since 1900, with the increasing use of motor vehicles and of fuel-burning industries, the percentage of sulfur dioxide in the atmosphere has greatly increased. In the presence of moisture the calcium carbonate is changed to calcium sulfate, whose volume is almost twice that of the original carbonate of the mural. As a result, disintegration in some areas of a mural can be rapid. In Italy this sort of disintegration has greatly increased and has made necessary the development of drastic though highly expert methods of transfer of frescoes from the original walls. These range from the method of *strappo* to that of *stacco*. While in practice they are not always clearly distinguishable, *strappo*, the more usual method, consists in gluing canvas firmly to the surface of the fresco, followed by pulling and easing away with long spatulas a thin layer of the plaster that contains the pigment particles of the fresco. The bond between the facing and the fresco must be stronger than the internal cohesion of the plaster. Excess plaster is removed, revealing the fresco in reverse. This is then fixed to a rigid support with synthetic resins, using inert substances mixed with resins as an intermediate layer to simulate optically the original underlying plaster. In the *stacco* method, a thicker layer of plaster is removed with the fresco and is smoothed flat on its back surface before sticking the rigid composite layer to a board. Where possible, consolidation without detachment is performed. The removal of previous repaintings and overlying whitewash is often the most tedious part of the work.

In humid, temperate climates, such as England's, limewater is usually used as a consolidant. Earlier consolidations, often of wax or natural resins, are not only difficult to remove but also have frequently accelerated deterioration. In dry parts of the world, synthetic resins such as polyvinyl acetate have been used with success as consolidants.

Paintings on paper and ivory. Environmental conservation for these objects, which are ordinarily painted in an aqueous medium, consists in maintaining a stable relative humidity in the region of 50–60 percent. At lower humidities, both paper and ivory (the latter often used as a thin layer for portrait miniatures) tend to shrink, and the former becomes more brittle. The thin ivory of miniatures, which often tends to crack, may crack along the grain if constrained under conditions of varying humidity. At higher humidities, there is a possibility, especially when ventilation is poor, of mold growth, which can occur above about 68 percent relative humidity. Watercolour paintings are particularly vulnerable to light, which ideally should not exceed 5–10 lumens per square foot (5–10 footcandles, 50–100 lux). Some pigments fade rapidly, whereas others do not alter, and there is inevitably not only a loss of colour but also a distortion of the artist's intention. It should be noted that a warm light, as from an ordinary incandescent lamp, is less damaging in general than an equal amount of daylight or light from a fluorescent lamp. Daylight should preferably be avoided. The ultraviolet component should, in any event, be removed. Ultraviolet filters are available for windows and fluorescent tubes.

Restoration of paintings on paper has many detailed variations. After removal of the material on which the paintings are mounted, local brown stains, usually known as foxing, are sometimes reduced. The painting is freshened by washing gently in water with or without a little neutral detergent (which should not be of the household variety). Often a watercolour painting will resist washing without loss of colour, but it is generally advisable merely to damp the back before proceeding to reduce the stains locally with an oxidizing bleach. A mild form of bleaching agent known as chloramine-T is sometimes used. Other, stronger oxidizing bleaches can be used subsequently, but there are various disadvantages. Stains other than the characteristic foxing must be identified and the specific solvents used. Japanese prints may be treated similarly, though with even more care, since some colours (notably, a range of mauves) must never be damped. Portrait miniatures on ivory require expert treatment, and it is even possible to damage them irrevocably in removing them from their lockets.

The conservation of fine prints is dealt with in the article PRINTMAKING. (N.S.B.)

SCULPTURE

Until the mid-1960s, painting conservators led the whole field of conservation with their technical expertise and experience. Today that imbalance has shifted, and it is in the field of sculpture conservation that many of the most complex and exciting technical developments are taking place. This change can be attributed partly to the growing international concern with the problems of stone decay

and partly to the rapid development of more sophisticated synthetic resins for use in repair.

The realization that the heritage of stone sculpture and buildings might not remain in a recognizable form for more than another 50 years has stimulated an interest in the conservation of sculpture that would not have occurred were the problems merely confined to museum collections. New pressures also are being exerted on museums. As the main centres of conservation expertise, they are called upon to lend their aid to the organizations dealing with these problems in cathedrals, churches, and historic houses. They also act as places of refuge for endangered sculptures. The scale of the problem (the rapid increase in pollution worldwide and the vast number of sculptures involved) has caused the profession of sculpture conservation to develop rapidly. Although it has been possible to borrow from painting conservation many surface cleaning techniques for use on terra-cotta, plaster, and polychrome wood sculpture, very few of these techniques are suitable for marble and limestone, with their particular susceptibility to damage from water and soluble salt migration.

The main problems facing the stone conservator are stabilization, consolidation, and further protection against pollutant gases and soluble salts. Stone is extraordinarily unstable in the modern environment. Once it has been attacked by pollutant gases, such as sulfur dioxide, or migrating salts, such as nitrates or chlorides, it is difficult to return the stone to a stable condition, even when it is placed in a museum environment. Although some temporary stability may be achieved by putting a damaged sculpture in a temperature- and humidity-controlled glass case, it is commonly found that degradation will continue and, in certain instances, even accelerate.

The mechanism by which soluble salts cause damage in stone and marble is complex, but the broad effect can be simply characterized. These salts, which can be derived from pollutant gases, a marine climate, or natural ground salts, dissolve in water and seep through the pores of the stone. The water may derive from the ground, from the atmosphere, or from oversaturation of the stone by inappropriate water-washing. As the salts move to the drier air at the surface of the stone, they begin to crystallize either on the surface (efflorescence) or beneath the surface (subflorescence). It is this crystallization and resultant expansion of the salts that breaks open the pores of the stone and creates the damage.

To have any hope of halting this activity, the stone conservator must interfere with the deep structure of the stone, sealing it against moisture movement and strengthening it against salt damage to the pore structure. Probably the most popular means of stabilizing stone is the introduction of a consolidant. In the past, consolidants such as wax and shellac have been tried. These do not penetrate deeply into the stone and often aggravate the problem. Various synthetic resins, such as acrylics, epoxies, polyesters, and silicones, have been used with greater success. By far the most successful, however, have been the alkoxy–silanes. These have several distinct advantages over other consolidants. They penetrate deeply into the stone (two to three inches in some limestones), and they deposit a hard, almost indestructible network of silica in the pore structure of the stone, which waterproofs and strengthens it.

There are many forms of alkoxy–silanes in use and many ways of applying them. The commonest methods are simple brushing, spraying, and vacuum impregnation. Of these, the first is the most controllable and delicate approach, while the last is the least controllable and most potentially dangerous.

The full treatment of a sculpture must, of course, include cleaning. Any consolidation treatment or attempt to remove salts from stone must be carefully integrated with an appropriate cleaning system. In the past, the most common way to clean and desalinate stone was to immerse it in a tank of water for a period of weeks or months. This process can cause considerable damage because it loosens friable stone and pigment from the stone surface. A better method was developed in the 1960s, by which a clay poultice (magnesium silicate and deionized water) is used to suspend a thin layer of water over the surface of the sculpture, like a cosmetic mudpack, sucking out both dirt and salts. This treatment minimizes the contact with water and also does less harm to the fragile surface of the sculpture. The use of sophisticated tools and techniques such as ultrasonic dental descalers and abrasion by air-blasted microscopic glass beads helps to give the conservator much greater control over the cleaning process. Lasers were first used for removing pollution deposits from stone in 1970. Improved laser technology, decreasing cost, and the concurrent development of fibre optics suggests that it may soon be possible to produce a flexible precision tool that is capable of removing dirt and other encrustations from the surface of sculpture by vaporization, without harming the stone itself.

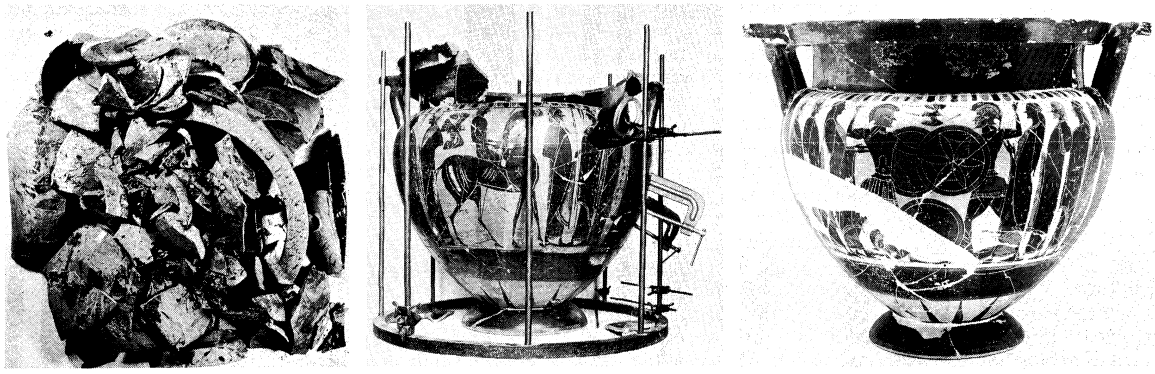
The other main advance in sculpture conservation has been the growing use of the binocular microscope. When the surface of a sculpture is examined under a surgical microscope of from 10× to 40× magnification, the immense damage that can be inflicted on sculpture by indiscriminate cleaning becomes strikingly apparent. Tiny fragments of original pigment often remain on sculptures that to the naked eye appear completely unpainted. From these pinpricks of paint, technicians can determine the exact composition of ancient sculptural pigments and their binding mediums. (J.H.L.)

DECORATIVE ARTS

Furniture. Apart from physical damage, effects of woodworm, and failure of glued joints, deterioration of furniture today is mainly attributable to the dry air (*i.e.*, low relative humidity) of centrally heated rooms in winter. In the past, the air in houses was not much different from that of the cabinetmaker's workshop. Dry conditions, however, cause the carcass wood to warp and split, joints to loosen, and surface decorations such as veneer to detach. Further damage can be eliminated by humidity control (within 50–60 percent relative humidity) the year round; a simpler alternative is the use of humidifiers in winter. Furniture should not be situated close to either humidifiers or sources of heat.

Restoration of fine furniture requires a range of skills similar to those of the original cabinetmaker, coupled with restraint and judgment with regard to the integrity of the object. Replacements required for mechanical or aesthetic reasons are preferably of matching material finished to simulate the original patina. Replacement of worm-eaten wood can now often be avoided by impregnation of the damaged original with resin, usually an epoxy. For live woodworms there are proprietary liquid pesticides, although care must be exercised during application to avoid damaging original finishes. A gas chamber using, for example, methyl bromide or a mixture containing ethylene oxide is more effective. Repair requiring specialized skills may be needed for a range of materials. These include tortoiseshell, ivory, pewter, mother-of-pearl, and the materials of upholstery. Oriental lacquer presents unsolved problems that are still the subject of international discussion.

Stained glass. Stained glass windows may include pieces of coloured glass, uncoloured glass to which a thin coloured layer has been applied (flashed glass), and, in examples from the 15th century and later, uncoloured glass with a painted and fired design of coloured enamel. Deterioration other than physical damage results mainly from rain and condensation and consists of overall corrosion on both sides and pitting, generally on the outside only. Susceptibility to corrosion varies with glass composition. The corrosion layer darkens the glass, but its removal is a matter of some controversy. A variety of methods of local abrasion may be employed. Missing pieces can be replaced with new coloured glass and glass enamelled and fired in accord with the design. Original glass from which enamel has flaked owing to imperfect firing cannot be refired and has to be retouched with a cold-setting paint. The tendency of the enamel to flake makes the simplest operation—grime removal by washing—a matter for particular care. Breaks, formerly mended with so-called strap leads, are now fixed with epoxy resins. Sometimes badly broken pieces are dry-plated with cover glasses, the pieces



Greek black-figure vase from Orvieto destroyed during the flood of 1966 in Florence, shown (left to right) before, during, and after restoration.
By courtesy of the Soprintendenza alle Antichita d'Etruria, Florence

being inserted together into the leading. At the Cologne cathedral workshop, all the pieces are plated, using a clear resin adhesive. Deterioration of leading makes releading necessary every century or so.

Organic surface coatings have been used for protection from the weather, but opinion is turning toward isothermal glazing, in which the window is replaced with plain glass, and the stained glass panel is hung inside with a narrow gap between the two, open to the interior atmosphere.

Textiles. Environmental requirements for textile preservation are similar to those for paintings on paper, but neglect of textiles is in general more damaging. Fading is a serious problem, but light also weakens the fibres of the material, especially silk. Gaseous air pollution is harmful, and soiling from airborne grime leads to the need for washing, which is best avoided. Where washing is necessary, nonionic detergent formulations are used, never ordinary commercial detergents. Dry cleaning with selected solvents may be substituted in particular cases. Handling and storage of fragile textiles require special care: loose wrapping with acid-free tissue paper; storage containers ventilated to avoid local humidity buildup; folding with sharp edges avoided; for tapestries, rolling with weft (design weave) along the axis; and so forth. New acquisitions and stored material require inspection for insect infestation. The feasibility of insect poisons and repellents in textile preservation remains uncertain.

Restoration of valuable textiles, generally by means of skilled needlework, does not normally involve the replacement of worn or decayed materials. When this has to be done for structural reasons informed judgment is required. Material that is so decayed that it cannot be reinforced by stitching to a backing material may require an adhesive bond. After decades of discussion over the use of synthetics, research now points to hydrolyzed starch (an old Japanese recipe) or, when use of water is inadvisable, methylcellulose in an organic solvent.

Ceramics. Ceramic restoration, generally of breakages and losses, sometimes has to begin with the often difficult removal of material from old repairs, and the reduction of ingrained dirt and stains. Rivets and dowels (from a former method of repair) are always removed. Choice of adhesive depends on type of object and probable degrees of stress; epoxy and polyester resins and polyvinyl acetate emulsions are among those used. For hard-bodied ceramics, missing portions or decorative details may be replaced with a suitably bodied and pigmented epoxy. A proprietary plaster filler is preferable for earthenware. Sticking and filling operations require skill to secure an exact fit and a perfect surface. Retouching, with brush or spray, is equally exacting; medium and pigments must be lightstable and the hardened coating physically durable. In museum practice the extension of the retouching to nearby original surfaces is avoided and only cold-setting media are used. Heat-curing risks damage to original glazes.

The apparently simple process of washing ceramics should in fact be carried out with care. Porous and soft-paste ceramics should not be immersed but cleaned with swabs of cotton-wool dampened with lukewarm water and, if necessary, a small amount of nonionic detergent. Hard-paste ceramics may be similarly cleaned but can be immersed unless there is gold decoration (which must be cleaned with special care) or metal mounts, which should never be wetted. Biscuit ware with ingrained dirt requires expert treatment. (N.S.B.)

BIBLIOGRAPHY

Conservation of buildings: BERNARD M. FEILDEN, *Conservation of Historic Buildings* (1982), a fully illustrated work on preservation techniques, one in the authoritative series of *Technical Studies in the Arts, Archaeology, and Architecture;* JAMES MARSTON FITCH, *Historic Preservation: Curatorial Management of the Built World* (1982); JOHN F. SMITH, *A Critical Bibliography of Building Conservation: Historic Towns, Buildings, Their Furnishings and Fittings* (1978); JACK BOWYER, *Vernacular Building Conservation* (1980), a technical guide to restoration; JANE FAWCETT (ed.), *The Future of the Past: Attitudes to Conservation, 1174–1974* (1976); DONALD W. INSALL, *The Care of Old Buildings Today: A Practical Guide* (1972); GREAT BRITAIN. PRESERVATION POLICY GROUP, *Report to the Minister of Housing and Local Government* (1970), a concerted attack upon the problems of historic city conservation in Britain; ORIN M. BULLOCK, JR., *The Restoration Manual: An Illustrated Guide to the Preservation and Restoration of Old Buildings* (1966, reissued 1983); and JANE JACOBS, *The Death and Life of Great American Cities* (1961, reissued 1972).

Conservation of paintings: HELMUT RUHEMANN, *The Cleaning of Paintings,* with a comprehensive bibliography by JOYCE PLESTERS (1968, reissued 1982); HAROLD J. PLENDERLEITH and ANTHONY E. WERNER, *The Conservation of Antiquities and Works of Art,* 2nd ed. (1971, reissued 1976); NORMAN BROMMELLE and PERRY SMITH (eds.), *Conservation and Restoration of Pictorial Art* (1976); NORMAN BROMMELLE, ANNE MONCRIEFF, and PERRY SMITH (eds.), *Conservation of Wood in Painting and the Decorative Arts* (1978); PAOLO MORA, LAURA MORA, and PAUL PHILIPPOT, *Conservation of Wall Paintings* (1984; originally published in French, 1977); and GREAT BRITAIN. NATIONAL GALLERY, *National Gallery Technical Bulletin* (annual).

Conservation of sculpture: SOPRINTENDENZA ALLE GALLERIE DI BOLOGNA, *La conservazione delle sculture all'aperto* (1971); *Preprints of the Contributions to the New York Conference on Conservation of Stone and Wooden Objects, 1970,* 2nd ed., 2 vol. (1971); *Deterioration and Preservation of Stones: Proceedings of the 3rd International Congress, 1979* (1979); NORMAN BROMMELLE, GARRY THOMSON, and PERRY SMITH (eds.), *Conservation Within Historic Buildings* (1980); *Adhesives and Consolidants: Preprints of the Contributions to the Paris Congress, 1984* (1984); and *Science and Technology in the Service of Conservation: Preprints of the Contributions to the Washington Congress, 1982* (1982).

Conservation of other works of art: The most valuable accounts are contained in the quarterly *Studies in Conservation,* published by the International Institute for Conservation of Historic and Artistic Work, London. See also ROBERT F. MCGIFFIN, JR., *Furniture Care and Conservation* (1983); S. LANDI, *The Textile Conservator's Manual* (1985); JUDITH LARNEY, *Restoring Ceramics,* 2nd ed. (1978); JOHN M.A. THOMSON *et al.* (eds.), *The Manual of Curatorship* (1984); and GARRY THOMSON, *The Museum Environment* (1978). The routine care of portable works of art is comprehensively treated in HERMIONE SANDWITH and SHEILA STAINTON (comps.), *The National Trust Manual of Housekeeping* (1984), based on experience in maintaining the contents of England's historic houses.

(D.W.I./N.S.B./J.H.L.)

Arthropods

Arthropoda is the largest phylum in the animal kingdom and includes such familiar forms as lobsters, crabs, spiders, insects, centipedes, and millipedes. About 84 percent of the known species of animals are members of this phylum, and they are very diverse in structure, in life-styles, and in types of habitat.

The distinguishing feature of arthropods is the presence of a skeletal covering composed of chitin (a complex sugar) bound to protein. This nonliving exoskeleton is secreted by the underlying epidermis (which corresponds to the skin of other animals). The body is usually segmented, and the segments bear paired, jointed appendages, from which the name arthropod ("jointed feet") is derived. More than 879,000 arthropod species have been described, of which about 86 percent are insects. This number, however, may be only a fraction of the total. Based on the number of undescribed species collected from the treetops of tropical forests, zoologists have estimated the total number of insect species alone to be as high as 10,000,000. The 30,000 described species of mites, another group of arthropods, may also represent only a fraction of the existing number.

The phylum Arthropoda may be divided into four subphyla: Trilobita, Chelicerata, Crustacea, and Uniramia. The subphylum Trilobita contains only the trilobites, which were the dominant arthropods in the early Paleozoic seas (570,000,000 to 225,000,000 years ago) but became extinct during the Permian Period (280,000,000 to 225,000,000 years ago), at the end of the Paleozoic Era.

Most members of the subphylum Chelicerata belong to the class Arachnida, containing the spiders, scorpions, ticks, and mites. They are largely terrestrial arthropods, living beneath stones and logs, in leaf mold, and in vegetation, but there are some aquatic mites that live in fresh water and in the sea. There are also many parasitic mites. Two small classes of chelicerates, the Merostomata, containing the horseshoe crabs, and the Pycnogonida, containing the sea spiders, are entirely marine. The merostomes are an ancient group and probably gave rise to the arachnids. Indeed, the earliest known fossil scorpions were aquatic.

The subphylum Crustacea contains mostly marine arthropods though many of its members, such as the crayfish, have invaded fresh water, and one group, the pill bugs (sow bugs) has become terrestrial, living beneath stones and logs and in leaf mold. In the sea, large crustaceans such as crabs and shrimps are common bottom-dwelling arthropods. Many minute species of crustaceans are an important component of the zooplankton (floating or weakly swimming animals) and serve as food for other invertebrates, fishes, and even whales.

Uniramia is the largest of the arthropod subphyla. It contains not only the class Insecta but also four closely related classes of long-bodied arthropods collectively known as myriapods: class Chilopoda (centipedes); class Symphyla (symphylans); class Diplopoda (millipedes); and class Pauropoda (pauropods). They are mostly terrestrial and, in contrast to the other arthropod subphyla, the uniramians are believed to have had a terrestrial origin. Centipedes, symphylans, millipedes, and pauropods live beneath stones and logs and in leaf mold; insects are found in all types of terrestrial habitats and some have invaded fresh water. The sea has remained the domain of the crustaceans, however, and only at its very edges are insects found.

This article discusses the arthropods as a group; for specific information on the most significant subphyla and classes of arthropods, see the *Macropædia* articles CRUSTACEANS; ARACHNIDS; and INSECTS; see also the *Micropædia* article MYRIAPOD.

For coverage of related topics in the *Macropædia* and *Micropædia,* see the *Propædia,* section 313.

The article is divided into the following sections:

GENERAL FEATURES

Size range. Representative groups of arthropods are shown in Figure 1. Most arthropods are small animals. Only aquatic forms are able to attain substantial sizes, because their bodies are supported in part by the surrounding water. The extinct chelicerate Eurypterida, for example, reached a length of 1.8 metres (5.9 feet), and some modern spider crabs may weigh up to 6.4 kilograms (14 pounds) and span 3.8 metres. Terrestrial arthropods do not grow very large. The largest insects and spiders do not weigh more than 100 grams (0.22 pound). The beetle *Goliathus regius* measures 15 centimetres (5.9 inches) in length and 10 centimetres in width, while the butterfly *Ornithoptera victoriae* of the Solomon Islands has a wing span exceeding 30 centimetres. One of the longest insects is the phasmid (walkingstick) *Pharnacia serratipes,* which reaches a length of 33 centimetres. The smallest arthropods include some parasitic wasps, beetles of the family Ptiliidae, and mites that are less than 0.25 millimetre (0.01 inch) in length, despite their complex structures.

Terrestrial arthropods

Distribution and abundance. Arthropods are found in almost all of the habitats that cover the Earth's surface. Many crustaceans live in the sea at depths exceeding 4,000 metres, while the insect collembolans and jumping spiders have been found on Mount Everest at heights exceeding 6,700 metres. Collembolans and the oribatid mites are among the permanent inhabitants of Antarctica. Brine shrimp are found in some saltwater lakes, and beetles, mites, and various crustaceans have been taken from hot springs. Minute crustaceans inhabit underground waters in many parts of the world, and deserts support a large arthropod fauna, especially insects and arachnids.

The numbers and diversity of arthropod insect pests are enormous. A bag filled with leaf mold from a forest floor, for example, will contain hundreds of arthropods, including mites, spiders, false scorpions, myriapods, a great variety of insects, and crustacean pill bugs. In the spring a temporary pool often teems with minute crustaceans. Planktonic copepods, which range in length from less than 0.5 millimetre to at least 10 millimetres, can reach densities of 30,000 individuals per cubic metre in surface waters of productive oceans, as around the Antarctic.

Insect pests

Importance. Arthropods are of great direct and indirect importance to humans. The larger crustaceans—shrimps, lobsters, and crabs—are used as food throughout the world. Small planktonic crustaceans, such as copepods, water fleas, and krill, are a major link in the food chain between the photosynthetic phytoplankton and the larger carnivores, such as many fish and whales. Although many species of insects and mites attack food crops and timber,

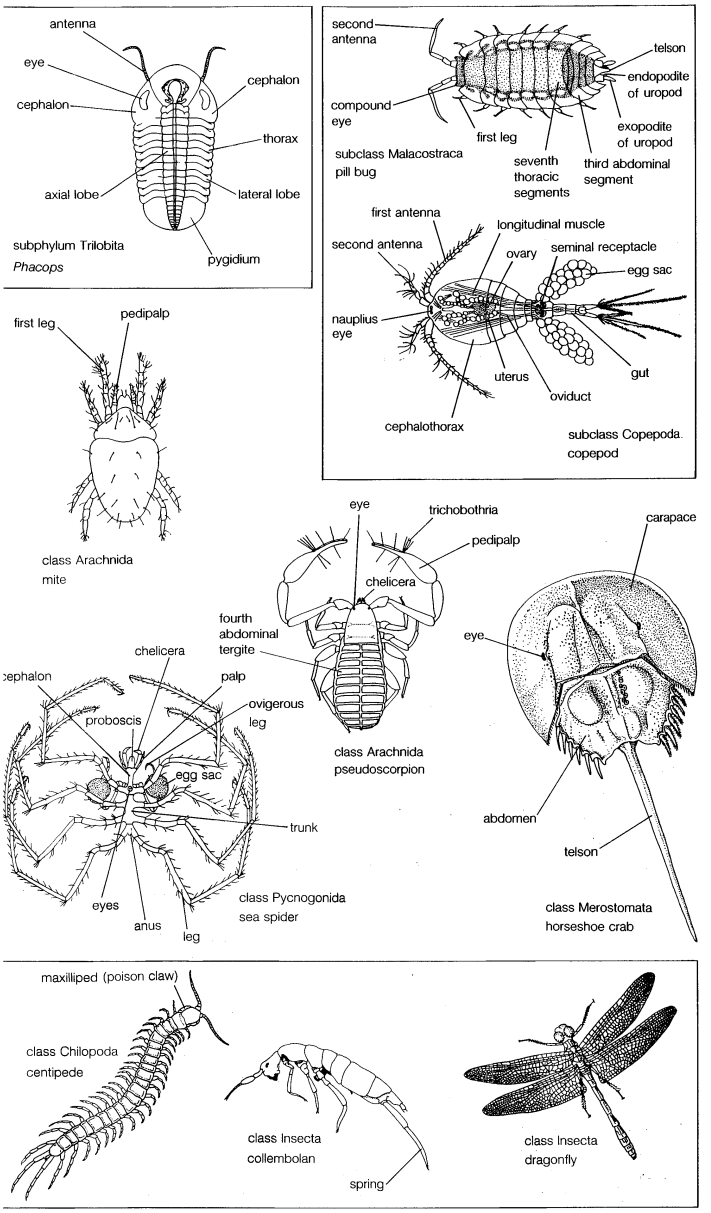


Figure 1: Representative arthropod groups.

From *Invertebrate Zoology*, 5th ed. (Figures 12-13A, 14-81B, 14-13A, 13-50A, 13-56, 13-41A, 13-1, 15-1A, 16-16C, and 16-17B (left) by R.D. Barnes, copyright © 1987 by Saunders College Publishing, a division of Holt, Rinehart and Winston, Inc., reprinted by permission of the publisher, after (*Phacops*) W. Sturmer and J. Berstrom, "New Discoveries on Trilobites by X-Rays," *Palaeontologische Zeitschrift* (1973), vol. 47½, Schweizerbartsche Verlagsbuchhandlung; (pill bug) F.C. Paulmier in W.G. Van Name, "The American Land and Freshwater Isopod Crustaceans," *American Museum Bulletin* (1936), American Museum of Natural History, New York City; (copepod) Matthes in A. Kaestner, *Invertebrate Zoology*, vol. 3, Crustacea, copyright © 1970 John Wiley & Sons, Inc., reprinted by permission of John Wiley & Sons, Inc.; (mite) E.W. Baker and G.W. Wharton, *An Introduction to Acarology* (1952), Macmillan Publishing Co.; (sea spider) G.O. Sars and (horseshoe crab) J. Van der Hoeven from L. Fage, "Classe des Merostomaces," in P. Grasse (ed.), *Traite de Zoologie*, vol. 6 (1949), Masson et Cie, Paris; (pseudoscorpion) M. Beier in P. Weygoldt, *Biology of Pseudoscorpions* (1969), Harvard University Press, Cambridge, Mass.; (centipede) R.E. Snodgrass; (collembolan) Willem; (dragonfly) Kennedy in H.H. Ross, *A Textbook of Entomology*, 3rd ed., copyright © 1965 John Wiley & Sons, Inc., reprinted by permission of John Wiley & Sons, Inc.

arthropods are of enormous benefit to human agriculture. Approximately two-thirds of all flowering plants (angiosperms) are pollinated by insects, and soil and leaf-mold arthropods, which include insects, mites, myriapods, and some crustaceans (pill bugs), play an important role in the formation of humus from decomposed leaf litter and wood.

The stings and bites of arthropods may be irritating or painful, but very few inject dangerous toxins. Medically, arthropods are more significant as carriers of diseases such as malaria, yellow fever, dengue fever, and elephantiasis (via mosquitos), African sleeping sickness (via tsetse flies), typhus fever (via lice), bubonic plague (via fleas), and Rocky Mountain spotted fever and Lyme disease (via ticks). Many diseases of domesticated animals are also transmitted by arthropods.

FORM AND FUNCTION

The exoskeleton and molting. The success of arthropods derives in large part from the evolution of their unique, nonliving, organic exoskeleton (Figure 2), which not only functions in support but also provides protection and, with the muscle system, contributes to locomotion. The exoskeleton is composed of a thin, outer protein layer, the epicuticle, and a thick, inner, chitin–protein layer, the procuticle. In most terrestrial arthropods, such as insects and spiders, the epicuticle contains waxes that aid in reducing evaporative water loss. The procuticle consists of an outer exocuticle and an inner endocuticle. In the exocuticle there is cross-bonding of the chitin–protein chains (tanning), which provides additional strength to the skeletal material. The hardness of various parts of the exoskeleton in different arthropods is related to the thickness and degree of tanning of the exocuticle. In crustaceans, additional rigidity is achieved by having the exoskeleton impregnated with varying amounts of calcium carbonate.

Epicuticle and procuticle

The formation of an exoskeleton required the simultaneous solution of two functional problems in the evolution of arthropods: If the animal is encased in a rigid covering, how can it grow and how can it move? The problem of growth is solved in arthropods by molting, or ecdysis, the periodic shedding of the old exoskeleton (Figure 3). The underlying cells release enzymes that digest the base of the old exoskeleton (much of the endocuticle) and then secrete a new exoskeleton beneath the old one. At the time of actual shedding, the old skeleton splits along specific lines characteristic of the group, and the animal pulls out of the old skeleton as from a suit of clothes. The old skeleton is usually abandoned but in some species is eaten. The new exoskeleton, which is soft and flexible, is then stretched by localized, elevated blood pressure augmented by the intake of water or air. Hardening occurs by stretching and especially by tanning within a number of hours of molting. In crustaceans, calcium carbonate is deposited into the new procuticle. (Soft-shell crabs are simply newly molted crabs.) Additional endocuticle may be added to the exoskeleton for some days or weeks following molting.

Molting is under hormonal control, and there is a long preparatory phase that precedes the process. The steroid hormone ecdysone, secreted by specific endocrine centres and circulated in the blood, is the direct initiator of molting. The timing, however, is controlled by other hormones and commonly by environmental factors. The interval between molts is called an instar. Because of the frequency of molts, instars are short early in life but grow longer with increasing age. Some arthropods, such as most spiders and insects, stop molting when they reach sexual maturity; others, like lobsters and crabs, molt throughout their lives. Most of the larger spiders of temperate regions, for example, molt about 10 times before reaching sexual maturity. As a result of molting, the length and volume of an arthropod display steplike increases over the life span, but internal tissue growth is continual as in other animals.

Instars

Loss of a limb is a common hazard in the life of many arthropods. Indeed, some arthropods, such as crabs, are capable of amputating an appendage if it is seized by a predator. The limb is then regenerated from a small, nipplelike rudiment formed at the site of the lost limb. The new limb develops beneath the old exoskeleton during the premolt period and then appears when the animal molts.

Muscles, appendages, and locomotion. The problem that a rigid external covering imposes on movement has been solved by having the exoskeleton divided into plates over the body and through a series of cylinders around the appendages. At the junction, or joints, between the plates and cylinders the exoskeleton is thin and flexible because it lacks the exocuticle and because it is folded. The folds provide additional surface area as the joints are bent. The arthropod's exoskeleton is therefore somewhat analogous to the armour encasing a medieval knight.

Most arthropods move by means of their segmental appendages, and the exoskeleton and the muscles, which attach to the inside of the skeleton, act together as a lever system, as is also true in vertebrates. The external skeleton of arthropods is a highly efficient system for small animals. The exoskeleton provides a large surface area for

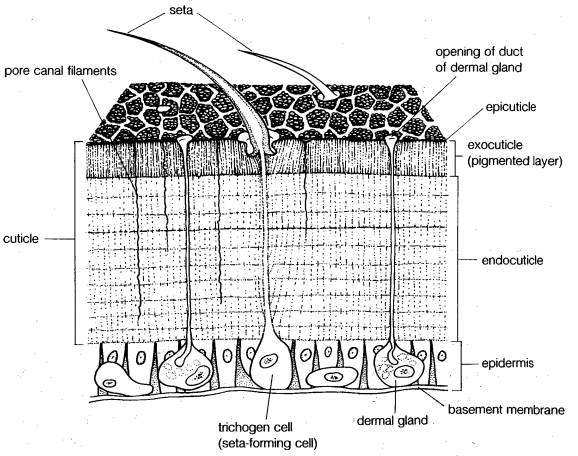


Figure 2: Diagrammatic section through the arthropod integument.

From *Invertebrate Zoology*, 5th ed. (Figure 12-3) by R.D. Barnes, copyright © 1987 by Saunders College Publishing, a division of Holt, Rinehart and Winston, Inc., reprinted by permission of the publisher, after R.H. Hackman in M. Florkin and B.T. Scheer (eds.), *Chemical Zoology*, vol. 6 (1971), Academic Press

the attachment of muscles and, in addition to functioning in support and movement, also provides protection from the external environment. The cylindrical design resists bending, and only a relatively small amount of skeletal material need be invested in thickness to prevent buckling. The external skeleton imposes limits on the maximum size of an arthropod, especially in those that live on land. The largest arthropods live in the sea, where they gain considerable support from the buoyance of seawater. On land, an excessive amount of skeleton would be required to support a large bulk and, in addition, the new soft skeleton might collapse following a molt.

Appendages

Appendages of arthropods have been adapted for all types of locomotion—walking, pushing, running, swimming, and burrowing. In most arthropods the legs move alternately on the two sides of the body; *i.e.,* when one leg is in a power stroke, its mate on the opposite side of the body is in the recovery stroke (the same is true of mammals when walking). The legs in front or back are a little ahead or behind in the movement sequence. Because of the lateral position of the legs, the body of an arthropod tends to hang between them. Leg interference and trunk wobble tend to be problems in an animal with a long trunk and many legs, such as a millipede or centipede. Most arthropods have evolved more compact bodies and a smaller number of legs. The number of pairs of legs used in walking is not more than seven (crustacean pill bugs), four or five (shrimps and crabs), four (arachnids), and three (insects). This reduces the problem of mechanical interference. When a ghost crab, for example, is running rapidly across a beach or dune, only the second, third, and fourth pairs of the five pairs of legs (counting the claws) are employed. Leg interference is further reduced in most arthropods by varying limb length and placement. For example, in *Scutigera,* the centipede commonly seen in houses, the legs increase in length from front to back and thus pass over or under one another in stepping. The tendency for the trunk to wobble has been reduced in some centipedes by having overlapping dorsal plates and in millipedes by having pairs of segments fused to form double segments. Many arthropods are capable of walking on vertical surfaces. Some simply grip minute surface irregularities with the claws at the end of the legs. Others, such as certain spiders and flies, have an array of specialized gripping hairs at the ends of the legs.

Insect wings

Insect wings are not segmental appendages as are the legs. The paired wings arise as lateral folds of the integument, one pair above each of the last two pairs of legs. Each wing thus consists of an upper and lower sheet of exoskeleton closely applied to each other. The two skeletal sheets are separated at various places, forming tubular supporting veins. Unlike the wings of an airplane, the wings of insects are flat plates, and lift is obtained by changing the angle at which the front margin of the wing meets the oncoming air stream. The evolution of flight is one of several adaptations that have enabled insects to become the most diverse and populous group of terrestrial animals.

A burrowing habit has evolved in some insects, such as mole crickets and ants, but the largest burrowers are crustaceans. Mole crabs and box crabs are rapid burrowers in soft marine sands, and various species of mantis shrimps, mud shrimps, and snapping shrimps create elaborate burrows below the bottom surface. Crustaceans also include the largest number of arthropod tube dwellers, surpassed only by certain marine worms (polychaetes). Most of the tube-dwelling crustaceans are amphipods. Their tubes are usually composed of sand or mud particles secreted together and attached to bottom objects; there are, however, some amphipods that carry their tubes with them like a portable house.

Digestive system and feeding. Arthropods exhibit every type of feeding mode. They include carnivores, herbivores, detritus feeders, filter feeders, and parasites, and there are specializations within these major categories. Typically, paired appendages around the mouth are used for collecting and handling food and are usually specialized in accordance with the particular diet of the animal. For example, the insect family Aphididae has mouthparts adapted for piercing vegetation and sucking out plant juices. The crustacean fiddler crabs, which emerge from burrows on sand flats at low tide, scoop up the surface sand with their small claws (only one in the male) and place the sand within their mouthparts, where it is sifted with fine hairs. The organic material is consumed, and the mineral material is ejected as a small "spitball." Where there is a large population of crabs, ejected material may cover the surface of a flat by the end of the low-tide period. The crustacean mole crabs, or sand crabs, of surf beaches use their antennae to filter plankton from the receding waves after reburying themselves. Planktonic crustacean copepods only a few millimetres long can collect up to several hundred thousand diatoms every 24 hours with certain appendages (maxillae) near the mouth. A number of carnivorous arthropods, notably spiders, pseudoscorpions, and centipedes, capture prey with poison, which is usually delivered with a pair of appendages; scorpions use a single stinger at the tip of the tail. In spiders, the poison is introduced through a pair of fangs (chelicerae) flanking the mouth, and in centipedes the poison claws lie beneath the head. Few of these species have a venom that is fatal to humans (see the *Micropædia:* MYRIAPOD).

The front and back parts of the digestive tract (foregut and hindgut) are lined with the same skeletal material that is found on the outside of the body and that is molted with the rest of the skeleton. Only the relatively small middle section (midgut) lacks a chitinous lining. The digestive tract varies greatly in structure, depending upon the diet

From *Invertebrate Zoology*, 5th ed. (Figure 12-4A–D) by R.D. Barnes, copyright © 1987 by Saunders College Publishing, a division of Holt, Rinehart and Winston, Inc., reprinted by permission of the publisher, modified from M. Locke in J. Bereiter-Hahn, A.G. Matoltsy, and K.S. Richards (eds.), *Biology of the Integument* (1984), Springer-Verlag, Berlin

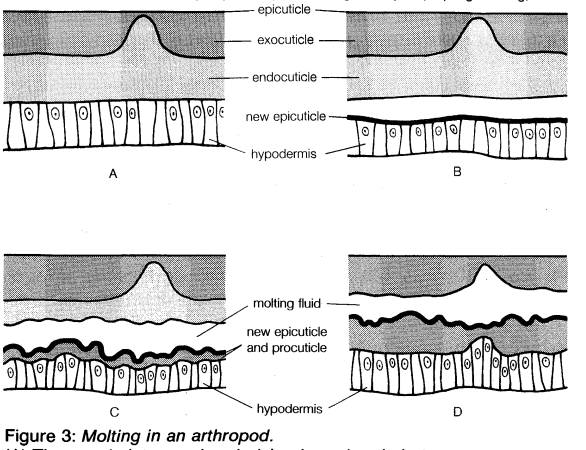


Figure 3: *Molting in an arthropod.*
(A) The exoskeleton and underlying hypodermis between molts. (B) The hypodermis separates, and the molting fluid and a new epicuticle are secreted. (C) The old cuticle is digested and a new procuticle is secreted. (D) Just before molting, the animal is encased in new and old skeleton.

and feeding mode of the animal. In general, however, the midgut region is the principal site of enzyme production and absorption of digested food. The enzymes may pass forward into the front part of the gut and even outside into the body of the prey, in the case of spiders.

Respiratory system. Aquatic arthropods (crustaceans and the chelicerate horseshoe crabs) possess gills for respiration. Although they vary in structure and location, the gills are always outgrowths of the integument (skin) and are therefore covered by the exoskeleton, which is thin in this area and not a barrier to the exchange of gases. Terrestrial arthropods possess tracheae and book lungs as respiratory organs. Tracheae are a system of tiny tubes that permit passage of gases into the interior of the body. In some arthropods the tracheal tubes are bathed by blood, but in insects the minute terminal endings (tracheoles) are embedded in the tissues, even within muscle cells. The tracheal tubes (but not the tracheoles) are molted along with the rest of the exoskeleton. Tracheae are a unique arthropod invention and undoubtedly evolved numerous times in the phylum, for they are found in myriapods, insects, and arachnids. Tracheal systems are highly efficient for these small, terrestrial animals. The small, external openings (spiracles) reduce water loss, the chitinous lining prevents collapse, and the small size of the arthropod and consequent short length of the tubule eliminates the need for moving gases in and out by active ventilation (diffusion usually being sufficient). Book lungs are chitin-lined internal pockets containing many blood-filled plates over which air circulates. Most spiders possess tracheae and book lungs, but large spiders (such as tarantulas) and scorpions possess book lungs alone.

Tracheae

Circulatory system. Arthropods possess an open circulatory system consisting of a dorsal heart and a system of arteries that may be very limited (as in insects) or extensive (as in crabs). The arteries deliver blood into tissue spaces (hemocoels), from which it eventually drains back to a large pericardial sinus surrounding the heart. A varying number of paired openings (ostia) are located along the length of the heart and permit blood to flow in when the valves are open. When the heart is contracting, closed valves prohibit the blood from flowing back and force it into the arteries of the tissues, from which it flows to other hemocoels. In the larger crustaceans, the blood then passes through the gills (where it becomes oxygenated) on its return to the heart. The blood of large arachnids and crustaceans contains the blue, oxygen-carrying pigment hemocyanin; insects lack a respiratory pigment since the tracheal system delivers oxygen directly to the tissues. A few insect larvae and some small crustaceans have blood containing hemoglobin.

Hemocoels

Excretory system and water balance. Crustaceans and arachnids possess paired excretory organs (maxillary, antennal, or coxal glands) that open at the bases of certain appendages. Myriapods, insects, and some arachnids, such as spiders and mites, possess another type of excretory organ, Malpighian tubules, which open into the intestine. Thus in these animals both excretory and digestive wastes exit from the anus.

Water loss through evaporation is a major problem for animals that live on land, especially small ones like arthropods, and an array of defenses against desiccation have evolved. Both arachnids and insects possess waxy compounds in the epicuticle, the outer layer of the exoskeleton, which greatly reduce evaporative water loss. Arthropods that lack a waxy epicuticle, such as the pill bugs, and very small arthropods, such as mites, pseudoscorpions, and collembolans, live in leaf mold and soil, beneath logs, under stones, and in other areas where the danger of desiccation is reduced. The waxes in the epicuticle not only reduce water loss but can also act as a water repellent, reducing the danger of submersion in droplets of rain or dew. This resistance to wetting enables aquatic insects, such as beetles, to carry below the surface a film of air, which can then be used in respiration. It also contributes to the ability of water striders to move over the surface of water without breaking through the surface film.

Both insects and spiders eliminate their nitrogenous wastes as compounds insoluble in water (uric acid, guanine), thereby not requiring that water be excreted. Insects share with birds and mammals the ability to produce a urine that is saltier than the blood, which is of great value in conserving water because it permits the production of a concentrated urine.

Nervous system and organs of sensation. The arthropod nervous system consists of a dorsal brain and a ventral, ganglionated longitudinal nerve cord (primitively paired) from which lateral nerves extend in each segment. The system is similar to that of annelid worms, from which arthropods may have evolved. The neuromuscular organization of arthropods is quite different from that of vertebrates, in which one neuron supplies a number of muscle cells, together forming a functional motor unit. The small size of the muscles prohibits such an organization in arthropods. Instead, the state of contraction of an arthropod muscle is determined by which of several different types of neurons supplying one muscle cell are fired.

The sense organs (sensilla) on the body surface involve some specialization of the exoskeleton barrier. The sensory nerve endings are lodged in cuticular hairs (setae), peglike projections, cones, pits, or slits, which may occur in large numbers on antennae, mouthparts, joints, and leg tips. Changes in the tension of the surrounding cuticle stimulate the nerve endings. For example, the legs of spiders and scorpions possess slits in the exoskeleton that are covered by a thin membrane to which a neuronal receptor is attached below. Tension changes in the exoskeleton cause slight movements in the cuticular membrane and stimulate the receptors. Slits of varying length may be grouped together like the strings of a harp. Slit sense organs enable spiders to detect web vibrations produced by trapped insects, and they permit scorpions to detect ground vibrations produced by approaching prey. Chemoreceptive sensilla (taste and smell) have holes in the cuticle permitting the chemical substances being monitored to enter.

Most arthropods possess eyes, but in most species they function only to detect the intensity of light and the direction of the light source. The ability to detect objects is more restricted. Among arthropods the greatest visual acuity is found in the predaceous mantis shrimp, some crabs, and many insects, all of which possess compound eyes. Compound eyes are effective in detecting motion. The eight eyes of spiders are not of the compound type, but in the case of the cursorial (hunting) wolf spiders and jumping spiders they are effective in locating and tracking prey.

Eyes

Reproductive system and life cycle. With few exceptions, the sexes are separate in arthropods; *i.e.,* there are both male and female individuals. The paired sex organs, or gonads, of each sex are connected directly to ducts that open onto the ventral surface of the trunk, the precise location depending upon the arthropod group.

In arthropods, sperm are commonly transferred to the female within sealed packets known as spermatophores. In this method of transfer the sperm are not diluted by the surrounding medium, in the case of aquatic forms, nor do they suffer from rapid desiccation on land. Among some arachnids, such as scorpions, pseudoscorpions, and some mites, the stalked spermatophore is deposited on the ground. Either the female is attracted to the spermatophore chemically or the deposition of the spermatophore occurs during the course of a nuptial dance, and the male afterward maneuvers the female into a position in which she can take up the spermatophore within her genital opening. Centipedes also utilize spermatophores with an accompanying courtship behaviour. Among insects there are some primitive wingless groups, such as collembolans and thysanurans, in which the spermatophore is deposited on the ground, but in most insects the spermatophores are placed directly into the female genital opening by the male during copulation. Many arthropods transfer free sperm rather than spermatophores. These include many crustaceans, millipedes, some insects (such as dipterans and hemipterans), spiders, and some mites.

Spermatophores

Arthropod eggs are usually rich in yolk, but in all groups there are species whose eggs have little yolk. Some specialized methods of reproduction found among certain arthropods include the development of unfertilized eggs

(parthenogenesis), the birth of living young (viviparity), and the formation of several embryos from a single fertilized egg (polyembryony).

The eggs of many crustaceans hatch into larvae which have fewer segments than the adult. The earliest larval hatching stage is a minute nauplius larva, which possesses only the first three pairs of appendages. Additional segments and appendages then appear at regular intervals with molting. There are several advantages of larval stages in the development of aquatic animals: Currents disperse the larvae, enabling some to settle in different locations from the parents; because many larvae are capable of feeding, less yolk is required in the egg; and, moreover, planktonic larvae do not compete with benthic adults.

In most chelicerates and insects, almost all of the segments are present at hatching, although in insects the body form may differ from that of the adult. Primitive insects, such as collembolans, have the adult form on hatching. Many insects, such as grasshoppers, crickets, and true bugs, hatch as nymphs, which superficially resemble the adult but lack wings. They gradually acquire these adult features during the nymphal instars. Other insects, such as beetles, butterflies, moths, flies, and wasps, hatch as larvae (grubs, caterpillars, maggots) that differ markedly from the adult. The larvae inhabit different environments and eat different foods than the adults. In these insects a pupal stage with metamorphosis bridges the gap between the larva and the adult form.

Nymphs

Myriapods have the general body form of the adult on hatching though they may lack some of the segments. Most millipedes hatch with only seven trunk segments. Some centipedes hatch with all of the adult trunk segments, but others have fewer than the adult.

The young of most arachnids are similar to the adult. The female scorpion gives birth to her young, which immediately climb onto her back. Female wolf spiders also carry their young, and prior to hatching they carry the white egg case attached to the posterior spinnerets. Unlike other arachnids, mites and ticks hatch as six-legged larvae, which acquire the fourth pair of legs at a later molt.

EVOLUTION AND PALEONTOLOGY

The arthropods share many features with the phylum Annelida. Both are segmented, and members of the annelid class Polychaeta have a pair of appendages on each segment. The plan of the nervous system in arthropods is very similar to that of annelids, and the basic plan in both groups shows a tubular, dorsal heart, which is then lost or modified in some. Annelids possess a coelom, which in arthropods is present only in the embryo. Its absence is probably related to the evolution of the exoskeleton and to the change in the mode of locomotion.

The first fossil arthropods appear in the Cambrian Period (570,000,000 to 500,000,000 years ago) and are represented by trilobites, merostomes, and crustaceans. Also present are some enigmatic arthropods that do not fit into any of the existing subphyla. The earliest terrestrial arachnid is from the Devonian Period (395,000,000 to 345,000,000 years ago), but it does not belong to any living order. Though a myriapod-like fossil has been found from the Devonian Period, it is not until the Carboniferous Period (345,000,000 to 280,000,000 years ago) that there is a good record of centipedes, millipedes, and insects.

Most zoologists recognize the trilobites, chelicerates, crustaceans, and uniramians as four major lines of arthropod evolution, but there is little agreement as to how these lines are related to one another or, indeed, if they had independent evolutionary origins from the annelids. The Uniramia takes its name from the proposition that the appendages of these arthropods are primitively unbranched. This idea is not accepted by all zoologists, especially some entomologists.

CLASSIFICATION

Distinguishing taxonomic features. Modification, specialization, number, and appearance of body segments and appendages (especially anterior ones such as antennae and mouthparts) are important criteria in distinguishing arthropod classes. Other structural features of taxonomic importance include location of the gonopores, structure of the head, and adaptations of the respiratory and excretory systems. In the classification below, the group marked with a dagger (†) is wholly extinct and known only from fossils.

Annotated classification.

PHYLUM ARTHROPODA

Bilaterally symmetrical invertebrates with jointed exoskeleton covering body and appendages; cilia absent; body segmented, though segmentation commonly reduced as a result of fusion; appendages typically specialized for different functions; coelom greatly reduced; nervous system consists of dorsal brain and a double or single (fused) ventral nerve cord; eggs typically rich in yolk; development highly modified.

†Subphylum Trilobita (trilobites)

Extinct; head (or cephalon) composed of 5 segments bearing a pair of antennae and compound eyes; oval, flattened body composed of cephalon, thorax, and pygidium, each segmented; dorsal surface molded longitudinally into 3 lobes; each segment bears a pair of similar, branched appendages; marine; Cambrian Period to the end of the Paleozoic Era; more than 4,000 fossil species known.

Subphylum Chelicerata

Body divided into prosoma (cephalothorax) and opisthosoma (abdomen); no antennae; first pair of appendages consists of chelicerae flanking the mouth; in most chelicerates the other prosomal appendages are a pair of pedipalps and four pairs of legs.

Class Merostomata

Large marine chelicerates with book gills on the underside of the opisthosoma; prosoma covered by a dorsal carapace; opisthosoma bears a long terminal spine; 2 orders, Xiphosura (horseshoe crabs, 4 species) and Eurypterida (Gigantostraca), which is extinct and includes 200 fossil species from the Paleozoic Era.

Class Arachnida (scorpions, spiders, ticks, mites)

Chiefly terrestrial; book lungs and/or tracheae as gas exchange organs; opisthosoma (abdomen) segmented or unsegmented externally and broadly or narrowly joined to the prosoma; prosomal appendages consist of 1 pair of chelicerae, 1 pair of pedipalps, and 4 pairs of legs; gonopore always on the lower side of second abdominal segment; about 70,750 species; 0.25 mm–18 cm.

Class Pycnogonida (sea spiders)

Marine; narrow trunk of 4 to 6 segments; greatly reduced abdomen; cephalon (head) with proboscis bearing a pair of chelicerae, palpi, and egg-carrying legs; usually 4 pairs of walking legs attached to lateral projections of the trunk; tubercle with 4 eyes located dorsally between the first pair of legs; no gas respiratory organs; commonly found crawling over sessile animals, such as hydroids and bryozoans; about 1,000 described species; 1 mm–10 cm.

Subphylum Crustacea

Chiefly aquatic; head bearing 2 pairs of antennae, a pair of mandibles, and 2 pairs of maxillae; trunk highly variable but commonly covered in part or entirely by a posteriorly directed fold of the head (carapace); paired appendages biramous, often with 1 branch lost; 2 stalked or stalkless compound eyes present in most; when present, gas exchange organs are gills; mostly marine, but many freshwater species; some isopods terrestrial; 44,000 described species in up to 10 classes.

Suphylum Uniramia

Chiefly terrestrial; segmental appendages primitively unbranched; head appendages comprise a pair of antennae, a pair of mandibles, and 1 or 2 pairs of maxillae; trunk and appendages variable; respiratory organs are tracheae.

Class Chilopoda (centipedes)

Elongate; many trunk segments, each with 1 pair of legs; 2 pairs of maxillae covered by a large pair of poison claws representing the first pair of trunk appendages; eyes, if present, are simple ocelli; gonopore on last segment; 5 mm to almost 30 cm; about 2,500 living species.

Class Symphyla

Mouthparts consist of a pair of mandibles and 2 pairs of maxillae; 12 leg-bearing trunk segments; terminal segment carries a pair of spinnerets; gonopore on fourth segment; 1–8 mm; about 160 living species.

Class Diplopoda (millipedes)

Elongate; trunk containing many diplosegments, each bearing 2 pairs of legs and spiracles; single pair of maxillae fused to form a flattened plate (gnathochilarium); first 4 trunk segments not diplosegments, and third bears the gonopores; simple eyes (ocelli) present or absent; 2 mm–28 cm; about 10,000 living species.

Class Pauropoda

Antennae branched; a pair of maxillae; 9–11 trunk segments bearing legs; gonopores on third trunk segment as in diplopods; 0.5–1.5 mm; about 500 described species.

Class Insecta

Body composed of a head, thorax, and abdomen; head bears simple eyes and usually a pair of lateral compound eyes; 2 pairs of maxillae, the second pair fused (labium); thorax of 3 segments, each with a pair of legs, and the second and third usually bearing wings; abdomen of 11 segments without appendages in the adult; gonopore at end of abdomen; 0.25 mm–33 cm; about 751,000 described species.

Critical appraisal. The classification adopted above is the most widely accepted system and the one followed in *Synopsis and Classification of Living Organisms* edited by Sybil P. Parker. An older and frequently encountered system unites all of the mandibulate arthropods (crustaceans and uniramians) within one subphylum, the Mandibulata, but since crustaceans and uniramians are now believed to constitute independent lines of evolution, the Mandibulata is an artificial taxon. If the four major lines of arthropod evolution—trilobites, chelicerates, crustaceans, and uniramians—evolved independently from annelidan ancestors, then each should constitute a separate phylum and the Arthropoda, a superphylum.

Most zoologists agree that the myriapod classes, which contain the uniramians with long trunks, are related to insects, but the precise interrelationships of these five classes are still unknown. Members of the phylum Onychophora, caterpillar-like animals of the tropics and the Southern Hemisphere, are believed to be near the ancestors of the uniramians, and they are sometimes included within the Arthropoda. Onychophorans, however, lack a well-developed exoskeleton, the most basic feature of the phylum.

Members of the phylum Tardigrada, called water bears, are microscopic animals living in algae, soil, and moss. Although they have sometimes been included within the Arthropoda, the similarity to arthropods is probably superficial, and they are relegated to a separate phylum by most zoologists. Another group sometimes called an arthropod subphylum or class is the Pentastomida (Linguatulida), wormlike parasites that live in the lungs, gut, and coelom of reptiles, birds, and mammals. The body is ringed, and there is no chitinous exoskeleton; on the side of the head are four hooks, but the mouth region lacks jaws and other appendages. These peculiarities are undoubtedly related to their endoparasitic habit; if they are aberrant arthropods, their relationship to the other classes is uncertain.

BIBLIOGRAPHY. Introductory texts include ROBERT D. BARNES, *Invertebrate Zoology,* 5th ed. (1987); VICKI PEARSE *et al., Living Invertebrates* (1987); RICHARD S. BOARDMAN, ALAN H. CHEETHAM, and ALBERT J. ROWELL (eds.), *Fossil Invertebrates* (1987); ALFRED KAESTNER, *Invertebrate Zoology,* vol. 2 and 3, trans. from German and adapted by HERBERT W. and LORNA R. LEVI (1968–70); "Arthropoda," in SYBIL P. PARKER (ed.), *Synopsis and Classification of Living Organisms,* vol. 2 (1982), pp. 71–728, taxonomic classifications; and FRIEDRICH SCHALLER, *Soil Animals* (1968; originally published in German, 1962). Advanced treatments of arthropod characteristics include KENNETH U. CLARKE, *The Biology of the Arthropoda* (1973), emphasizing the unity of functions among arthropods; CLYDE F. HERREID II and CHARLES R. FOURTNER (eds.), *Locomotion and Energetics in Arthropods* (1981); ANTHONY C. NEVILLE, *Biology of the Arthropod Cuticle* (1975), on the phylogeny, physical properties, and physiology of cuticle; and J. BEREITER-HAHN, A.G. MATOLTSKY, and K. SYLVIA RICHARDS (eds.), *Biology of the Integument* (1984), vol. 1, *Invertebrates,* with several chapters on the arthropod integument. Opposing viewpoints on the evolution of arthropods are presented by H. BRUCE BOUDREAUX, *Arthropod Phylogeny with Special Reference to Insects* (1979), favouring a monophyletic or single-ancestor origin; and S.M. MANTON, *The Arthropoda: Habits, Functional Morphology, and Evolution* (1977), taking a polyphyletic origin.

(R.D.Ba.)

Classification of the Arts

As long as the name arts is given to a realm so vast and indefinite as to embrace literature, music and dance, theatre and film, the visual and decorative arts, and other equally diverse activities, their classification—the ways in which each is regarded as being either unique or like others—will remain a controversial but necessary undertaking. Classification is a useful approach to the organization of knowledge in any field: the classification of plants and animals in the 18th century led to the discovery of evolution in the 19th. In the arts, classification can be of immense help in understanding the interrelations between the arts and in drawing attention to characteristics of each that might otherwise go unnoticed. Whether it is consciously devised or largely unconscious, some sort of classification is implicit in any serious exploration of the arts. Related to it is an appraisal of the importance and value of each of the arts. Admirers of a particular art often feel that it is unique and deny that it is like any other. Logically, however, anything correctly described as a work of art, whether a poem, a picture, or a sonata, belongs to that class and hence resembles other members of the class to some extent. Some works or types of art are more nearly unique than others in resembling fewer others or in resembling others in fewer respects. On the other hand, no type or example is exactly like any other, but the differences may be negligible for all practical purposes.

There is no one correct way of making a classification. The aims and interests of the person making the classification, and his philosophical orientation, are usually significant factors. The emphasis may be placed on sensory qualities, for example, or on moral and religious ideals. Regardless what classification is used, however, as it is subdivided into smaller groups, detailed similarities and differences are brought out in the emotional, intellectual, and social features of the works or art involved.

The classification of the arts is closely related to subjects that receive more extensive treatment in the articles AESTHETICS; PHILOSOPHIES OF THE BRANCHES OF KNOWLEDGE; and ARTS, CRITICISM OF THE. Discussions of specific classification systems are in ARTS, STYLE IN THE; FOLK ARTS; and ARTS, PRACTICE AND PROFESSION OF THE. The relationship between the arts within a given culture is discussed in SOUTH ASIAN ARTS and comparable articles on the arts of other peoples.

The article that follows will attempt only to draw the reader's attention to the importance of the subject of classification in the study of the arts and to offer him means of applying thoughtful scrutiny to a process that has too often remained subconscious.

It is divided into the following sections:

BASES OF CLASSIFICATION

To classify things is to arrange them in groups or sequences according to a plan, especially on the basis of some characteristic that they are thought to have in common. Thus, the names in a directory may be arranged according to the letter of the alphabet with which each last name begins. A biologist classifies the kinds of animal or

plant within a certain area. A museum of art may place its sculptures in one gallery, its paintings in another. It may also classify them according to historical period, such as Italian Renaissance or 18th-century French.

A base or basis of classification is a concept of what the members of a certain class are thought to have in common, a factor that can be used either for grouping them or for separating them from other classes. "Subject represented" is one such base: it may be used to separate "paintings" into three or more groups, such as portraits, landscapes, and narrative scenes. This involves classification in a narrow sense; that is, as grouping items or subgroups together under certain headings. It also involves division, which can be regarded as the opposite of classification, but the latter is usually made to cover both processes. A large, complex art such as literature can be divided into smaller ones, such as poetry and prose. Each of these also can be subdivided, as into epic, dramatic, and lyric poetry. A relatively complex, detailed division and classification of the arts from a philosophic point of view is called a system of the arts.

Taxonomy in the arts

The study and scientific practice of classifying phenomena in an extensive field is called taxonomy. In biology and other exact sciences, it is highly complex and consistent, with precise definitions. Thus a "genus" in biology comes between a "family" and a "species." In classifying arts, however, such terms are used more loosely. Thus, genus can mean, in aesthetics, any fairly large, inclusive group, capable of including subgroups and also of being subsumed under a still larger group. A species, in the broad sense, is one of the main divisions of a genus, and the traits or characteristics of a genus are generic traits. Differentiae are ways in which individuals or subgroups in a larger group are significantly unlike. Aristotle distinguishes between epics and tragedies in poetry in terms of differentiae.

To define or locate a particular example or subgroup in a system of classification, it must be ascertained, first, what genus (inclusive group) or genera it belongs to and then how it differs from others in that group or groups. A species is one of the main divisions of a genus, but it may also act as a genus (in the broader sense of that word) by including smaller groups.

Problems of definition. What are the arts, and what is art in general? This is still a controversial question after centuries of debate. No particular definition commands universal assent. Several meanings are still frequently used, of which the oldest is the broad, technical sense. In this sense, the English word art and its equivalents in Greek and Latin covered not only what are now called "fine arts" or "aesthetic arts" but any kind of transmitted, useful skill, such as agriculture, medicine, and war. This sense of "art" survives in such terms as Bachelor of Arts, a degree that is often awarded for a course of study that involves no aesthetic arts at all.

In the 18th century, the so-called beaux arts, the beautiful or high arts (also called "elegant" or "polite" arts), were distinguished from the merely useful arts on the ground that they were aimed at giving aesthetic pleasure to the beholder. In the 19th and 20th centuries, there has been a tendency to abandon the name art in speaking of the purely utilitarian skills and to call them instead "industries," "technics," "branches of engineering," or "applied sciences." Without the prefix fine, the word art alone is now commonly understood to mean the fine or aesthetic arts. To produce an experience of beauty or aesthetic satisfaction is said to be their distinguishing function or characteristic but not necessarily their only one. In this moderately broad, technical sense, some, but not all, architecture, furniture, and clothing can qualify as arts in spite of their useful purposes. They can be called useful arts, rather than merely useful skills, because they combine aesthetic and utilitarian forms and functions. On the other hand, such technics as coal mining and placing metal pipes underground do not qualify as arts at all, since they do not ordinarily involve aesthetic perception.

In psychology, anthropology, and other sciences, a particular product or performance does not have to be beautiful or aesthetically satisfying to qualify as art. The same can be said of much primitive art, children's art, and art produced by the insane. They may be classed as art if they belong to types of product or performance that have been socially recognized as having an aesthetic function. In this sense, any picture, clay figure, dance, or traditional song can be accepted as a work of art, whether beautiful or not. Being nonevaluative, this conception makes it possible for the scientist to study the arts as a field of cultural phenomena for investigation, without having to show in advance that they are pleasant, good, or beautiful.

In another sense of "art," called the expressionist theory, art has been defined as the expression and transmission of remembered emotion. This definition is not inconsistent with the technical meanings, but it puts an emphasis on the artist's procedure rather than on the effects and functions of the product.

In a third sense (an extremely narrow one), the concept of art is limited to painting and drawing alone, sometimes to the visual arts alone. This definition has the disadvantage of excluding music, poetry, dance, and many other arts that have long been recognized as "fine" or "aesthetic." It is confusing in that painting and the other visual, manual arts were themselves long excluded from the category of "liberal arts."

Decorative and industrial arts

The decorative arts are a species of visual art whose main function is to combine utility with beauty or aesthetic satisfaction. They tend to emphasize visual ornamentation and design along with fitness for some useful end or ends. Although Western painting and sculpture in the past traditionally tended to emphasize representation, the decorative arts used both abstract and representational design. Utility, design, and representation appear with varying degrees of emphasis in such arts as medieval book illumination, jewelry, Greek or Chinese vase painting, Persian rugs, and French rococo furniture. Some styles of decorative art are comparatively plain and simple in order to achieve an effect of visual design without superficial ornament.

Ironworking is usually classed as an industry not an art, because most of its activity is devoted to products that have no aesthetic intent. A small branch of the industry, however, may be devoted to making decorative designs in iron and steel, perhaps for use in architecture. This branch may be called a useful art or a decorative art; the two categories overlap. If old-fashioned hand methods are used there, the term handicraft is also applicable.

The term industrial art is now applied chiefly to arts in which machinery and mass production are commonly employed, with many specialized workers cooperating, as in motion pictures. It may be applied to the process of making large numbers of colour-print reproductions from paintings. Such art used for advertising, however, as in newspaper layouts and street posters, is called "commercial art." The industrial arts include many types of large-scale manufacture in which an aesthetic appeal is sought, as in the manufacture of books, magazines, refrigerators, typewriters, furniture, television sets, automobiles, airplanes, or appliances. The aesthetic factor in such work is sometimes called "styling."

Bases in status. Until the 18th century, the most common basis for classifying the arts was in terms of their social and psychological status. This system, which was devised by the ancient Greek philosophers, involved a wide separation between the so-called liberal and servile arts. In ancient Greece and Rome, the liberal arts were conceived as: (1) those that befitted a freeman and were thus comparatively noble, aristocratic, or genteel; (2) those requiring the exercise of superior mental ability, rather than mere hand labour, however skillful; and (3) those tending to elevate the minds of the artist and of his patrons, rather than merely providing material comforts, pleasures, and conveniences. The servile arts were those befitting only a person of a lower class, at least as far as the work itself was concerned. It was beneath the dignity of a lady or gentleman to work in them but not to use and enjoy such products. For the wealthy aristocrat, even to practice painting as a hobby exposed him to ridicule. Aristotle notes that even if the gentleman practices music for his own pleasure, he should not do it too well.

Liberal and servile arts

From the medieval period through the Renaissance, the

class distinction between liberal and servile was retained in modified form. The "seven liberal arts" of the Middle Ages were composed of the trivium, or literary group—consisting of grammar, dialectic, and rhetoric—and the quadrivium, or mathematical group—consisting of arithmetic, geometry, music, and astronomy. Music was theoretical in emphasis, with little attention to its sound and much to the numerical relations among the tones. In the 12th century, literature, including poetry, tales, and dramas, was classed as a mere supplement or aid to philosophy. Modern taste admires the visual arts of the Middle Ages, but many churchmen and philosophers of the time condemned them as sensuous, perhaps idolatrous. The artist in a material medium such as stone or metal was low in social status until well along in the Renaissance.

In and after the Renaissance, art was conceived as being more hedonistic, more devoted to providing aesthetic pleasure through the sensuous perception of beautiful forms. This conception partly replaced the emphasis on moral and intellectual qualities in the ancient Greek and Roman traditions. The manual arts of decoration and design gradually rose in prestige, along with the arts of music, poetry, ballet, and theatre. Visual artists in such media as architecture, landscape, design, and pageantry gained in respect and in financial rewards. High status in them depended more on success in pleasing a luxury-loving aristocracy than on moral and intellectual virtues. But some philosophers urged that art at its best combines both sets of values.

Elite and popular arts

A hint of social status persists today in the distinction between "elite" and "popular" art, which includes products of the "mass media" such as film, newspaper and magazine illustration, radio, and television. But this distinction refers more to the level of taste and education supposedly required to appreciate these arts, and to their different publics, than to social or financial status.

Certain arts have been grouped from time to time as "lively," suggesting animation and gaiety, in contrast with the supposed solemnity of classical, "highbrow" art. Such arts are, on the whole, suited to popular taste, though often appealing also to the scholarly. They often are easy to grasp and involve a simple narrative. The comic strip is static, but it suggests movement, often playfully exaggerated into mock violence. Some, but not all, lively arts involve directly presented motion, as in the film, which has rapidly ascended from a popular to a serious art. At its best, it can achieve all types of aesthetic value, including emotional expression, character, and plot, but it does not always try to avoid banality.

Major and minor arts

The antithesis between "major" and "minor" arts is partly dependent on the social status and cultural context of the arts concerned, though it is often mentioned as if certain arts were inherently and permanently greater than others. The ground of superiority of an art, according to some philosophers, is its greater ability to express thoughts and feelings of universal, lasting value. Poetry and other forms of literature have been regarded as superior in this respect to the decorative arts, which were associated with superficial, thoughtless luxuries. As the decorative and useful arts rose in critical esteem, however, a tendency to reject these evaluations has become evident. A Persian rug or a sonata, it is said, is not necessarily inferior in aesthetic, moral, or spiritual values to a poem, play, or novel. Its value depends on what it does in its own medium. What seems at first to be mere trivial ornament may have deep, symbolic meanings for the culture in which it originated. Certainly the relative size of the products provides no sure measure of value, as, for instance, between a badly designed cathedral and a well-designed miniature painting. The major–minor antithesis may be useful if applied to the relative status of an art in its own cultural context. The status of an art may change within a culture and differ widely from one culture to another. Tattooing has been a major art in some primitive cultures, notably among the Maori people of New Zealand. Mosaic was a major art in Byzantine culture. Poetry, for the present, has declined in importance in Western civilization. The creative energies of man flow at different times through different channels.

Mode of presentation as a base. One way to avoid questions of value in a descriptive classification of the arts is to use as a base the concept of the sense to which the work is primarily addressed, instead of fineness or beauty. Painting, sculpture, and architecture are said to be addressed primarily to the sense of vision; hence they are now increasingly known as "visual arts." Music is an "auditory" art. Opera and sound film with colour are "audiovisual." An armchair is addressed not only to vision but also to the sense of touch and to muscular sensations of comfort and fitness for the posture or range of postures desired. Thus, a throne, a theatre seat, and a dentist's chair will have somewhat different forms and functions. From the aesthetic standpoint, the visual characteristics of physical objects are usually emphasized. Some philosophers have pursued this approach into the realm of "lower senses," providing for the "gustatory" art of cuisine and the "olfactory" art of smell, as in perfume and ritual incense.

Painting is one species of visual art, and it can be subdivided in various ways. One is according to style, such as baroque or romantic. Another is based on the amount of representation, whether the work is realistic, abstract, or nonobjective. The words abstract and nonobjective are sometimes used interchangeably, to mean nonrepresentational or intentionally devoid of resemblance to any outside object. At other times, an abstract painting is said to be one that started with a representational conception but omitted some or all its representational details, while a nonobjective picture is one conceived and executed from the beginning in terms of lines and colours without any definite outside reference.

Abstract and nonobjective art

Painting can also be divided according to the subject represented; that is, into portraiture, landscape, fantasy, and so on. Some paintings are colouristic in style; others are linear. Some emphasize perspective, with imaginary vistas into deep space. The other arts can be similarly divided; narrative literature can be in prose or verse; prose narrative can be divided into novels, short stories, and so on.

Poetry is sometimes presented aurally, as in speaking it aloud; at other times, visually, as in reading it silently; at still other times, tactually, as in reading Braille type for the blind. In opera, poetry is presented audiovisually. Originally, when few persons could read, it was presented to the sense of hearing. Now, it is more often read silently. A poem is fundamentally the same whether read silently or heard. In reading it silently, the word sounds are imagined rather than heard. For purposes of classification, the art of poetry may be described as primarily auditory, but now, often visual or audiovisual; in short, it is variable.

The "performing arts" are so designated because of the ways in which they are presented to the attention of observers. Most of the visual arts perform automatically, so to speak. The artist who created the painting or statue did his performing once and for all time, when he made it. There is a temporal sequence in perceiving any complex, three-dimensional work of art, especially a large one such as a cathedral, since the observer must walk around it and perhaps inside it in order to see it fully as a three-dimensional form. The form itself may change in appearance from moment to moment, through changes in sunshine, atmosphere, and shadows. Such changes are usually considered to be indeterminate and superficial, however, not enough to characterize the work of art itself as mobile. In mobile sculpture, such as that of the modern American Alexander Calder, the form as a whole changes in accord with air currents and other pressures from outside. Unless moved in a definite way, as by a motor, the succession of arrangements in mobile sculpture is somewhat indeterminate.

Mobile and static arts

Arts in which the objects do not ordinarily have to move or change radically in order to be properly observed have been called "static arts," "space arts," or "arts of rest," in contrast with "mobile," "dynamic," or "time" arts such as music and dance. Some arts, notably cinema and ballet, unfold actively in both space and time. In cinema, the product—the film projected on the screen—is made to perform automatically; the operator does not have to influence the showing unless something goes wrong. The main performance was done permanently in the photographing, editing, and other processes involved in produc-

ing it. Much the same can be said of a phonograph record or tape, since the performance was completed by the musician before the record was made or marketed. Slight adjustments in tone, balance, volume, and the like may be required at the start; after that, the record performs automatically.

What is meant in speaking of "performing arts," however, is, for example, what the actor does in speaking his lines and moving or gesturing as directed, what the musician does in playing a certain sonata on his piano or violin, and what the dancer does in moving his body as required by the choreography. In these examples, there is usually a set of directions, prepared by the designer, the creative artist or group of artists, and made available to the performer. As a rule such directions allow some scope for original interpretation or minor departures from the score, so that the performance is not purely automatic. Different performances of the same script or printed score may vary considerably as to nuances of expression and yet be recognizable as renditions of the same composition. Such variations help to distinguish an original interpretation (good or bad) from a merely mechanical reproduction. Of course, a creative performer may at times compose his own score or improvise without the aid of any directions.

Early theories contrasted "arts of motion" with "arts of rest" and put painting in the latter category. With the advent of the film, especially such works as Walt Disney's "animated cartoons" in colour, photography and painting began their rapid development as arts of time. They are now arts of both space and time, rest and motion. This development illustrates the need for flexibility in classification, to allow for the accelerating evolution and unpredictable variation of the arts.

Arts in which works are shown or sounded in a definite temporal sequence tend to acquire an appropriate notation to guide that sequence. Chief among these are the performing arts of spoken literature (especially poetry and drama), music, dance, pantomime, and other temporal arts of the theatre. Stage design, costume, and lighting are sometimes classed as theatre arts but are not necessarily shown in temporal order.

The first notation of art

Poetry was the first art for which a definite notation, that of writing and literature in general, was developed. As sung or recited by the ancient bards, poetry left a great deal to be supplied by tradition and individual taste. Indications of rhythmic variation—by separating lines, sentences, and phrases with the aid of punctuation, capital letters, and italics—were slowly developed. Some of the earliest musical notation consisted of a few conventional marks above the words of a poem. Dance steps, which were simple and regulated by tradition with the aid of music in songs and rituals, remained long without notation.

The notations of music, from the Renaissance to the mid-20th century, became very precise in directing pitch, melody, chord structure and progression, rhythm, metre, phrasing, dynamics (loud and soft), and many nuances of expression. Because of its scope, precision, and adaptability to many styles of music, the printed score was widely used as a framework for forms that combined music with words, such as song, cantata, oratorio, or opera. It was found impossible, however, to record Oriental, primitive, and other exotic styles of music and dance in Western musical notation. Such music includes pitches that cannot be precisely indicated in Western scales; its rhythms tend to overflow European bar lines, and the timbres of its instruments and voices cannot be exactly described in conventional Western notation. Such limitations were even more keenly felt as Western avant-garde composers sought to use a great variety of sounds, instrumental and otherwise, that had not been previously used in serious music—sounds of nature, city life and countryside, birdsongs and traffic noises, flowing water, thunder and lightning, and many new sounds derived from electronic machinery. In music of the 20th century, timbre and tone quality tended to be emphasized rather than conventional melodic and chordal progressions. Timbre, rather than rhythm or pitch, became the principal component of experimental music. Physical scientists cooperated with composers in adapting the new knowledge of electronic sound to music.

For the dance also, conventional musical notation came in the 20th century to seem less than adequate for guiding movements. Traditionally, the mobile designs of dance had been based on a set of conventional postures and movements, mostly devised in the 18th century. Light and graceful whirls, leaps, and glides on tiptoe could be fairly well symbolized in a simple choreographic notation, which is still used to some extent in traditional solo dance and ballet. As in music, however, style leaders felt the need of a more flexible notation, capable of recording new types of emotional expression.

Another motive for breaking with the past in this regard was the feeling that dance should not be a handmaid to music or limited to expressing musical forms and feelings in bodily movement. Experiments were tried in dancing without music or with rhythm alone, and new bases were sought for recording the moods and movements of the dance in visual symbols.

Imperfections of notation

It would be unreasonable to expect a perfect notation in so complex an art as modern ballet. Even in such a long established notation as the printing of poetry, many qualities—*e.g.*, tempo—cannot be specified. Verbal notation is always incomplete, partly because both poet and reader usually prefer to have much of the word-sound content, along with other meanings, left to the imagination. Printed words are enough to start the reader on an approximate path toward understanding. The sound film and the phonograph offer new, partial substitutes for printed notation as guides to performing the temporal arts.

Other types of base. Three main types of base for classification and division of the arts may be distinguished. They may also serve as aids in defining particular arts or be combined into one system of the arts.

The first is that of medium or material. It is used in the names of such arts as painting, metalwork, woodcarving, jewelry, and ceramics. These refer to the physical materials out of which the work of art is made. Sound waves in music and light waves in cinema are also physical media. Tools and instruments, such as pianos, voices, brushes and canvas, chisels and marble, the human body (as in dance), are all distinctive materials of the various arts. Perceptual qualities such as rhythm, pitch, and colour are parts of the medium of sound film. Poetry also uses rhythm.

The second type of base is that of the process or technic employed, for example, the hands or other parts of the body, or instruments or machines, such as the cameras and projectors used in motion pictures. Shaping, sounding, and verbalizing are three main types of artistic process. They often overlap and combine in various arts.

When photography is conceived merely in terms of operating a camera, it is a medium of art rather than an art itself. It can be used either for purposes of art, as a scientific instrument, as a means of recreation, or in other ways. It approaches art when used as a device for giving someone an aesthetic, visual experience by means of the resultant picture.

Some technics are overt, as in dance, employing the whole body; others are mental and inner, as in memorizing a dramatic role. Some artistic skills are professional or managerial, some mechanical or manual. An architect plans and oversees the construction of buildings and the making of plans and specifications. A landscape architect works with plants, roads, levels of ground, hills and valleys, roads, lawns, bridges, and sprinkler systems. An author may do little that is overt muscular work but much dictating or typewriting.

Some art production is solitary, some cooperative. Some emphasizes visual shaping; some sounding; some verbalizing, usually organizing words and meanings approximately in accord with the rules of grammar and syntax in the language employed.

The third type of base is that of the form, design, and functions of the product, including its function as an aesthetic object; that is, as a stimulus to aesthetic perception and imagination. Arts can be grouped or distinguished as to the amount of emphasis on the following: (1) Presentative (*i.e.*, directly perceptible) and suggestive factors; (2) modes of suggestion (*e.g.*, mimesis symbolism, common association); (3) components in aesthetic form (*e.g.*, melo-

Types of formal emphasis

dy, harmony, rhythm, perspective, plot, characterization); and (4) modes of composition: utilitarian (*e.g.,* as church or palace), representational (*e.g.,* narrative, dramatic, lyric), expository (*e.g.,* essay, symbolic picture), or thematic design (*e.g.,* fugue, sonata, Persian rug, sonnet).

HISTORICAL DEVELOPMENT

In European philosophy, the classification of the arts, or "system" of the arts as it is usually called, has formed an integral part not only of the philosophy of art but of philosophic systems in general. This has been especially true since the work of the late 18th-century German philosopher Immanuel Kant. European philosophers have not limited themselves to making superficial, verbal arrangements but have undertaken to show the role of each art in the mind of man, in world history, and in civilization. A system of the arts may also be used in an attempt to evaluate the arts and to list them in a hierarchy according to their metaphysical and moral roles. In the Western democracies, the classification of the arts has usually been given a more modest, secondary role in philosophy, limited to showing empirical relations.

Thinkers as different as Saint Augustine of Hippo (AD 354–430), one of the chief influences in Christian thought, and Francis Bacon (1561–1626), the English philosopher who was instrumental in the development of modern science, felt the need of a systematic survey of the aesthetic arts as part of a general survey of human knowledge and experience. Separated by more than 1,000 years, Augustine and Bacon disagreed on the value of worldly knowledge obtained through empirical science. Augustine disapproved of such knowledge, while Bacon approved of it. Looking back over his youth, Augustine found much to regret and repent in the pleasures afforded by the various arts to the five senses. Bacon, in his monumental survey the *Advancement of Learning,* admired the progress of the sciences and foresaw their future benefits to man; to poetry, painting, and music he gave a fairly high place, but in the "voluptuary arts," which appeal to the lower senses, he found little to praise.

Kant, writing toward the end of the 18th century, covered the field in a tolerant, empirical way, avoiding moral and metaphysical dogmatism. Beginning with "art" in general (in the broad, technical sense including all transmitted human skill), he distinguished it from nature, science, and paid handicrafts. He then distinguished between aesthetic art and mechanical art, the first of which he subdivided into fine, or beautiful, art and agreeable, or pleasant, art. His division of fine art into the arts of speech, the shaping arts, and the arts involving the beautiful play of sensations led him eventually to further subdivisions into poetry, music, landscape gardening, the art of colour, buildings, and furniture. Under agreeable art, he assigned a place for dinner music, table arrangement, and entertaining narrative.

Georg Wilhelm Friedrich Hegel (1770–1831), another German philosopher who was vitally concerned with the arts, offered a system combining the ancient Greek philosophical conception of a cosmic mind (which embraced all man recognizes as reality) with the theory of evolution. In the world process envisaged in his scheme, he assigned an important role to the arts: architecture is most capable of expressing the early, symbolic stage in world history; sculpture, the classic stage; painting, music, and poetry, the romantic stage.

The symbolic, classical, and romantic stages

The article on the Fine Arts, in the 11th edition of the *Encyclopædia Britannica* (1910–11), by the English man of letters Sir Sidney Colvin, is one of the few attempts in English to deal with the subject in a detailed, systematic way. He proposed three main divisions: the first, into shaping, moving, and speaking arts; the second, into imitative and nonimitative arts; and the third, into serviceable and nonserviceable arts. This triple division expressed Colvin's belief that no one formula could adequately describe the manifold interrelations of the arts.

On the other hand, Max Dessoir, a leading German aesthetician in the first part of the 20th century, tried to combine several bases in one pattern. In his construct, space and time arts are arranged in two vertical columns, with sculpture, painting, architecture, and plastic arts listed under "space" and mimicry, poetry, music, and poetic arts under "time." A third parallel column contrasts the arts "of imitation and definite associations" with the "free arts of indefinite associations." He commits the same error that many previous writers made in supposing that the space arts coincide with the arts of rest and the time arts with the arts of movement and succession and that sculpture and painting are necessarily imitative (*i.e.,* representational). He did not recognize the extent to which the traditional arts had changed, making new modes of classification necessary.

Instead of rectangular diagrams, Étienne Souriau, a French aesthetician of the mid-20th century, offered a wheel-shaped one. By means of this form, he showed how seven basic types of perceptible data (lines, volumes, colours, sounds, and so forth) were developed into complex arts. Using concentric circles, he showed how each type of datum developed into a nonrepresentational and a representational art. Though no such diagram can offer a complete picture, Souriau presents a large number of interrelations in a simple pattern.

BIBLIOGRAPHY. ST. AUGUSTINE, *Confessions,* trans. by E.B. PUSEY, pp. 223–239 (1838), one of the earliest extant classifications of the arts, was made in disapproving of them on moral and religious grounds; FRANCIS BACON, *De Augmentis Scientiarum,* is an enlargement in Latin of *The Advancement of Learning,* bk. 2, ch. 1, vol. 4 of *The Works of Francis Bacon* (1858), a profound, far-reaching survey of the state of science and technology in the early 17th century, which includes a short classification of the arts; SIDNEY COLVIN, "Fine Arts," in *Encyclopædia Britannica,* 11th ed., vol. 10, pp. 355–375 (1910–11), is one of the few attempts in English at a detailed, systematic classification of the arts on various grounds; MAX DESSOIR, *Ästhetik und allgemeine Kunstwissenschaft* (1906; Eng. trans. by S.A. EMERY, *Aesthetics and Theory of Art,* 1970), is a system of the arts reduced to a short, neat pattern, by a leading German aesthetician; G.W.F. HEGEL, *Vorlesungen über die Aesthetik,* trans. by F.P.B. OSMASTON, *Philosophy of Fine Art: Introduction* (1920), is a system that is remarkable for its cosmic breadth; IMMANUEL KANT, *Kritik der Urteilskraft,* part of this work trans. by J.C. MEREDITH as *Critique of Aesthetic Judgment* (1911), is a pre-evolutionary empirical system, based on common sense and broadly tolerant toward the many manifestations of art in his time; THOMAS MUNRO, *The Arts and Their Interrelations,* rev. ed. (1967), is a critical survey of the various classifications of the arts up to the mid-20th century, with reference to more than 400 arts and types of art; *Oriental Aesthetics* (1965), on types and classifications of the arts in India, China, and Japan; and ETIENNE SOURIAU, *La Correspondance des arts* (1947): a wheel-shaped diagram exhibits the system of the arts from a modern French point of view.

(Th.M.)

Criticism of the Arts

Criticism of literature, of music, of the visual arts, or of any other of the arts can be a controversial enterprise. For one thing, disputes arise about the purpose and nature of the judgments made by critics; for another, disputes arise about the nature and properties of what it is that critics discuss—the works of art themselves; and for still another, disputes arise about the possibility of generalizing about *all* the arts, each of which differs significantly from the others in many respects.

Each of these disputes is really a complex cluster, involving nearly all of the persistent issues involved in the practice of criticism. Although this article cannot pursue all of these issues, it will attempt to provide an orientation that will facilitate their investigation.

The article is divided into the following sections:

THE NATURE OF AESTHETIC CRITICISM

The concerns of the critic. It may be safely said that criticism in all the arts is concerned with the description and assessment of particular works. Although some hold that the critic's proper function is only the appraisal or evaluation of works of art, it is certain that nothing can be evaluated unless it is describable beforehand. Accuracy of description can in fact be so extremely difficult, particularly for complex or unfamiliar works and traditions, that some critics devote their efforts largely to matters of a descriptive sort.

It may also be said that all critics of the arts are concerned with the description and evaluation of works of art *as works of art,* even if they are more interested in moral, political, religious, ideological, or other considerations. For example, the moral criticism of works of art presupposes that the actual properties and features of a given work first may be fixed and then may be examined further according to moral considerations. Similarly, religious, political, or other considerations can only follow some relatively objective determination of what it is that is being examined. It is in this sense that criticism of the arts is primarily aesthetic. Although the objectivity of a description or an evaluation may be argued, as well as the propriety of concentrating on only the aesthetic aspects of a work, the central point is not open to dispute; *if* given works of art are to be appraised on any grounds, they must be independently describable beforehand. One cannot judge what one cannot identify and describe.

Primacy of aesthetic issues

The properties of works of art. Quarrels also arise about what may rightly be construed as the properties of works of art. For example, to say that a novel *expresses* a certain Gothic longing or that a painting *symbolizes* the end of feudalism or that a work of sculpture *represents* a bird in flight presupposes some theory of what a work of art is. The object cannot be described truthfully unless it is of a sort that can exhibit the properties ascribed, and quarrels often arise about what may rightly be construed as the properties of works of art as such. To see this is to see the sense in which descriptive statements are theory-laden. The theory of criticism and the theory of the nature of a work of art are largely aspects of the same question.

Provisionally, then, the criticism of works of art may be considered, at least minimally, as aesthetic criticism—that is, centred on the actual properties of the works considered. Correspondingly, appreciation of art may be said to be aesthetic when the art is savoured or enjoyed in terms of the properties that may be discriminated in it. Appreciation, in this sense of the term, may be informed by criticism, since both are focussed on the same properties. This point should be emphasized because it has been held by Leo Tolstoy and others that a genuine appreciation of art must be a naïve, direct, or uninformed response. If serious thought and labour go into the creation of works of art, however, there is no reason why a similar effort will not be needed to appreciate or criticize them.

Identifying and describing works of art. The crucial conceptual issues in criticism of the arts are what is meant by a work of art and how its properties may be ascertained. The difficulties involved in these issues are principally of two sorts. One concerns the identity and individuation of works of art. For example, different performances of the same piece of music may be perceived to differ from one another. A given piece may be transposed for different instruments, or it may be performed on instruments that have been modified since the date of composition, and performances by different artists will regularly exhibit noticeably different qualities. Nevertheless, it is normally considered one and the same piece of music as long as it is identified by reference to a particular score, even though aesthetically there may be more interest in the subtle variations in the performances than in what they have in common. Some critics hold that all distinct performances must be compatible with some ideal performance or that all admissible performances must correspond to the fundamental score, and deviations from it result in serious logical paradoxes. It does seem possible, however, to admit as instances of the same work performances so different from each other as to appear incompatible with any idealized performance.

The same conclusion may be drawn for all the arts that rely on a notational scheme. Variant versions of a poem, for example, may count as instances of the same poem, and analogous instances may be found in theatre, dance, film, and in view of the increasing importance of the blueprint, architecture. Analogies may be found in the plastic arts as well.

The referent of criticism

It is important to emphasize that an entire range of questions regarding the objectivity of criticism rests on puzzles respecting the referent of criticism. Both in describing and in evaluating particular works, it must be made clear whether different versions of the same work of art or several distinct works of art are being discussed. Identity may sometimes be established by detailed analysis, for example, by providing evidence on notational and cultural grounds that what seemed to be different folk songs may be construed as variant versions of the same song.

The principal issues of criticism concern not the identification of works of art but their description and evaluation. It is not possible to segregate entirely what may be correctly said about a given work of art and what it is that individuates that work of art—that is, makes it a distinct entity within a class of related entities.

To illustrate this point, imagine that the central image of a given work may be construed in two radically different ways. Some of Anton Chekhov's plays, for example, may be construed as either comedies or tragedies, depending on whether the views of the author or those of the director Konstantin Stanislavsky are preferred. If the possibility of such competing views be admitted (even though the mere admission of them presupposes a certain theory of criticism), then any account of what a critic does must accommodate this possibility.

When critics say what a certain image means and their

(defensible) accounts conflict with each other, it is clear that they cannot be *describing* the work at hand. It should be possible to confirm a description by an independent study of the object described. Obviously, it is not possible to confirm descriptions that are incompatible.

It might be argued that the differences in conflicting critical accounts are merely apparent, as a coin might be described as circular from one view and elliptical from another. To resolve such an apparent difference in a physical object like a coin, one could invoke some commonly accepted canon such as saying that when it is seen "under normal circumstances" the differences disappear. In criticism of the arts, however, there is no such common canon in terms of which the competing views may be sorted. Moreover, in some schools of criticism, such conflicting views would not be ruled out as impossible.

Interpreting works of art. The solution is to distinguish between *describing* works of art and *interpreting* works of art: a critic's description may be true or false of the work in question, but his interpretation of the work can be only plausible or implausible. Each of several incompatible interpretations may be plausible, but they cannot all be true. In rendering an interpretation of a work of art, then, the critic must be imputing to it properties that cannot with certainty be found in it. This view of critical practice implies a theory of the nature of a work of art that allows some properties to be considered either truly in the work or only plausibly imputed to it.

Distinction between description and interpretation

A substantial body of criticism is intelligible only if the critics who purport to be describing given works—incorporating analysis, comparison, historical or biographical explanation, and the like—are, to some extent at least, interpreting those works. The enormous body of criticism of the greatest poetry, drama, and fiction shows how dissimilar the apparently descriptive efforts of critics may be. The mere existence of this vast body of diverse views would appear to refute the claim that interpretation is merely the unearthing of what is descriptively true of a work of art. It may include this but it must be more.

Interpretive criticism would be unintelligible without a counterpart theory of the nature of a work of art. The key to that theory lies in not confusing works of art with the physical objects associated with them. Physical objects having different physical properties cannot be the same object. As has been noted, however, works of art with different physical properties, and even critical aesthetic differences, may count as instances of the same work of art. This fact is most easily seen in literature, drama, music, and, to some extent, dance and architecture, arts in which the identity of a given work depends on some notational system. It is less easily seen in the visual arts, even though different tokens of the same work, as in etching, may be admitted there too. Clearly, then, works of art are individuated in ways that substantially depart from those of physical objects and in ways that vary from art to art.

A further consideration is that a work of art is a purposive system, whose internal organization may be seen to reflect the systematic decisions of the artist. Criticism explicates this purposive organization, and critics are sometimes obliged to impute a design to a work of art rather than merely to find it inhering in the work. Under those circumstances, the critic turns from description to interpretation. Thus, the very nature of art leads to interpretation as a legitimate, and even inescapable, practice of critics.

An even more decisive difference between physical objects and works of art lies in the sorts of properties that are normally attributed to them. For instance, *expressive* qualities may be assigned to works of art: a novel such as Gustave Flaubert's *Madame Bovary* may be said to express a certain "bourgeois consciousness." Or *symbolic* qualities may be assigned: the castle in Franz Kafka's novel of that name has been said to symbolize divine redemption. Or, *representational* qualities may be assigned: Pablo Picasso's painting "Guernica" is called a representation of the horrors of war. Also, *meanings* may be assigned in various ways: in William Shakespeare's *Hamlet,* for example, the meaning of Laertes' exchanges with Polonius and Claudius is said to be in their contrast with Hamlet's self-doubt; and the meaning of Bigger Thomas' dilemma in Richard Wright's novel *Native Son* is said to derive from the plight of the Negro in America.

As has been noted, it is difficult to say whether properties such as these are to be found *in* works of arts or only may be imputed *to* them. Out of respect for the artist, critics may assume their own role is subsidiary—*i.e.,* that of identifying properties of a work that might not otherwise be appreciated. Because it cannot be said with certainty, however, whether expressive, symbolic, or representational qualities or meanings and the like are actually *in* works of art, criticism must go beyond the severer canons of description to those of interpretation. This issue of the autonomy of criticism has been debated in the works of T.S. Eliot, Oscar Wilde, and many others.

Thus, as has been shown, critics may well offer incompatible accounts that purport to describe the same work. Even in cases in which the identity of a work of art is closely associated with its physical properties, as a work of sculpture might be associated with a mass of cast bronze, pure description may be adequate only for its physical properties but not for the properties that may be ascribed to a work of art. If a line in a painting is called "humorous," it is not always clear whether the line is being described or interpreted. What is clear is that critical interpretation must conform to certain minimal constraints. It must be compatible with whatever is descriptively true of the work in question; any interpretation that is compatible only with what is descriptively false is inadmissible. Interpretation, therefore, depends on some indisputable range of critical description.

Constraints on interpretation

The limits of this range, however, are not easy to establish. To do so requires deciding which among competing critical practices is the correct one. Such a comparison of critical hypotheses leads to metacriticism—that is, to the evaluation of the critical norms themselves, an extremely difficult matter. Without becoming entangled in this problem, some pragmatic rules of critical practice may be laid down: it should be possible to formulate its canons, these canons should be used by a community of practitioners, they should be effective in confirming or denying relevant claims, and it should not be necessary to distort them to examine new works or works that were not anticipated when they were adopted.

These rules permit a number of critical approaches and, therefore, allow alternative and incompatible interpretations of given works of art. It would be difficult, however, to justify the restriction of critical practice to any highly specialized way of proceeding. In literary criticism, for example, it would be difficult to reject out of hand the legitimacy of many diverse but fruitful approaches, such as the archetypal criticism of the Canadian critic Northrop Frye or the Marxist criticism of the Hungarian statesman and writer György Lukács or the Freudian criticism of the British psychiatrist Ernest Jones or the neo-Aristotelian criticism of the U.S. critic Ronald S. Crane or countless others.

An extreme view of criticism holds that the interpretation to be preferred is the one that maximizes the value of the work in question. This view, however, does not recognize the rigour and independence of criticism or its obvious analogies to other endeavours, such as science, that are concerned with claims to truth and validity.

The objectivity of critical practice may be conceded not only to description but also to interpretation. Since interpretation is concerned with plausibility rather than with truth, plural approaches to works of art may reasonably be tolerated. The appreciation of a given work may entail a canvass of all plausible ways of construing it. Moreover, different schools of criticism may prove variously fruitful for each of the arts or for the movements within them. A Freudian approach to music or architecture, for example, promises less than formalist approaches; without question, however, the Freudian approach would be more fruitful in literature than in architecture. To examine the diversity of tenable critical practices (and there are new ones all the time) is to doubt that all but those consistent with some supreme critical principle can be disqualified.

Evaluating works of art. In addition to the identification, description, and interpretation of works of art, the

problem of their evaluation has been the subject of a number of disputes. The issues are those of value theory in general: there seems little reason to suppose that value judgments of works of art differ significantly from those of conduct, for instance. Immanuel Kant held that aesthetic and moral judgments are of fundamentally different logical kinds, but he was chiefly concerned with taste (which, despite Kant, appears in the moral, as well as the aesthetic, domain, in judging such matters as tact, decency, and personal ideals) and he failed to consider the implications of professional criticism and connoisseurship.

Nature of values

Although all the basic issues of the nature of values are not within the scope of this article, two fundamental issues are vital to a survey of the theory of criticism. For one thing, expressing one's tastes, preferences, likes, and dislikes must be carefully distinguished from judging merit. And for another, the principal varieties of value judgments must be sorted out. Quarrels about the critic's appraisal and evaluation of works of art are linked to these issues.

There is no prevalent philosophical view regarding value judgments that can be formulated in a simple and straightforward way. The account that follows puts forward one viable thesis, which may clarify how disputes about values can or cannot be resolved in a domain dominated by considerations of taste.

Value judgments are such in virtue of their predicates—that is, what they affirm or deny about the subject at hand. The predicates themselves, however, are not inherently valuational or nonvaluational; they are construed either way on the basis of governing theories. For example, to judge that Peter is tubercular may constitute a value judgment if good health is accepted as a norm. If tuberculosis is not regarded as a normative concept, however, but as a purely biological phenomenon, the judgment remains factual though not valuative. Factual judgments may be assigned truth values—that is, they may be determined to be true or false. Value judgments, on the other hand, are judgments that entail reference to a norm, or standard of merit. Some value judgments may also be factual judgments as well, as in the case of the judgment about Peter's tuberculosis.

Statements about one's likes or dislikes—one's tastes, in short—are usually not considered normative statements but statements of fact. They concern an aspect of a person's disposition and behaviour.

In making value judgments, however, the person puts a value on things by judging them in accord with norms that usually lie outside his own tastes. Of course it is possible that the values he places depend entirely on his tastes—his likes, dislikes, preferences—but it is not necessary. Herein lies a crucial distinction, for value judgments that depend on personal tastes are strikingly different from those that do not. Value judgments that depend upon taste may be called appreciative judgments and those that do not may be called findings.

Findings differ from factual judgments only insofar as they are linked to standards of merit; otherwise they are alike. In principle, critical judgments of the merit of a work of art, like other value judgments, may be as rigorously objective and confirmable as judgments of fact. Logically, findings are judgments of fact, in that truth values may be assigned to them.

It should be emphasized that questions of the relativity of values, or of their variability, are not relevant here. The logical properties of findings are not affected by the nature of socially institutionalized values. Value judgments about beauty or fashion may well derive from the prevailing tastes of a community, but, as findings, they will be made in accordance with those norms and not with one's personal taste (if we disregard coincidence).

In contrast, appreciative judgments are based entirely on one's own tastes. The well-known Latin maxim *de gustibus non est disputandum,* "there is no disputing concerning taste," applies to appreciative judgments only; it does not bear at all on findings.

Distinguishing between these two fundamentally different kinds of value judgments permits some rational debate respecting the merit of a given work of art. To the extent that a critical tradition is standardized (as the standards of wine tasting or of dog breeding have been established), critical findings may be regarded as objective relative to such a tradition. (The entire system may be challenged, of course, but that is another matter.) But appreciative judgments cannot be objective in that way; they cannot be simply true or false. Appreciative judgments depend on personal tastes, which vary greatly among individuals, so that seemingly contradictory judgments are common. If these judgments were treated as findings, such contradictory judgments might be found to be jointly true, a logical impossibility. If such a contradiction were resolved by regarding each judgment as true relative to given taste, then appreciative judgments could not be distinguished from mere statements about one's tastes. Since this distinctive kind of value judgment is prominent in aesthetic criticism, appreciative judgments may be conceded much weaker confirmation than findings. In a finding, one holds that a given work has (or lacks) a certain merit; the matter can be decided by attention to relevant criteria and evidence, as factual matters are. In an appreciative judgment, one assigns a certain value to a work in accord with one's personal likes, dislikes, and preferences. In findings, the supporting reasons must be publicly compelling; in appreciative judgments, the supporting reasons need only be relevant and reasonable to a critical public. The public may attest to the coherence of an appreciative judgment and its supporting reasons, but, because of differences in taste, it is not bound to share that judgment, as it must in the case of findings.

The problem of conflicting judgments

Incompatible findings, like incompatible descriptions, are inadmissible. Seemingly incompatible appreciative judgments, however, are admissible insofar as they are coherent (in accord with informal canons of such judgments) and compatible with the actual properties of the given work. In the context of an academic tradition, judgments of beauty may well be findings; but, in the context of personal taste, judging beauty is rendering an appreciative judgment. One may (appreciatively) say that a certain woman is beautiful and offer supporting reasons that may seem forceful or not. But, if one claims, in the mode of findings, that a certain woman *is* beautiful relative to a particular tradition of values, the evidence will serve flatly to confirm or deny the claim.

Disputes may arise about judgments of both sorts, and there are procedures for appraising each type. The precept about the pointlessness of arguing about taste should therefore be amended to say that the kind of rational dispute possible for findings is not possible for appreciative judgments, or judgments of taste. Nonetheless, rational dispute of a different sort is possible for appreciative judgments as well.

The two types of judgment—findings and appreciative judgments—cannot be segregated on the basis of their respective range of predicates; an assertion that something is beautiful, for example, may be construed as either type of judgment. The conditions under which relevant claims may be confirmed is the key factor in distinguishing between the two types.

Other aspects of value judgment

There are, of course, deeper questions about values than can be treated in this article. Especially pertinent is this question: are values discoverable in nature, or the real world, or can they be construed only as an expression of the interests of a particular community? If it were possible to prove the existence of "true" or "real" values, there would be no need to account for appreciative judgments; practicing critics would be concerned only with findings—that is, with making judgments that are in accord with those true values. In fact, however, no such true values have been proved, and the critic is chiefly concerned with appreciative judgments, in which his personal tastes are systematically articulated. In any culture there may be a number of critics operating at cross purposes, each imprinting his own distinctive taste on a community of enthusiasts. Revolutionary tendencies in taste appear, however, resulting in significant alterations in critical practice. Thus, the judgments of critics tend to be informal and variable; because of the logical properties of their appreciative judgments and the conventional nature of the norms that support their findings, it could not be otherwise.

These, then, are the principal conceptual issues bearing on the systematic practice of criticism of the arts. The critic's judgments may be either valuational or nonvaluational. With respect to the former—that is, his value judgments—findings must be distinguished from appreciative judgments; with respect to the latter, the nonvaluational judgments, descriptive judgments must be distinguished from interpretive judgments. Because appreciative and interpretive judgments do not allow truth values to be assigned to them in the way that findings and descriptions do, the critic must tolerate logically weak judgments at the heart of his enterprise. This tolerance, however, should not be misconstrued as a lack of rigour in applying critical canons. On the contrary, rigorous critical practice requires such tolerance.

EXTERNAL INFLUENCES ON AESTHETIC CRITICISM

The influence of doctrine and ideology. Actual critical practice raises certain technical questions that tend to divide considerations that are internal from those that are external to aesthetic criticism—that is, those that bear on the work of art as such from those that involve ulterior appraisals or characterizations. In spite of disputes regarding the proper scope and orientation of criticism, the diverse schools of criticism exhibit broadly parallel practices in both the internal and the external considerations to which they address themselves. For example, Marxist critics such as Lukács and Christopher Caudwell use the socioeconomic categories of their philosophy to analyze works of art as art, to characterize the artist's life, and to study the role of the work within the dynamics of social change. Clearly, it is not always easy to discriminate between internal and external criticism.

The issue here is conceptual; it is not open to ideological quarrel. For instance, it has been said that the ancient Greek dramatist Aeschylus in his trilogy the *Oresteia* was presenting an idealized version of emerging notions of political justice. This external interpretation presupposes an internal interpretation of the work that does not violate any canons of aesthetic criticism, or it would not be tenable. Precisely the same consideration applies to the attempt by the Freudian critic Jones to psychoanalyze Shakespeare on the basis of *Hamlet* or Sigmund Freud's attempt to analyze Leonardo da Vinci from his drawings and paintings. The principle that the external and the internal criticism must match remains the same whether the criticism is concerned with libidinal conflicts as are the Freudians, with archetypes as are the followers of the Swiss psychologist Carl Gustav Jung, with class conflicts as are the Marxists, or with theological and moral resolutions of modern poetry as is the French Thomist philosopher Jacques Maritain. In the light of the pluralism of viewpoints that has already been shown inevitable, it is not surprising that internal and external criticism cannot be sharply opposed to each other. Although adherents of some particular approach may disagree, it is difficult to see how any one approach can be exclusively vindicated. In a very real sense, the preference for any given approach may be construed as an appreciative judgment in itself.

Advantages of pluralism

It is not at all necessary that a theory about some aspect of human society be true in order for it to be fruitful and defensible in criticism. In describing, interpreting, and evaluating a work, the critic requires only that the theory is relevant to it. For instance, the critic may appropriately characterize or appraise a work in terms of Marxist, Freudian, Jungian, or Thomist doctrines to the extent that the imagination of either the artist or the audience is informed by such doctrines. Such doctrines need only be recognizable and socially significant to be useful in criticism; scientifically valid theories, on the other hand, may be entirely uninstructive when applied to the description and interpretation of a work of art. False theories may be so rewarding in discussing particular works that they cannot be dismissed solely because they are not true. To understand William Wordsworth in terms of some version of Platonism is not to subscribe to Platonism itself, and to read the *Oresteia* in Marxist terms need not presuppose that Aeschylus held similar sociological convictions. The truth, or at least the plausibility, of the descriptive, interpretive, and evaluative judgments of critics must be carefully distinguished from the truth of the doctrines advanced by them or by artists. The validity of Thomism, for instance, is an entirely separate issue from the validity of interpretations of Dante from the Thomist point of view.

Those critics who believe they subscribe to a science of values, whether it may be Freudian, Thomist, Marxist, or some other, would quarrel against such a use of diverse doctrines. Inasmuch as the foundations of their beliefs do not compel the belief of all others, however, the use of doctrines of doubtful validity in criticism must be tolerated.

A doctrine may be applied by a critic if it is compatible with his own practice of criticism and if it provides an account of the design of that work, especially those elements that seem to call for interpretation.

In sum, the discipline of criticism has been shown to exhibit significant rigour, despite the logical weaknesses of the judgments of interpretive criticism and the uncertainty of the doctrines it applies. Contributing to this rigour are the requirements that an interpretation not deny any descriptive elements; that it be plausible; and that it deal with the traditionally salient problems of the work—that is, the interpretive puzzles that obstruct the assignment of a coherent and comprehensive internal order to the work.

Historical and biographical influences. A final issue may serve to round out this account of the criticism of the arts. As has been seen, an interpretation must be compatible with what is descriptively true of a given work of art. It is sometimes supposed that both description and interpretation should be similarly constrained by historical and biographical considerations. In fact, some versions of this thesis hold that interpretations prove to be no more than description provided under relatively difficult conditions of discovery. In the historical thesis, the "meaning" (*i.e.*, the correct characterization) of a work of art is the one that accords with its critical reception in the culture in which it was produced. The biographical thesis holds that "meaning" is what accords with the artist's intention.

Historical and biographical criticism

A number of things may be said about both theses. For one thing, the two theses may be construed as the same wherever the artist's intention is interpreted in terms of the conventions governing such works rather than in terms of his own avowals of what he intended. It is extremely difficult, however, to suppose that one and only one construction can be put on a given work, especially in complex societies in which artists often intentionally depart from academic or other conventions. The problem of plausible alternative interpretations haunts critical practice even when the problem is confined within historical limits, even though the elimination of such alternatives is often claimed to be an advantage of historical criticism.

Furthermore, plausible interpretations that are put forward long after the work of art is produced may also be construed as the work of historical criticism; the restriction of interpretation to the period in which the artist worked represents a somewhat arbitrary constraint on the practice of criticism. Some works of art attract critics and audiences over generations, or even ages, without requiring close attention to historical or biographical origins. The force of accurate historical criticism cannot be denied, but criticism that ignores the *evolving* historical reception of a work of art unjustifiably restricts the scope of criticism.

A historical hypothesis, particularly one concerned with cultural significance, bears a striking similarity to interpretive criticism; like interpretive criticism, history as a discipline must rely substantially on criteria of plausibility. It should be noted that the same arguments that apply to historical and biographical criticism also apply to criticism based on cultural anthropology or the history of ideas.

Historical misunderstandings and inaccuracy about a given work of art must be distinguished from the historically changing significance of that work. A Freudian interpretation of *Hamlet,* for example, cannot be dismissed solely because the Freudian system was unknown to Shakespeare. If audiences that share Shakespeare's cultural and linguistic traditions are deeply influenced by the Freudian outlook, then a Freudian reading of his work, if independently plausible, cannot be regarded as a historical mistake. Interpretations that are constrained by *special*

historical and biographical considerations must be distinguished from those that are not. Interpretations of both sorts may be welcomed.

Intentional criticism

Further objections may be raised against biographical criticism that narrowly construes the artist's intention as what may be documented in conversation, letters, and the like. Quite often, there is simply no such documented knowledge of the artist's intention. Even when there is, it may prove vague and ambiguous or even irrelevant to the appreciation of the work. In any case, any interpretation based on a statement of the artist's intention is subject to precisely the same sort of confirmation that other interpretations would be. Intentional criticism may be viewed as a form of historical criticism that assumes certain types and modes of artistic endeavour within a tradition (to which the artist's intentions are said to correspond).

Some critics opposing what the American critics W.K. Wimsatt and Monroe C. Beardsley called the "Intentional Fallacy" would not admit any criticism based on the artist's intention; they would rule it out on the grounds that it violates the proper aesthetic concern of the critic. In view of what has already been said in this article about the nature of a work of art and of its boundaries, however, intentional criticism may be viewed as simply another selective mode of critical practice, one that is by no means exclusively correct but not for that reason inadmissible.

In conclusion, then, the principal issues affecting critical practices in the arts may be said to be (1) the identity and individuation of works of art; (2) the nature and characteristic properties of works of art; (3) their description and interpretation; (4) value judgments about them; and (5) the admissible varieties of, and constraints upon, criticism of the arts.

BIBLIOGRAPHY. Perhaps the most comprehensive canvassing of the current philosophical literature respecting criticism of the arts is to be found in M.C. BEARDSLEY, *Aesthetics: Problems in the Philosophy of Criticism* (1958); and J. MARGOLIS, *The Language of Art and Art Criticism: Analytic Questions in Aesthetics* (1965). Each of these two books (with extensive bibliographies) proposes its own theory of criticism; the first reduces interpretation to description, while the second opposes such reduction. Of relatively recent theories of art bearing on the problems of criticism, the following are among the most penetrating: N. GOODMAN, *Languages of Art: An Approach to a Theory of Symbols* (1968), primarily addressed to notational problems in identifying works of art; R. WOLLHEIM, *Art and Its Objects* (1968), pursues the paradoxes of numerical identity regarding works of art; S.K. LANGER, *Feeling and Form* (1953), attempts a systematic ordering of the different arts in terms of symbolic form; and H. OSBORNE, *Aesthetics and Criticism* (1955), construes criticism primarily in evaluative terms and the work of art as the object of such practice. Recent anthologies of the most widely discussed papers include: W. ELTON (ed.), *Aesthetics and Language* (1954), a linguistically oriented attack on the Italian philosopher Benedetto Croce and on Idealist aesthetics; and J. MARGOLIS (ed.), *Philosophy Looks at the Arts: Contemporary Readings in Aesthetics* (1962).

(J.Ms.)

Practice and Profession of the Arts

The practice of the arts goes back to the remote prehistory of mankind; the profession of the artist is of more recent origin. The oldest surviving works of art such as the cave paintings at Lascaux, France, and Altamira, Spain, dating from the late Paleolithic Period between 10,000 and 15,000 BC, were presumably made by men who never thought of themselves as artists. Though these works can be viewed as art today, it is probable that their creators were intent only on executing a magic ritual designed to aid them in the hunt.

At some point in the development and differentiation of human society, the professional artist emerged. The monumental buildings and stone carvings of ancient Egypt or Assyria and the intricately decorated pottery and elegant murals of Crete were undoubtedly done by trained and experienced craftsmen. They were no longer occasional practitioners of art, as the cave painters had been, but full professionals—*i.e.,* men skilled in a specialized occupation, practicing it full-time and probably earning all of their livelihood from it.

But even the relatively simple definition of a professional as a person receiving pay for carrying out a specialized occupation on a full-time basis is difficult to apply to the arts. In no other field do the categories of professional and amateur so overlap. One of the 20th century's major poets, Wallace Stevens, and one of its most innovative composers, Charles Ives, earned their living as insurance executives. By every criterion of proficiency and accomplishment, both deserve to be ranked as professional artists, yet in a real sense, both were amateurs. Similarly, it is difficult to classify Thomas Jefferson as a professional architect, even though he designed some of the most beautiful buildings of his time.

The phenomenon of amateurism in the arts represents only one of the key problems in an attempt to discuss the professional artist. There is also the question as to whether the arts are professions, as distinct from crafts or skills. This aspect of the concept involves consideration of matters of the status or prestige of the arts and decisions on what constitutes proper preparation for careers in them. On these issues, attitudes have varied enormously from one art to another, from one historical period to another, and from one culture to another. Nearly everywhere and always, music, architecture, and poetry have been regarded as professions, while pottery making has been regarded as an art and granted the dignity of a profession only in some non-Western cultures.

Since antiquity, there has been a continuing debate as to whether painting and sculpture demand skills of eye and hand alone or intellectual grasp and training as well. The painter has often spoken contemptuously of the sculptor, describing him as a mere stonecutter, and the sculptor, in turn, has looked at the painter as simply an illusionist and trickster.

The attempt to gain for some of the arts the status of learned or quasi-learned professions resulted in a distinction between "the fine arts" and "the applied arts," a distinction that has done harm to both. The notion that a painter has a profession, but a cabinetmaker or silversmith only a craft, helped to isolate the former from everyday life and to limit the creative enterprise of the latter. The gradual breakdown of the attitude that permitted such exclusive categories to be created and the general rejection of the false distinctions between art and utility may be among the most encouraging developments in the arts in recent times.

Of even greater importance has been the lessening of the separation that exists between the professional and the practitioner. The "cultural explosion" following World War II not only made the arts more readily accessible to the general public but also stimulated artistic activity in great numbers of people. Never have there been so many professional artists active in so many fields, and never have there been so many nonprofessional practitioners of the arts.

The present article treats the practice and profession of the arts in terms of training, working conditions, and social and economic roles and influences. The phenomenon of fraudulence in the arts is also discussed. The principles of aesthetics and other theoretical aspects of the arts are treated in the articles AESTHETICS and PHILOSOPHIES OF THE BRANCHES OF KNOWLEDGE: *Philosophy of art.*

This article is divided into the following sections:

Preparation of the artist

Formal education for a profession in the arts is a relatively new development. The school of art or architecture, the conservatory of music, the academy of dramatic art, and the university's creative-writing program have emerged in modern times, some within the last generation or two. To most artists of the past it would have seemed incomprehensible that they should go to school to learn their profession. Neither would any of the great educational institutions before this century have thought it part of their function to teach artists. Indeed, serious doubts continue to be expressed as to whether academic schooling is the best training for most branches of the arts.

Formal instruction in the visual arts—painting, drawing, sculpture, and architecture, as well as the many arts of decoration and design—has the longest and most diversified history. A consideration of the patterns of development that have occurred in these related fields will reveal the general directions of training in all the arts as well as many of the basic problems running throughout the history of art education.

THE VISUAL ARTS

The ideal of the universally trained artist

Throughout much of history, the insistence that the artist be learned in many fields has constituted a deliberate attempt to raise his status, to rank him as a practitioner of the liberal rather than the mechanical arts. The Roman architectural theorist Vitruvius (flourished 1st century BC) declared:

> [The architect] should be a man of letters, a skilful draughtsman, a mathematician, familiar with historical studies, a diligent student of philosophy, acquainted with music; not ignorant of medicine, learned in the responses of jurisconsults, familiar with astronomical calculations . . . (From *De Architectura,* F. Granger [tr.], Harvard University Press, 1962.)

Such a prescription represented the ideal of the universally trained artist, and, so far as is known, no institution existed in antiquity to provide this or any other kind of formal training in the arts.

The traditions of apprenticeship. Until the 16th century, when the earliest academies of the arts appeared, the artist acquired his skills mainly through various systems of apprenticeship. He learned his trade as he practiced it under the instruction and supervision of a master. Sometimes the apprenticeship was regulated by a craft guild. Typically, a boy was bound to a master at the age of 14 and served for seven years. This system was in force throughout most of the Renaissance, and it was under such rules that Michelangelo entered the workshop of Ghirlandajo and Leonardo da Vinci that of Andrea del Verrocchio. Gradually the system changed so that, in the 17th century, the beginning artist came to be considered a pupil rather than an apprentice. He lived and studied in the home of a master for an indefinite period and was free to leave when he felt he had learned enough.

Functioning of the systems. Whether as apprentice or more informally attached pupil, however, the young artist was trained on the job by his master. As an apprentice, he began by doing the most menial jobs: grinding the colours, cleaning the brushes, or preparing the wood panel or the plastered wall. Gradually, he was trusted with more responsible tasks, given instruction in the technique of the art itself, and allowed to paint in some of the decorative details of an altarpiece or fill in large neutral areas of a fresco. Finally, he might be assigned the subordinate figures or landscape background of a painting by the master. It is said that Leonardo as an apprentice painted so beautiful an angel in a "Baptism of Christ" by his master that the master resolved to renounce painting and devote himself to sculpture.

The student, from menial to master

Strengths and weaknesses of the system. The training of an apprentice in a Renaissance workshop was not confined to a single art. The workshop of Antonio and Piero Pollaiuolo, perhaps the most active art establishment in Florence of the mid-1400s, accepted commissions in sculpture as well as in painting of all sorts. The graduate of such a workshop, then, if he had kept his eyes open and his hands busy, could have developed a wide range of artistic skills. The apprentice system, however, had serious weaknesses. In a good workshop run by a conscientious master, an apprentice could receive excellent training. Far too many apprentices, however, undoubtedly were condemned to years of repetitious drudgery, receiving little systematic instruction and, at best, being trained purely as craftsmen. Moreover, the guilds deliberately limited the number of apprentices.

Academies and the artistic elite. The founding of the first academies of art, in the late 16th and early 17th centuries, was a step forward. Although later such institutions became restrictive and rigid, they were designed to break the monopoly of the guilds, to regularize the training of artists, and to lift artists from the category of mere craftsmen by liberalizing their education.

The French Academy as model. The most famous and influential of all academies of art was the Académie Royale, commonly known as the French Academy, founded in 1648 in imitation of the Accademia di S. Luca, which had been set up in Rome in 1593. Under the patronage of Louis XIV's great minister, Jean-Baptiste Colbert, the French Academy by the 1660s had become a major force. Under the directorship of the painter-decorator Charles Le Brun after 1683, it exerted a virtual dictatorship over French art. Its success inspired the creation in other countries of official academies that aped its organization and program.

Curriculum and rewards. The curriculum of the French Academy, fixed under Colbert and Le Brun, centred around drawing. The student drew from the drawings of his professors, then from casts, and finally from life. In addition, he attended academic lectures, analyzing pictures from the royal collection. These discourses were designed to instill in the student certain aesthetic criteria and a fixed hierarchy of values in which subject matter played a leading role. In the scale, historical paintings were regarded as the noblest, still lifes as the meanest.

The academy's students competed for a series of prizes culminating in the Prix de Rome, a four-year scholarship

in Rome that assured its holder a successful career at the top of his profession. Even those who won no prizes could anticipate a secure future in the service of the state or of individual patrons, for the steady production of uniformly trained and indoctrinated artists was of great importance to the regime.

The academy performed a great service in breaking the stranglehold of the guilds on the training of artists, in raising the standards of art instruction, and in improving the status of artists. On the other hand, it locked the artist into a closed system, made conformity a virtue, and treated individualism and originality of style as sins to be avoided at all costs. By standardizing methods of instruction, it also standardized the bases of critical judgment. By admitting only 200 students and thus creating a small artistic elite, it neglected the need for trained artists and designers in the decorative and applied arts. Academicism established above all the principle that the artist succeeded not through native genius but through a correctness of technique that could only be acquired through proper training.

Academicism as a negative force

The English Royal Academy. In England, a number of schools of art were established during the early 18th century, notably one founded by the painter and caricaturist William Hogarth. In the employment of female models, these schools went beyond anything yet attempted in France. It was not until 1768, however, that the patronage of the crown was obtained for a Royal Academy. By the end of the century, this academy had achieved immense prestige and influence, and a painter who could sign the coveted "R.A." after his name was certain of a prosperous career. The *Discourses* delivered over a period of 15 years by the painter Sir Joshua Reynolds, the academy's first president, remain the best statement of the academic philosophy.

As a teaching institution, however, the Royal Academy fell far short of its French prototype. The limitation to 40 on the number of academicians increased the social and economic desirability of membership but narrowed the scope of the academy's teaching functions. The restrictive curriculum and the discouragement of "eccentricity" of style were not calculated to foster true talent. Of the outstanding English artists of the early 19th century, only J.M.W. Turner was trained and supported by the Royal Academy. The painter-poet William Blake despised it, and landscapist John Constable became an R.A. only late in his career.

Countermovements: the applied arts and crafts. It was not only its self-perpetuating and oligarchical exclusiveness that provoked criticism of the Royal Academy. With the growing impact of the Industrial Revolution, the academy was attacked by reformers in and out of Parliament for its failures to encourage the applied arts and to remedy the growing shortage of trained artisans in such fields as china manufacturing, ornamental metalwork and plastering, and carving in wood and stone. Parliamentary hearings sharply critical of the Royal Academy led in 1837 to the establishment of a government School of Design and an accompanying museum.

Schools and museums of design. The School of Design and its branch schools were succeeded, after 1852, by a network of art schools under a government Department of Practical Art. By 1864 there were 90 such schools instructing about 16,000 students. The founding, in 1852, of the Victoria and Albert Museum, which came to have vast collections of design in all fields, was an important step forward in art education. This combination of schools and museums of applied art was soon imitated in many of the leading cities of Europe, especially in Austria, Germany, and The Netherlands. Despite attacks on the basic principle of separating schools for the applied arts from those for the fine arts, such schools continued to flourish in the 20th century. Most of them, however, modified their curricula to include such studies as art history and aesthetics.

Waning of academic dominance. In painting and sculpture, however, the academic idea lost some of its vitality and dominant influence in the 19th century. By the mid-18th century, the French Académie Royale was under attack from the philosophers of the Enlightenment, and in 1793, at the height of the Reign of Terror following the French Revolution, it was dissolved and its functions handed over to a Commune des Arts led by the painter Jacques-Louis David. The Institut de France, set up in 1795 to supervise the arts, revived the academy in everything but name. The academic establishment, under different forms, remained a powerful though no longer controlling force in the French art world. As the École Nationale Supérieure des Beaux-Arts, it continues today to train artists to compete for the Grand Prix de Rome.

Continued influence of the masters. Even in the heyday of the academy, the system of training the artist in the master's studio had continued to flourish. In 1771 the Irish painter James Barry estimated that 5,500 persons were receiving some kind of art instruction in Paris alone and that of these, 1,500 were being trained to work in such industrial enterprises as the Gobelin tapestry works. Thus most artists were still trained in the studios of established artists or in elementary technical institutes, such as an industrial school founded by Jean-Jacques Bachelier in 1762. France's dominant position in the decorative arts during the 18th and 19th centuries can be attributed in large part to the effectiveness of the Bachelier school and similar training centres.

During the second half of the 19th century, the demand for training in painting and sculpture grew so great that large numbers of beginners congregated in such studios, or ateliers, as the famous Académie Suisse in Paris, where they painted from a model under the often cursory supervision of a well-known teacher.

Copying the old masters

Another training method became important during the 19th century. Much earlier, copying of frescoes such as those by Masaccio in the Brancacci Chapel in Florence had been a major source of instruction for later painters. With the opening to the public of such great museums as the Louvre, the practice of painting from the old masters again became a basic feature of the artist's training. Even so sophisticated a painter as Paul Cézanne was devoted to this method of teaching himself.

Synthesis of art and craft. Despite 19th-century advances in the formal training of artists in all fields, sharp criticisms of prevailing art education were made, notably by the critic John Ruskin and the poet-craftsman William Morris. They opposed the separation between training for the applied arts and education for the fine arts. Morris saw the division as creating a group of designers who turned out standardized patterns for machine-made objects and a group of fine artists who were constrained from producing things needed by people for an aesthetically pleasing environment in daily life.

As a direct result of Morris' teachings, a number of art schools were established in England to revive handicrafts and to bridge the gap between the applied and the fine arts. The most successful of these schools was the London Central School of Arts and Crafts, founded in 1896. After 1900, however, the leadership in the movement to reform art education passed from England to Germany, where there was a decisive change in emphasis. Contrary to Morris' rejection of the machine, an attempt was made to realize the aesthetic possibilities of machine-made objects and also to create a new architecture.

Impact of the Bauhaus. The major figures in 20th-century art education were the Germans Bruno Paul and Walter Gropius. It was Paul's strong conviction that art schools had a responsibility to artists and to society: to train artists in fields in which they could both earn a living and be socially useful. After World War I, he combined the Berlin Academy of Art with the School of Decorative Art. In a pamphlet on art education published in 1918, he declared that all students, whether they intended to become fine artists or applied artists, should receive basically the same training. He stated also that no one should be admitted to an art school without first having learned a trade in a workshop or trade school.

Walter Gropius, in his early career in Germany and his later work in the United States, exerted an incalculable influence not only on art education but also on the whole development of modern art and architecture. Already a well-known architect in 1914, he became principal of the Weimar School of Arts and Crafts. After service in World

War I, he returned to Weimar and in 1919 founded the Staatliches Bauhaus (State Architecture House), a school merging the Weimar art school with the arts and crafts school. In 1925 the Bauhaus was moved to Dessau.

Unison of arts and crafts

The Bauhaus ideal was to unite all arts and crafts to create a new architecture that comprised a living environment; to break down the false separation between the applied arts and the fine arts, between art and utility; and to train artists in the creative possibilities of machine design.

To these ends, the Bauhaus curriculum emphasized the use of different tools and materials, as well as academic instruction in geometry, principles of construction and design, and the history of art. Instruction was divided into three stages. The first, lasting for six months, covered an elementary survey of Bauhaus principles and methods. The second, taking three years, was the basic practical and academic course in which the student, in addition to the general training he received, was required to specialize in a trade under the supervision of a particular master. At the end of this course, the student had to pass one of the regular city trade examinations before, finally, he entered the third phase, *Baulehre* ("building instruction"), in which he took an active part in one of the Bauhaus' projects.

During its short life, which ended in 1933, the Bauhaus achieved remarkable results and assembled an extraordinary faculty that included such leaders and innovators as painter-photographer László Moholy-Nagy, architect Marcel Breuer, and painters Wassily Kandinsky and Paul Klee. It had created the ideal of a new type of art education and proved that it was workable. John Ruskin, as Slade Professor of Art at Oxford in the 1870s, had taken his students out to repair roads, but it remained for the Bauhaus to transform Ruskin's experiments into a coherent program.

Acceptance of the arts in higher education. Though the Bauhaus had no immediate successor, it affected the teaching of art in many schools in following decades. Such education achieved notable growth in university-based or university-sponsored schools of art, especially in the United States. Such schools attempt, in many instances, to combine a four-year academic education with studio work in drawing, painting, sculpture, and other media. Many offer specialized work in such fields as industrial design, book illustration, advertising art, and costume design. The graduates of these schools usually are granted bachelor's or master's of fine arts degrees. None of the schools, however, has attempted to realize the Bauhaus ideal of a student completely trained in both the fine and applied arts.

The same is true of schools of architecture, most of which are connected with universities. In these, the academic curriculum is generally more rigorous than in schools of art. A degree is usually granted after a five-year course, which includes considerable work in traditional liberal arts subjects. A number of schools offer special concentrations in city planning. Few if any, however, see architecture as Gropius did, as a discipline unifying all of the crafts.

THE OTHER ARTS

Music. Next to schools for the visual arts and architecture, the most widespread of contemporary schools for the arts are those for the training of musicians, whether composers or performers. Like the school of art, the school of music has a long history.

The Paris Conservatoire model. The first full-fledged music school of modern times was the Conservatoire de Musique, founded in Paris in 1795 by the revolutionary National Convention as a successor to the earlier École Royale de Chant et de Déclamation and Institut National de Musique. In 1797 it had 125 professors and 600 pupils. It has had a continuous and distinguished history to the present day.

Development of the music curriculum

The curriculum developed at the Paris Conservatoire has been followed with slight variations in schools of music elsewhere. Originally, classes in composition, harmony, singing, and instrumental performance were given. Later, classes in music history were added, and an increased emphasis was given to training in sight reading. Still later, instruction in aesthetics and musical analysis became part of the standard curriculum.

The success of the Paris Conservatoire and the greatly increased demand for trained musicians during the 19th century led to the creation of conservatories in major European and American cities. In the United States, the Boston Conservatory of Music was founded in 1867, the Peabody Conservatory in Baltimore in 1857 (first classes in 1868), and, later, others in New York, Philadelphia, and elsewhere. Some of these schools have remained independent, whereas others have become part of universities.

Private and institutional study. The conservatories of music have not altogether supplanted the older system under which a student worked under a particular master, as Ludwig van Beethoven did under Joseph Haydn at the end of the 18th century. Such a famous teacher of composition as Nadia Boulanger attracted students from all over the world, and singers and instrumentalists often prefer to attach themselves to the studio of a favourite teacher. Increasingly, however, the conservatories have absorbed the master's classes.

Literature. The idea that the college-bred writer is superior to the one who has not had a liberal arts training goes back at least as far as the Elizabethan Age, when Robert Greene, one of the so-called university wits, sneered at Shakespeare as an upstart crow. The notion that creative writing, as distinguished from the ancient discipline of rhetoric, is something that can be taught in an academic environment is a recent development.

In the United States, especially, many creative writing programs have developed in higher education, usually in conjunction with the appointment of a well-known writer in residence as a member of the faculty. The creative-writing student typically receives more broadly based, and less technical, training than the student at a school of art or music.

Theatre and dance. The training of dramatists has taken a direction somewhat different from that of other writers. It has been combined with practical work and experience in the crafts of the theatre.

Academic teaching and apprenticeship. The workshop set up at Harvard by George Pierce Baker and the Yale School of Drama, which he founded in 1925, served as models for later drama schools. In recent years, some university-based schools of theatre arts have developed very elaborate programs, with professional courses in scenic design, lighting, and direction, as well as in playwriting and acting. They often sponsor highly trained repertory companies that present new and often controversial plays as well as the classics.

There also exist separate schools for actors outside the university establishment. Among the best known of these is the Royal Academy of Dramatic Art in London, which has trained many successful actors. The Actors Studio in New York City has enrolled fledgling actors as well as established actors who wish to keep their skills sharp.

Training the performer

The old apprenticeship system remains more a vital factor in the training of actors than in any other of the performing arts. Particularly in countries in which repertory companies are well established, as in England, France, Germany, and the Soviet Union, actors can learn and gain experience in their craft by progressing from walk-on roles to more demanding ones. Often the actor can start in a provincial repertory company and move to one of the national companies. For the experienced actor, the repertory company provides the opportunity to do a variety of roles, and to go from a major to a minor part from week to week, as a member of an integrated and continuing ensemble. The system goes back to the beginning of stable acting companies in the Europe of the 16th and 17th centuries, when actors might enroll as boys and live out their lives as members of the same company.

Formal education for the dance in modern times dates from the establishment by Louis XIV of the Académie Royale de Danse in 1661. In 1669, the Académie Royale de Musique was founded, and a school to train professional dancers was added in 1672. These institutions made Paris, and specifically the Paris Opéra, the world centre of ballet training and performance in the 19th century.

Formal demands of the dance. Training for the dance has made less headway in universities than that for oth-

er of the performing arts. A few colleges in the United States offer a fully developed dance curriculum, and some schools of fine arts include programs in dance, but the professional instruction of the dancer, whether in ballet or modern dance, usually is carried on in the 19th-century tradition of teaching in schools associated with ballet companies or opera houses. It has changed little from the time when Edgar Degas painted the young dancers of the Paris Opéra at practice. With few exceptions, training for the dance has remained a relatively narrow and physically demanding regimen.

The technologically based arts. Relatively little attention has been paid to the development of professional curricula in photography. More than any other art, it has been treated as a hobby to be learned without formal instruction. Where it is taught, the focus is on photographic technique rather than on photography as an art. There are, however, signs of a growing acceptance of photography as an important and expressive art form, and it is now becoming part of the curriculum not simply of schools of journalism but of schools of fine arts as well.

The new arts of the 20th century—film, radio, and television—still depend largely upon people trained in the older arts. The foundation in recent decades of film institutes associated with universities, notably in New York state and California, has been a significant step. A few radio and television programs exist, sometimes as parts of a department of communication arts or outside academic settings. Most of the directors, actors, and writers in film and television either have developed in the theatre or the literary world or have grown up in the industry without formal instruction.

THE SELF-TAUGHT ARTIST

With all of the many systems and institutions for formal training in the arts that have come into being over the centuries, there always have been self-taught artists. This is especially true of writers, the overwhelming majority of whom, from Homer to the present day, have had no specific training in their art and many of whom, in fact, like Shakespeare or George Bernard Shaw, had comparatively little schooling of any sort. They have learned their trade by studying other writers and life itself, and by proceeding from imitation to original creation. Many people have turned to writing after education for careers in other fields. Medicine alone has produced such outstanding writers as the Englishmen Sir Thomas Browne and W. Somerset Maugham and the American William Carlos Williams.

Art without schooling

Self-trained artists have also achieved prominence in other fields, though far less frequently than in writing. The history of architecture, perhaps of all the arts the one that seems most to demand strict technical training, contains two spectacular instances of the self-trained professional. Sir Christopher Wren, among the greatest of English architects, was a mathematician and professor of astronomy at Oxford before he took up architecture. But whereas Wren was grounded in the principles of the exact sciences, Sir John Vanbrugh, the architect of Blenheim Palace, had been a soldier and a dramatist before he turned to architecture at the age of 35.

The completely self-trained professional artist is a rarity in painting and sculpture. More than any other kind of artist, the painter especially must learn from and with other painters. Although such painters as Paul Cézanne, Vincent Van Gogh, and Paul Gauguin were in many respects self-taught, they absorbed a great deal from studies of, and interactions with, their fellow artists. Occasionally a true "primitive" such as Henri Rousseau, unhampered by conventional instruction, has produced a style with extraordinary power and directness as well as naïveté.

TRAINING OUTSIDE THE WEST

In non-Western cultures also, professional training in the arts has had a long history. Much of it has been carried out under systems of apprenticeship similar to those found in the West. Both in the Orient and in Africa, the profession of the arts has often been a hereditary one, sometimes, as in the case of bronze casters and ceramics workers in China, with secret techniques passed on from one generation to another. Similarly, Japanese Nō and kabuki actors long passed their craft from father to son, with the aristocratic Nō being regarded for centuries as a secret tradition. In West Africa, an entire village frequently has specialized in the production of a particular kind or style of art object.

Oriental and African art practices

Education in the arts has also been diffused through written treatises. In China a number of practical manuals on painting are extant, the two most famous ones dating from the 17th century. In India there were noted teachers of arts and crafts as early as the 1st century. Most recently, art schools similar to those in the West have been founded in a number of Eastern countries, notably in Japan, where a conscious attempt has been made to combine ancient traditions and modern styles. During the period of French colonial rule in Southeast Asia, an influential École des Beaux Arts was set up in Hanoi. In many primitive and folk societies, however, such arts as dance have been passed on by imitative learning.

Conditions of work in the arts

In no way can artists be stereotyped as shivering recluses, painfully pursuing their craft in dimly lit garrets; as well-fed hangers-on to the coattails and whims of the wealthy; or as irresponsible wastrels, lurking in the outskirts and subcellars of respectable society. The conditions of their work and life have been shaped in large measure by the regard of their society for the arts in general or for their art in particular. High regard has not always brought success or satisfaction, however, and artists have turned into many different avenues in search of tangible or other rewards for their art.

THE STATUS OF THE ARTIST

The social and intellectual status of artists has differed considerably from one field to another, from one culture to another, and from one historical period to another. They have been regarded as entertainers, artisans, seers and prophets, and, often, as dangerous cranks or misfits. The history of each of the arts has been marked by a determined effort on the part of its practitioners to raise their status and to be accepted as honoured members of their society.

Prehistory. In preliterate cultures, the poet preserved and transmitted the beliefs and traditions that gave the culture its sense of identity and purpose. Less an original creator than a storyteller and singer, he kept alive the works that had grown up by gradual accretion and refinement, and often he contributed to the process. The Homeric epics of ancient Greece probably developed in this way over generations, though in the form in which they have survived they almost certainly were shaped by one or more supreme geniuses.

The primitive bard

The primitive bard seems to have occupied an anomalous position, held in awe because of his seemingly divine powers but without a fixed place in the social hierarchy—existing as a dependent of a tribal ruler or as a wandering beggar. It is no accident that the Greeks represented Homer as blind, thereby symbolizing not only the poet's inward-directed vision but also his separation from ordinary men. The peculiar contradiction that poets have complained of bitterly, that their gifts are venerated but that they themselves are neglected, has its most famous statement in the couplet, "Seven cities warred for Homer being dead,/Who living had no roof to shroud his head," by the English poet Thomas Heywood.

Antiquity. In the great age of classical Greece and, later, of Rome, the artist lost some of his legendary quality as seer, but he gained considerably in the area of social acceptance.

Greek poets, sculptors, and designers. The skill of the dramatist, like that of the athlete, was tested in competition and rewarded by prizes. The bard gave way to the historically defined individual. Sophocles, in the fifth century before Christ, was no blind beggar but a respected citizen, a soldier, and politician, as well as a tragic poet. Above all, the intellectual role of the writer was recognized and his claim to respect was validated by the celebrated

Emergence of the known artist

dictum of Aristotle that poetry is more philosophical than history. Aristotle's defense of poetry was perhaps a deliberate response to the celebrated attack by Plato in *The Republic* on the creative artist as a potential threat to the stability of the state.

In other fields, too, artists began at the same time to emerge from anonymity and to achieve social acceptance. The nameless artisans and stonecutters who erected the first crude Doric temples and carved the cult images that they housed were succeeded by renowned architects such as Ictinus (flourished 5th century BC) and Callicrates and by master sculptors such as Phidias (flourished 475–430 BC), who supervised the rebuilding of the Athenian Acropolis. Phidias is alleged to have represented himself and the Athenian tyrant Pericles on the shield of Athena. Such self-glorification would have been beyond the imagination of earlier generations.

The great mass of Greek sculptors of the classical and Hellenistic periods, however, continued to be regarded as craftsmen, paid standard wages for a day's work or for a specific piece. To Plato, for example, sculptors were common workmen. Even the recognized masters were expected to turn their hands to whatever their patrons demanded.

Rome and service to the state. Under the late Roman Republic and the empire, the artist continued to enjoy a favoured status. The great Roman dramatist Terence (flourished 2nd century BC) began his career as a slave, but his achievements earned him freedom and admission to the leading intellectual circles of his time. The organizing genius of the Romans enlisted poets, architects, and sculptors in the service of the state. Gaius Maecenas (died 8 BC), the first of the wealthy patrons of the arts, encouraged and supported such poets as Virgil and Horace. Vast building and engineering projects throughout the empire gave employment and prestige to many architects.

This was the age, too, that saw the emergence of the art connoisseur and the tourist in search of cultural treasures. The esteem for artists in Rome is reflected in the eagerness with which Pliny the Elder sought out and recorded biographical information on the great artists of Greece. For the first time the artist was felt to merit the same attention as the statesman or soldier. A similar rise in the prestige of the writer was signalized by the founding of such great libraries as the one at Alexandria, devoted to preserving the literature of the past.

The musician of antiquity. Much is known about the musician in antiquity, although practically none of the music itself has survived. There is considerable testimony to show that music itself apparently was considered among the highest of human accomplishments; in fact, of divine origin. The legend of Marsyas, the human being who dared to challenge Apollo to a musical contest and was flayed for his pains, was a favorite subject of Hellenistic sculpture. The myth of Orpheus, whose playing on the lyre could move even inanimate things to wonder and delight, reveals the almost superstitious awe in which the Greeks held the powers of the musician. Musical theory, with its close relationship to mathematics, was regarded as a branch of philosophy. The philosopher Pythagoras (flourished *c.* 530 BC), who saw the whole universe as a harmony of the spheres, was only one of many who gave the highest intellectual ranking to the study of music. The special status given to music in antiquity continued into the Middle Ages, when music alone, of all the arts, was ranked among the seven branches of learning.

Music as the highest art

In Greece, there emerged for the first time the distinction between the gentleman amateur and the paid professional that was to play so large a role in the later history of the arts. Every educated Greek was expected to be able to play an instrument, to sing acceptably, and to discourse on the theory of harmony. On a much lower intellectual and social level were the professional entertainers who competed for prizes at great public concerts, who performed in the dramas, and who supplied the musical accompaniment for athletic games.

Actors, dancers, and rhetoricians. The distinction between the creative artist and the entertainer or interpreter, which continues to be made even today, seems already to have existed in ancient civilization. The accomplishments of the actor-dancer were admired and applauded, but, aside from the citizen-dancer of the Greek festivals of tragedy, he had a relatively low intellectual and social status. The names and other details of some of the famous actors of this period are known, but references to them are in a tone altogether different from references to a Phidias or a Virgil, and everything indicates that actors were regarded as belonging to an inferior class.

One art virtually unpracticed today, that of the orator or rhetorician seeking to persuade, by speech or writing and according to a detailed formulistic pattern, occupied an especially favoured place in antiquity. The principles of the art were expounded in treatises by Aristotle, Quintilian, and others, and its greatest practitioners, Demosthenes among the Greeks and Cicero (first century before Christ) among the Romans, were accorded the same respect given to the poets. In the twilight of the Roman Empire, the young St. Augustine supported himself as a teacher of rhetoric. The prestige of rhetoric as a noble and useful art reached its highest point during the Renaissance, when every court had its master rhetoricians and orators.

The lost art of persuasion

The Dark and Middle Ages. With the collapse of the Roman Empire in the West in the 5th century, the professional artist virtually disappeared in Europe. Until the revival of Western culture under Charlemagne about 800, the practice of the arts was confined mainly to scattered monastic enclaves, which kept alive the traditions of a written literature and began to develop a new architectural style. Some of the arts, such as the written drama, were lost altogether, only to be revived later from church ritual and folk ceremonies. The sophisticated poet, a product of the stable and assured civilization of antiquity, vanished from the scene to be succeeded by anonymous bards, such as the creator of the Anglo-Saxon *Beowulf,* who, as in Homeric days, put into final form the results of centuries of oral shaping of legend and history. Only in the Byzantine Empire and, after the 7th century, in the rapidly growing world of Islām, was the professional artist able to flourish.

The rich burgeoning of medieval civilization following the early 9th-century Carolingian renaissance produced a magnificent harvest in the arts, but for the most part the creators remain unknown. The superlatively beautiful so-called Book of Kells from 9th-century Ireland was decorated, page by page, by a monk or group of monks who saw the task as their gift to God, their way of living their vocation. It is doubtful that they were conscious of themselves as artists or that they were so regarded by their fellows.

The builders and sculptors of the great Romanesque and Gothic churches were journeymen masons and carvers who worked under the driving direction of a forceful churchman, such as the remarkable Frenchmen Hugh of Semur, Abbot of Cluny from 1049 to 1109, and Abbot Suger, who built the Abbey Church of St. Denis from 1140 on. The radiant stained-glass windows of the Gothic cathedrals were made in the workshops of Chartres and other centres. Only rarely did a particular craftsman, such as the sculptor Giselbertus of Autun or the sculptor-decorator Nicholas of Verdun, emerge from anonymity.

The Renaissance and Baroque. An enormous increase in the prestige of the artist came about with the beginnings of the Renaissance in 14th-century Italy. The veneration for antiquity that so dominated intellectual life was in large measure a veneration for the arts of antiquity, for a life-style in which the written and spoken word, the visual image, the public monument, and those who created them were seemingly central elements.

Reverence for the artist

New stance of the poet. The crowning with laurel of the poet Petrarch at Rome in 1341 was a ceremony intended to establish a link with ancient civilization. Almost immediately after the death of Dante, *The Divine Comedy* became the object of an intense scholarly and critical study equal to that given the great epics of antiquity.

Throughout Italy, but especially in Florence, the humanist of the 15th century was dedicated to the purification and preservation of Latin as a living universal language. He gave to the profession of writer and scholar a distinction it had not had since the collapse of Rome nearly a millennium earlier. Men of letters were held in an esteem

that—in the case of that shown the Dutch humanist Desiderius Erasmus—approached reverence.

Regnancy of the visual arts. The new Renaissance attitudes were most striking in the visual arts. The painter and architect Giorgio Vasari, in his *Lives of the Most Eminent Architects, Painters, and Sculptors* (1550), saw the development of the arts in Tuscany as the working out of a divine plan. According to Vasari, God sent Michelangelo to the world endowed with so great a universality of power in each art that he might be considered of a divine rather than human nature. A century earlier, no one would have considered artists deserving of extended biographical treatment.

The artist had fought for this new status, however, from Filippo Brunelleschi, who stoutly defended his independence as an architect from the meddling of the Florentine government, to Michelangelo, in his defiance of Pope Julius II. Leonardo resented bitterly the placing of painting among "the mechanical arts" and attempted to raise the status of painting by deliberately downgrading sculpture, which he called a sweaty and fatiguing job for workmen. Even after he had become a world-famous artist, Michelangelo felt it necessary to defend his social and intellectual position. Forbidding his nephew to address letters to him as "Michelangelo the Sculptor," he insisted he had never been a painter or sculptor such as those for whom it was a business.

Status and the academy

The first academies of fine art were founded not primarily as teaching institutions but as means of enhancing the standing of artists. Florence's Accademia del Disegno, founded by Vasari in 1563, was headed jointly by a prince and an artist—the Grand Duke Cosimo and Michelangelo. In 1564 the academy publicly demonstrated the honour being paid to artists in its elaborate funeral ceremonies for Michelangelo. Such leading Venetian artists as the painters Titian and Tintoretto and the architect Andrea Palladio were glad to apply for membership in the Florentine Academy, and its advice on the design of the Escorial Palace was requested by King Philip II of Spain.

Particularly in northern Europe, however, the average painter, as distinct from the outstanding genius, continued to be regarded as a craftsman throughout the 16th century. In 1590 the Guild of St. Luke in Haarlem included not only painters, wood-carvers, and goldsmiths but also printers, slate layers, plumbers, and lantern makers. This extreme situation was bitterly complained of by a leading Dutch painter and theorist of the period, but it illuminates the attitude toward painting still held by many.

In the 17th century the intellectual and social status of the artist reached perhaps the highest level it has ever attained. Two of the greatest Baroque artists, the Flemish painter Peter Paul Rubens and the Italian sculptor Gian Lorenzo Bernini, lived like princes. Both were accepted in leading intellectual circles and were completely at home in this environment. The Spanish master Diego Velázquez enjoyed lifelong security as the favourite painter and friend of King Philip IV, who, despite the grumbling of many aristocrats, awarded him the Noble Order of Santiago.

Writers and performers. The change in the status of the writer during the Renaissance and Baroque periods was not nearly as spectacular. The poet and rhetorician were treated with respect, but it was still necessary for Sir Philip Sidney about 1581 to write *An Apologie for Poetrie* (first published in 1595), and the poet Edmund Spenser, like many poets of lesser stature, had to beg for patronage from powerful nobles. Shakespeare published his narrative poems with obsequious dedications to his patron, the Earl of Southampton.

Practitioners of such newer literary forms as the novella and the popular drama never achieved full social or intellectual acceptance in this period. The dramatist, in particular, linked as he was with the day-to-day commerce of the theatre, and often himself an actor, was considered simply as a hack writer. His plays were regarded lightly as literature, and when Ben Jonson dared to publish his plays as his *Workes* in 1616, he was jeered at for his pretensions. Actors, though they were loved and admired as entertainers, had escaped only recently from being classed as "rogues and vagabonds," and technically they were regarded as household servants of the nobility. An occasional actor, such as Edward Alleyn or Shakespeare, was able to attain the position of a gentleman if he was a manager as well or had powerful patrons.

The theatre as mere entertainment

Only in the France of Louis XIV did the dramatist attain the status of an honoured man of letters, and then only by adhering to the rigidly defined neoclassical standards of dramaturgy. The drama, like all the arts, was closely integrated into the political structure, and powerful ministers of the crown carefully supervised the work of the playwright. Leading dramatists such as Jean Racine were victims of political intrigues, but if they survived them, they could look forward to the ultimate reward of election to the Académie Française. Molière, however, who was an actor and theatre manager as well as a dramatist, was regarded socially as a member of Louis's household.

Revolutionary impetus. The great social and political changes of the late 18th century, climaxing in 1789 in the French Revolution, were as decisive for the artists as for other groups of men. The emphasis on personal freedom and the desire to break loose from mind-forged manacles as well as social shackles were key elements in the Romantic movement that dominated the arts for almost 100 years. A wide gulf separated the attitudes of Haydn from those of Mozart, though the two were only a generation apart in age. Haydn was content to serve as music master for the Esterházy family in Hungary from 1760 to 1790. Mozart could easily have made a similar career, but he rebelled against being a servant to the Archbishop of Salzburg. He determined to earn his living as an independent musician, and for a time his career flourished. The day of the free-lance composer had not yet dawned, however, and he died early, in poverty and loneliness.

For the first time in history, under the impetus of the French Revolution, there emerged the phenomenon of the revolutionary artist committed to an active political role. The most conspicuous example was the Neoclassical painter Jacques-Louis David, who not only expressed the ideology of the Revolution in his works and organized its great public festivals but was a political leader and a friend and ally of the Revolutionary leaders Marat and Robespierre. The ambivalent attitude expressed in the 20th century toward political activity by artists was illustrated earlier when David was caught up in Robespierre's downfall in 1794. Unlike other of Robespierre's followers who were executed summarily, David, after a stay in prison, was pardoned and restored to favour as virtual dictator of the arts. A similar situation befell the Spanish painter Francisco Goya, who had partially cooperated with the French puppet government in Spain. When the Napoleonic armies were driven out in 1814, he regained his position as court painter. There may have been in both cases an unwillingness to dispense with the services of a great artist. In addition, however, there also may have been at work the feeling, expressed even in this present century, that the artist is not to be taken altogether seriously as a politician.

The artist as political activist

The developing setting of modern art. A characteristic feature of artistic activity in the 19th and 20th centuries is that no generalizations about the status of the artist can be made. Trends are divergent and often completely contradictory. On the one hand, the artist became one of the representatives of the acquisitive society, a successful businessman satisfying the demands of the market. Whether he was a man of genius, such as the English novelist Charles Dickens and the French novelist Honoré de Balzac, or a talented and industrious hack, he wrote for a vastly expanded audience created by mass literacy and growing leisure. The English writer Anthony Trollope, sitting in his club and turning out his fixed quota of words, finishing a novel one day and beginning another the next, was a perfect embodiment of the respectable, punctual, middle-class writer.

Similarly, the official artist was recognized by the national academy and hung at the salon exhibitions. He created a standardized and approved product—whether portrait, landscape, or sentimental illustration—to meet the great demands of a newly rich clientele eager for culture but unsure of its tastes. The American portraitist John Singer Sargent not only answered completely the artistic needs

of his upper class patrons but identified with them as a person. At this point the social distinction between patron and artist virtually has disappeared.

On the other hand, the 19th century was above all the period of the alienated artist, the deliberate exile from society. The tone was set early in the century by the English poets Lord Byron and Shelley, both born aristocrats, who flouted the standards and conventions of their class. The contempt for accepted social behaviour and the adoption of the pose of the outcast are seen at midcentury in the French poet Baudelaire and in the closing decades in the French painter Gauguin. *Épater la bourgeoisie,* "to dumbfound the middle class," became the slogan of a whole group of French Romantics. The doom-ridden artist propelling himself toward destruction by drink, drugs, or a furious excess of behaviour became a characteristic figure of modern times, from Edgar Allan Poe to Jackson Pollock and many "pop" singers of the 1960s and 1970s.

Emergence of the alienated artist

Somewhat related to this phenomenon was the distrust with which the middle class came to view artists, above all painters and musicians. This is especially true of attitudes toward the avant-garde artist, whose rejection of academic conventions within his own art makes him suspect as an enemy of society itself. Napoleon III's superintendent of fine arts summed up for all time the attitude of officialdom toward such artists when he expressed his displeasure and disgust with their work, characterizing them as democrats who do not change their linen and hope to put themselves over on the world.

Impact of scientific dominance. As science became a dominant force, the intellectual prestige of the artist declined. The role of the prophet continued to be played by such writers as Thomas Carlyle and John Ruskin, both of whom were skeptical of the benefits of science and technology. No creative artist since Goethe, however, has gained the intellectual respect accorded to scientists. Modern art has produced no men of universal interests and abilities to match such Renaissance geniuses as Leon Battista Alberti or Leonardo, whose work and thought carried them into virtually every area of human activity.

The ideological gap between scientists and artists, between "the two cultures," as the English novelist C.P. Snow has called them, has widened dangerously over the past century. A number of artists, however, attempted to incorporate into their own work some of the methodology of science and to gain for themselves the intellectual status of the scientist. In a famous essay of 1880, *The Experimental Novel,* Émile Zola argued for a literature governed by science and claimed for the novelist the function of a biologist of society. The French Postimpressionist painter Georges Seurat, who set for himself the reconciling of art and science, believed that properly applied scientific theory could replace intuition as the basis of art. In the 20th century, apologists of Cubism, though not its leading practitioners, saw in it the artistic expression of the fragmented world of modern physics. Recently, electronic music, created mainly by university-based composers, has provided a new union of technology and art.

New areas of respectability. The actor-manager David Garrick had helped to make the theatrical profession respectable, but during the early 19th century, actors and actresses, however applauded or financially successful, continued to be regarded by many as slightly disreputable. By the end of the century, however, the situation had so changed that in 1895 Queen Victoria knighted Henry Irving, the first actor to be so honoured. Such an accolade became almost commonplace in following years, and in both Europe and America the actor is now accorded considerable respect.

The new arts of the 20th century—cinema, radio, television, and recording—have all produced their mass idols. The fantastic adulation lavished on film stars or pop singers, however, has often been accompanied by an openly or covertly expressed intellectual contempt. Of all the artists involved in the creation of a film, only the director seems so far to have achieved intellectual respectability. Such directors as the Italian Federico Fellini, the Swede Ingmar Bergman, and the Frenchman Jean-Luc Godard became subjects of a serious aesthetic discussion comparable to that focussed on the most important of contemporary writers or visual artists.

Artists of the new arts

In music, the last 100 years have featured the virtuoso performer, the pianist, the violinist, the soprano, or the conductor. The separation of composer from performer began early in the 19th century, and instead of a Bach, Mozart, or Beethoven performing his own works, there is the virtuoso who interprets the compositions of others. The most publicized musician of the 20th century and the one most widely admired for his intellectual gifts was not a composer but the conductor Arturo Toscanini. The famous pianist Ignacy Paderewski, when he served briefly as premier of Poland in 1919, became the first professional artist to head a national government. One of the most striking phenomena of the rock music of the 1950s and 1960s was the closing again of the gap between composer-lyricist and performer.

Another important development in this century was the growing activity of black artists and women artists to break down the social and intellectual discrimination that historically has been directed against them. There has been, as a result, not only increased public knowledge of the contributions made by both groups but also a greater understanding of the artist as an individual who is directly involved in—and in much of his work reflects—the problems and stresses of his society.

THE ARTIST'S LIVELIHOOD

Throughout history, the economic status of the artist has varied as widely as his social and intellectual status, though changes in the one have not necessarily paralleled changes in the other. Since conditions favouring one art may have been unpropitious for another, artists in different fields often did not experience the same measure of economic security at any given time. Similarly, though artistic activity in general has followed the economic development of a society, there have been periods in which a society has experienced great prosperity while at the same time artists have starved.

Forms of patronage. Until recently, artists in most fields were directly dependent upon patronage—whether governmental, church, or private—for their livelihood. They produced works specifically commissioned for a particular need or occasion or designed to appeal to the tastes of a well-defined group or known individual. The notion of an artist working to please himself and then attempting to sell his product on the open market, either directly or through an intermediary, is a comparatively modern one. In such fields as architecture and serious music the artist even now remains almost completely dependent on direct commissions.

The system of patronage has had a profound effect on the work of art itself. The idea that institutional sponsorship always produces bad works of art is refuted by the facts. The buildings on the Acropolis of Athens, the Gothic cathedrals throughout Europe, the Sistine Chapel frescoes in Rome, and the sculptures of the Benin kingdom of West Africa, for example, were all commissioned by a governmental or religious body. Artists working to express a shared system of ideas and beliefs for a community of which they are an integral part are likely also to share a common and stable style. The rapid shifts of style that characterized the arts in modern times reflect, to a certain extent, the new economic status of the artist as well as the present instability of artistic traditions.

Quality and patronage

The open marketplace. Nevertheless, many artists of the past have chafed at the loss of individual freedom entailed by institutional support. This was particulary true during the 16th and 17th centuries, when the intellectual and social status of the artist was improving. When artists gradually freed themselves from the tyranny of patronage, they often discovered that they had submitted to the equally confining tyranny of the marketplace. In 17th-century Holland, for the first time in the history of art, painters began to produce works to be sold at auction or in art markets. The dependence on commissions did not cease, but direct patronage was no longer the only source of the artist's income. Most of Rembrandt's work, for example, was probably done to order, but many of

his surviving works, including more than 60 self-portraits, probably were done without immediate commission.

In the 19th century, a number of artists attempted to capitalize directly on their own work, not by selling individual paintings to particular buyers but by exhibiting them for a fee to a mass audience. The American artist Rembrandt Peale earned $9,000 from the showing of his painting "The Court of Death," which was seen by 32,000 people during a 13-month tour in 1820–21. At virtually the same time, the French painter Théodore Géricault was exhibiting his famous, "The Raft of the Medusa" throughout England and Ireland. In 1855, Gustave Courbet constructed his own Pavillon du Réalisme at the Universal Exposition in Paris, charging the public an entrance fee to view 50 of his paintings.

In the other arts as well, the artist emerges as an individual entrepreneur from the 16th century. Alexander Pope was perhaps the first poet in history to earn a living entirely from the public sale of his work. The new commercial orientation of the writer was summed up in the famous aphorism of Samuel Johnson, "No man but a blockhead ever wrote, except for money."

Institutional patronage has never, even in modern times, lost its importance for the artist. Eugène Delacroix, the supreme French Romantic painter, accepted a number of commissions for large murals from the French government. One of the most celebrated paintings of the 20th century, the "Guernica" by Pablo Picasso, was done for the Spanish Pavilion at the Paris Exposition of 1937. Both Marc Chagall and Henri Matisse created masterpieces of religious art on commission.

Artists in the modern period have had, nevertheless, to appeal to the general public in order to survive. For those who succeeded, the rewards were great, and many artists, both academic and avant-garde, became wealthy. On the other hand, innumerable other arguably good artists have been driven to despair by lack of public recognition and support.

Governmental and other subsidies. Various attempts have been made to provide some kind of governmental subsidy for artists. Among the most conspicuous of these were the Works Projects Administration (WPA) programs for painters, writers, actors, and other artists supported by the United States government during the depression of the 1930s. These projects were attacked bitterly by some politicians and eventually discontinued.

State-supported performing companies

Today, the principle of government support for the arts is widely accepted. Most European countries have state theatres and opera houses, and they support symphony orchestras, ballet companies, and other groups of artists. The British government grants sizable subsidies to the National Theatre and the Royal Shakespeare Company. In the United States, the federal government and some state governments have given financial aid to artists through such organizations as the National Endowment for the Arts and the New York State Council on the Arts. Universities and some private foundations have become important sources of economic support for the artist. Poets, painters, and composers in residence are now found in many universities, sometimes with teaching duties, sometimes free simply to create.

Conditions outside the West. The social forms of non-Western cultures have varied so widely, from place to place and from period to period, that it is impossible to make any valid generalizations about the economic position of the artist in such cultures. At one end of the scale, in simple nomadic or agricultural communities, the artist hardly has existed as a specialized person. Such crafts as pottery making or weaving generally have been diffused throughout the community, sometimes limited to one sex or another. In the sophisticated and highly developed imperial, monarchial, or tribal societies of China, India, or Africa, on the other hand, the trained artist has always been regarded as a valued servant of the ruling or priestly group and has been integrated completely into the economic structure of the state. With the vast social and political upheavals in the Orient and in Africa in recent decades, a period of rapid change in the economic and social status of the artist has set in.

INTERACTIONS AMONG ARTISTS AND WITH THEIR PUBLICS

The mingling of artists working in the same or different mediums always has provided a leavening element for their creative energies and imagination. Similarly, collective or collaborative work, criticism and appreciation, and national and international awards, as well as collective action for the benefit of the artistic community, have contributed to artistic life and work.

The artist's working environment. The act of artistic creation is usually a solitary one, whether carried out in an isolated cell or in a crowded room. Most artists, however, have sought the stimulus of some special kind of environment, generally one in which they could have regular interaction with their peers. Only rarely has an artist who is cut off from others, by necessity or choice, been able to produce great work. It has been suggested that the American poet Emily Dickinson might have developed her poetic gifts even more significantly if she had not been so isolated from other writers. The battles of wit between Shakespeare and Ben Jonson at London's Mermaid Tavern may be apocryphal, but they describe a situation typical of the literary life.

A deliberately and formally organized artistic environment has proved only rarely to be conducive to creative work. Art colonies, either specifically limited to artists or existing within Utopian communities, have tended to fall apart rather quickly. One of the longest lasting of these, the MacDowell Colony at Peterborough, New Hampshire, has endured mainly because it offers a pleasant and quiet summer retreat for the artist. Brook Farm in Massachusetts, which Nathaniel Hawthorne joined briefly in 1841, and the Helicon Home Colony in New Jersey, founded by Upton Sinclair, were both short-lived.

The university has come to play an important role as a fostering environment for the artist. The many temporary or permanent positions it offers have provided not only a measure of economic security for artists but also an intellectually stimulating setting. There is some feeling, however, that permanent immersion in an academic environment may insulate the artist too much, eventually depriving him of his vital sources of inspiration.

Centres and schools of artistic activity

Historically, most artists seem to have thrived best in a city that is the centre of vigorous political, economic, and intellectual activity and in which they could interact with other artists when and where they chose. Florence in the 15th century, London in the Elizabethan era, Vienna in the late 18th and early 19th centuries, and Paris from the Revolution until World War II were such places. The Café Guerbois in Paris was the scene of innumerable heated discussions in the 1860s and the 1870s between many of the leading painters and writers of the period. Such interchanges undoubtedly provided much creative impetus for their art.

Sometimes, artists brought together by similar views have mutually inspired each other and worked in tandem to develop new styles. This was true probably of Giorgione and Titian in Venice, and certainly of Georges Braque and Picasso during the elaboration of Cubism in the early 20th century. There have also been looser groupings of artists with common interests, such as the Impressionist painters who exhibited together from 1874 to 1886. More recently, such groups as the Dadaists of immediately after World War I and the Surrealists of the 1920s for a time united various artists of sometimes diverse tendencies.

Artists have on occasion, particularly during the 19th century, combined out of a sense of shared poverty and neglect to express a common contempt for accepted values. They have formed a separate society of their own, a "bohemia" characterized by deliberately outrageous costume and behaviour. The ultimate expression of the bohemian disdain for middle class conventions was the glorification of suicide, and young French bohemians actually founded a Suicide Club in 1846 as a gesture of defiance.

Social integration outside the West

Such bohemian excesses generally have been unknown in non-Western cultures. In societies where the role of the artist has been clearly defined and established by long tradition, the notion of the alienated artist would be virtually incomprehensible. Like the European craftsman of

the Middle Ages, the non-Western artist has functioned not in a special environment set apart from his society but as a respected member of the community. Here again, of course, as in other aspects of the profession of the arts, the very recent period has begun to see a breakdown in traditional values and relationships.

Multiple creators. Related to the question of the artistic environment is that of group creativity. A collective approach has proved far less viable in the arts than in such fields as scientific research. In the performing arts, film, and architecture, however, cooperative activity has been not only possible but often necessary, although usually there is one guiding vision—that of the director or the chief architect, for example. Many great buildings, such as St. Peter's Basilica in Rome, represent the effort of a number of different architects. Almost always, however, these architects have been in charge successively rather than at the same time.

Collaboration

More successful in most arts has been the collaboration of two individuals. The Parthenon was the product of the architects Ictinus and Callicrates. Collaboration was a common procedure among Elizabethan dramatists, a most notable example being the works of Francis Beaumont and John Fletcher. A collaboration between composer and librettist, such as that between Mozart and Lorenzo da Ponte, created many great operas.

Art collectives have been organized, with varying degrees of success, in many of the Socialist or Communist states, particularly in the immediate post-Revolutionary periods. In such countries as China, North Vietnam, and Cuba, deliberate emphasis has been placed on "the democratization of art," an attack on the notion of art as the province of the individual genius, and an attempt to create a collective and anonymous art. Numerous political events, such as the 1968 student strike in France, stimulated collective activity in the production of propaganda posters and the organization of guerrilla-theatre companies that gave informal but impassioned performances usually on social or political themes.

Guilds and unions. Artists have also banded together for very practical purposes. It was, for instance, during the later Middle Ages and the Renaissance that European artists, like all other craftsmen, were organized into guilds that looked after their economic interests and regulated trade procedures. The bronze workers of the great African kingdom of Benin, which flourished in the 16th century, also had their guild.

Attempts to organize creative artists into modern trade unions generally have failed. During the depression in the United States, artists' and writers' unions enjoyed a brief period of growth, and during the 1960s black artists and women artists formed their own groups. Performing artists, in contrast to creative artists, have formed strong unions that have been able to exert considerable pressure to better their economic situation. American Actors' Equity Association and British Actors' Equity Association are probably the best known of these unions, and there are similar organizations in radio and television and in the film industry. On an international scale, there have been various attempts to bring together artists for political or economic activity. International PEN (*P*oets and playwrights, *E*ssayists and editors, and *N*ovelists) is an organization of writers that has taken vigorous public positions on matters of concern to professional men and women of letters. Founded in 1921 and actively supported in its early days by such prominent figures as H.G. Wells, Bernard Shaw, Anatole France, and Thomas Mann, the organization includes today more than 8,000 writers from 58 countries.

Impacts of criticism and appreciation. Among the necessary conditions of work for most artists are public criticism and appreciation. Although artists often have pretended to be scornful of critics, generally they have flourished best in an atmosphere in which their work was studied, understood, and encouraged. Almost every important artistic movement of modern times has had its critical spokesman and defender. Conspicuous among such champions have been the Frenchmen Émile Zola, for the Impressionists, and Guillaume Apollinaire, for the Cubists, as well as the American critic John Martin, for the modern dance movement.

Many creative artists have been discouraged, however, by lack of appreciation of their work. The great Baroque architect Francesco Borromini committed suicide at least partly because of critical neglect and attack. A legend that the English poet John Keats died because of savage criticism of *Endymion* is false, but Keats was affected seriously by the attacks. Cézanne, in his later years, was deeply hurt by what he thought was complete public disregard of his achievement.

National and international honours

Like individuals of distinction in other fields, artists have been singled out for awards of various sorts. The most prestigious of these honours is the Nobel Prize for Literature, which, with two exceptions (1914 and 1943), has been awarded annually since 1901. The prize has sometimes been refused, notably because of political pressure, but the writers chosen usually have regarded it as the climax of their careers. There is no comparable award in the other arts.

In the United States, the Pulitzer Prizes annually honour outstanding work in drama, fiction, poetry, and musical composition, as well as in scholarly and journalistic writings. The National Book Awards also honour literary work in the different genres.

In other countries, artists are given prizes or other marks of merit for distinguished achievement. The Prix Goncourt is the best known of French literary awards. Many creative and performing artists appear each year on the honours list in Great Britain, as recipients of knighthoods or other distinctions. In the Soviet Union, the title of Honoured Artist is conferred as a badge of accomplishment, and in Japan, actors are often honoured with such titles as National Living Treasure, indicating the prestige they have acquired.

Artists are also eligible for election to honorary societies, some of which are limited to the arts. The greatest tribute that can be paid to an artist or intellectual in France is elevation to membership in the Académie Française, the famous company of "immortals." In the United States, the National Institute of Arts and Letters constitutes the chief honorary society of the arts.

THE ROLES OF THE AMATEUR

Amateurism has always been an important factor in the development of the arts. In the oldest sense of the term, the "amateur of art" is simply the lover of art. He need not necessarily be a practitioner of any of the arts himself, functioning instead to encourage and support the arts in every way possible. What distinguishes this kind of amateurism from passive art appreciation is its active commitment. From its ranks come the students, the connoisseurs, and the patrons of art.

There is also a more common sense in which the term amateurism applies to the arts. There are vast numbers of people who, without special talent or extensive training, write, paint, or play musical instruments primarily for their own pleasure or as a leisure occupation. Often such amateurs band together in musical groups or theatrical companies that perform for themselves and their friends. The line of demarcation between amateurs of this kind and professionals is usually clear, although their economic contributions in the form of royalties are often important for playwrights, film performers, and the composers of popular music. Few Sunday painters, unlike Gauguin, ever abandon their businesses and families to risk everything on their talents.

The idea of the gentleman artist

Another kind of amateurism—rare today, though historically it has been of great significance—goes back to an idea that appeared first in antiquity and acquired force during the Renaissance: namely that proficiency in an art, particularly music or poetry, is an (ennobling) attribute of the gentleman. Art is valued as an accomplishment of the educated man, but not as a profession. Unlike the art amateur who practices an art simply for his personal enjoyment and basically is not concerned with how well or how badly he does it, the gentleman amateur aims at excellence. His ideal is the quality of *sprezzatura,* or effortless grace, described by Baldassare Castiglione in his

Courtier (1528) as one of the chief characteristics of the Renaissance courtier.

Chinese culture always has been unique in considering painting as the special avocation of the noble amateur. The man of letters, the statesman, and the upper-class gentleman have traditionally been trained in and expected to be proficient in the art of painting, and especially in calligraphy. In India, too, training in the arts has been considered a necessary part of the education of members of the highest caste.

The amateur in all of these senses has made invaluable contributions to the preservation and dissemination of art. The devotee of art has encouraged and supported artists, has collected and handed on works of art, and has founded museums, libraries, and schools for the arts. Above all, he has set a standard of taste in every generation, sometimes a false or artificial one but one that often has raised the level of artistic performance. If the skilled amateur at times has introduced a measure of snobbism into the appreciation of the arts, he has compensated for this by his defense of art against those who have attacked it as immoral, dangerous, or frivolous. Most of all, however, artists are indebted to the millions of amateur practitioners of art who, having experienced its value for themselves, can respond to it as practiced on the highest professional level. (Si.T.)

Social and economic aspects of the arts

It is necessary to begin a sociological analysis of the arts by identifying the various social frameworks within which artistic activities have been conducted and the influences that these frameworks have had on the style and content of the arts, the levels of creative attainment, the mode of living of the artists, and the uses to which their art has been put by society. This mode of analysis is not concerned, as the histories of the various arts are, with describing how the particular arts have historically evolved and what they have meant to their users. Rather, it is aimed at discerning the basic alternative patterns of organizing artistic activities and the consequences, for society and for the arts, of adopting one or another of them.

Most of the necessary knowledge for recognizing these patterns is still lacking or is ambiguous in its implications. Indeed, there is no generally accepted theoretical basis for encompassing all the arts in relation to all sociological variables in all types of societies, from the simplest to the most complex. There is, furthermore, hardly any other field in the whole area between the humanities and the social sciences as inviting to partisan sensibilities as the relationship between art and society. Any general statements about relationships between art and society must therefore be treated cautiously, not as established knowledge but as tentative hypotheses.

THE FIELD OF ART

There are two possible ways to conceive of art sociologically. In the nonhistorical view, art must be conceived as that which is defined by a society, or an artistically relevant part thereof, as art. This could be called the "labelling" view. The work itself may or may not claim to be art; it is recognized as such by those with an authority to do so (in modern societies, by artists and art critics) or by anyone interested. This definition implies that art did not exist in preliterate societies until it was recognized by the moderns, since what now appears to be art has been treated mostly as religious or utilitarian artifacts in such societies. In the absence of the label art, imposed by later cultures, these objects are indeed only utilitarian or religious artifacts. Art arises merely in our perception of them, but does not exist in the intrinsic qualities of the objects themselves.

Historical conception of art

The second, historical view of art is based on the assumption that art is what survives a series of "tests" given to objects that function as art. When such an object is initially presented to the view of people other than its creator, it could be viewed as representing a "claim to art." When it is accepted by large numbers of people in a society or by its established elites or by other artists and art critics, it could be said to have become a popularly, or authoritatively, or professionally validated work of art. But the ultimate test of its artistic quality is whether it can transcend the boundaries of time and space and be accepted by other peoples in other areas.

The historical conception of art implies that art represents a universe with *some* shared characteristics that are everywhere recognizable as artistic (whether or not the concept of "art" is consciously acknowledged). Experimental psychological studies provide some support for this view. It has been found, for example, that traditional Japanese potters agree to a high degree with U.S. art students on the relative merits of a series of works of art shown to them. The agreement is closer than that between American art and nonart students. Thus, practitioners of art appear to agree across cultural boundaries on the quality of artistic attainments.

The agreement on what constitutes a work of art has the following characteristics:

1. It is partial. Each society and period has its own particular standards, as well as generally accepted criteria, by which it judges works of art. The patrons and publics of art (and even art critics) probably insist on these special standards to a greater extent than do the artists themselves. Hence, local criteria should have greater weight in judging works of art when the artistic enterprise is dominated by nonartists. This is one reason for the great fluctuations over time in the economic evaluations of particular works of art, for persons who are not artists usually determine these judgments.
2. The artistic consensus is hierarchical. When artists are left alone, they can, in the long run, roughly agree on the ranking of individual works of art. This happens even in quite egalitarian societies, such as the Australian Aborigines, where most men participate in artistic activities but differences in the quality of achievements and in individual capacities are recognized. The notion of a hierarchy of artistic qualities is therefore not a superimposition of a social hierarchy on artistic experiences.
3. The consensus on art has expandable boundaries. The artists of a society can incorporate works they have been unfamiliar with into their notion of the "field" of art, with a discriminating sensitivity as to their merits. The entrance of new claims to art generates efforts at evaluation, clashes of particularistic values, and the rise of artistic schools and movements. Such claims are ultimately tested in terms of what is considered to be the total structure of the field of art. In turn this structure evolves, by a process of self-testing, through the evaluations it gives to works claiming the right to enter it.

THE AESTHETIC FUNCTION

The nature of the system of art seems to derive from sustained experience with the practical problems of making objects or acts that perform the functions of art and from a sense, which craftsmanship seems to generate, for what transcends mere craftsmanship. A mid-20th-century survey in the United States has shown that craftsmanship by itself is more highly regarded, in judging works of art, by nonexperts than by artists and art scholars.

Art and craftsmanship

Since craftsmanship is a purposeful activity, it appears that the art public is more apt to judge art by some presumed purpose, while artists judge it in terms of what transcends any presumed purpose. If it is ultimately the consensus of artists that determines what is included in the field of art, artistic value must lie in something that is not recognized, even by the artists themselves, as "the purpose" of art.

It follows that works of art cannot be understood by the manifest functions they have been specifically intended to perform. Where they function most purely as works of art, they perform latent functions—unintended and unrecognized. If this observation is valid, art could be regarded as the generalized system of the society for the performance of unintended and unrecognized (but nevertheless needed) psychological and cultural functions. If these functions are effectively performed by other systems of the society, art can remain implicit in them. To the extent that other systems become explicit about the functions they perform and rationalize them to exclude everything that does not

clearly contribute to their purpose, art has to emerge as an autonomous system.

The autonomy of art from the social forces that would control it derives, as an immediate consequence, from the structural requirements of good design peculiar only to art; and, ultimately, from the origin of the aesthetic function in the intuitive organization of unintended and unrecognized functions of an interrelated psychological and cultural nature. To the extent that art performs an aesthetic function, it is not subject to intentional social control. A variety of consequences follow from this conception of the aesthetic function:

1. If its very essence arises from performing unrecognized functions, art must be a less self-conscious, a less "rationalized," and indeed a less professionalized activity than any other in the cultural sphere.
2. Since it must be ready to perform unrecognized functions as they unpredictably arise, the system of the arts cannot be a specialized one, adapted to a particular set of circumstances. It must remain generalized, to some extent maladapted to the existing state of society, and able to function in a wide range of areas of ambiguity.
3. The survival over time, and perhaps the aesthetic quality, of works of art depends on how wide a range of unintended and unrecognized functions they can effectively perform. It is because they have a wide aesthetic range, in this sense, that great works of art function for us even when it is not known exactly what they have meant for their producers, as is the case with prehistoric art. The latent functioning of a work of art is not dependent on the grasp of its intended meanings.
4. It could be argued that an effective organization of any unrecognized function constitutes the aesthetic aspect of the social or psychological system in which it is embedded. Thus, scholars could analyze successful works of art as diagrams of effectively constructed but hidden psychological processes occurring in personalities individually or collectively, and could discern how these diagrams are used, by artists and art consumers, to deal with otherwise raw and chaotic psychological processes. The diagrams may be seen as providing rehearsals for effective organization of these psychological processes or as suggesting alternative models for them or as providing focal points around which they can crystallize.

If aesthetic value depends on consciously unrecognized functions, does an explication of these functions erase the possibility of an aesthetic experience (or prevent it from arising for those who will, in the future, be conscious of the functions the work of art presumably performs)? Not necessarily—if the work of art, after one or more of its functions have been explicated, can still function effectively in other unrecognized and unintended ways. The interpretation of art could be viewed, then, as a struggle against its inexhaustibility, but the functions that have been fully explicated would seem to become more cognitive than aesthetic.

SOCIAL USES OF ART

Relation of art to other disciplines

While, to be artistically effective, art must function as art—filling many unrecognized and unintended functions—it does not generally operate as "pure" art. First, it may overlap other cultural systems—religion, philosophy, science, a secular ideology—and perform, in part, the more clearly identifiable functions of these systems. It is then shaped, to some degree, by the superimposed functions it performs as a part of those systems.

In general, art tends to become increasingly differentiated from other cultural systems in the course of social evolution. Yet in some periods a closer integration of art and some other cultural system may be sought. Thus, the system of science has, in most cases indirectly, affected much of modern visual art. It cannot be taken for granted that the most complete differentiation of art from other cultural systems is most conducive to its authenticity. Nor does a self-conscious "integration" of art with another cultural system enhance art, as is shown by the failure of Socialist Realism, tightly bound to Communist ideology. It could be inferred from the notion of the aesthetic function that an indirect, unintended mutual interaction between art and some other cultural system would be most stimulating to the arts.

Second, art may be used by various social agencies or groups to perform functions these agencies are interested in. Such use of art by social agencies structures the content and style of art and influences its level of creative attainment and its total repertoire of functions. Again, it is not to be taken for granted that it is necessarily disastrous for art to be used for extra-artistic purposes. On the one hand, by using art to fulfill their purposes, social agencies may restrict art's capacity to function as art. This seems likely to happen to the extent to which social agencies successfully limit art to performing any set of consciously recognized and intended purposes.

On the other hand, by using art for their own purposes, social agencies may also stretch art's limits in directions that artists might not otherwise have been inclined to explore, enhance certain of its expressive potentialities at the cost of others, and increase its capacity to communicate with contemporaries while perhaps reducing its ability to communicate transhistorically and cross-culturally. By being forced to struggle against purposes imposed from the outside, artists may become more aware of what is both peculiar and essential to art.

Like other symbolic systems, art can function as a means of attaining the purposes of any system of society. Thus in the economy, art can be used as a means for attaining or symbolizing the possession of wealth but also as a critique of particular ways of using it. In the political system, it can encourage or condemn a particular distribution or use of power. In the community, it can be used as a means of reinforcing or protesting against the existing order of sensibilities, expectations, social rankings, and social distances. In an ideological system, it can operate to strengthen the hold of established values by filling the imagination with forms or content suggestive of these values—or to question them by presenting forms and content that are irreconcilable with existing values. But even if art "objectively" functions to promote particular social ends, it is not necessarily consciously employed to promote these ends.

Extra-artistic functions

None of these extra-artistic uses of art seems by necessity aesthetically superior to any other. What seems important is whether artists accept the legitimacy of the extra-artistic expectations directed to their work and incorporate them, unselfconsciously, into their own notion of the artistic task. If they do so, they should be able to produce good art regardless of the type of extra-artistic expectations imposed on it.

The critical question, nevertheless, in evaluating the aesthetic quality of works of art is whether they can resist or transcend the social uses to which they have been put by their makers and whether they retain their own character and transmit their own message even when manipulated to serve the purposes of their sponsors. An art that has not been subjected to manipulation during its making could well be deprived of a powerful stimulus for acquiring a "resistable" character, an artistic toughness that ensures its aesthetic survival.

THE COGNITIVE CHARACTER OF ART

To what extent do the particular social uses of art shape its content and style? To what extent do content and style, therefore, correctly reflect or distort the "objective" realities of the society in which they have been produced? Does use of art by a social agency or group necessarily imply a distortion, in its interests, of the reality that art may be presumed to reflect? Or can some groups (as, in the Marxist views, the "progressive" ones, which identify themselves with the direction of history) use art in their own interests without thereby forcing it to distort reality?

The most general response to these queries is probably that art reflects either subjective affirmations or subjective denials, symbolic invalidations, of the existing reality. It can therefore be read only as a record of the history of subjective attitudes toward objectively existing reality. The ways in which art has been used can be assumed to influence the subjective responses it will express. The subjective responses that exist in the environment in which art is produced but that are not "useful" to the groups

Inherent subjectivity of art

or individuals that provide resources for art creation, or for the artists themselves, are less likely to be reflected in art. Yet, insofar as art necessarily performs unintended and unrecognized functions as well, it may reflect even the subjective responses that it is not useful to anyone involved in the artistic process to reflect. These responses may contradict the conscious intentions of the artist. Art is never an objective record, and it is never fully controlled by those who use it—or it ceases to be art.

The most significant art may well express both the most striking characteristics of objective social reality and the sense of what is most missed in it—reflections of reality as well as utopian denials of it. The significance of such art may arise from its discovery of ways to articulate these mutually contradictory subjective responses to social reality, without suppressing one in favour of the other. Cognitive distortions in art arise not from an introduction of subjective attitudes but from a denial of the ineradicable contradiction between existing "objective" realities and their possible "subjective" negations. Art distorts the totality of human experience, in any social setting, when it is biassed in favour of either "objective" recording or "subjective" expression.

Art reflects, or compensates for the deficiencies of, objective reality not only in its content but also in style. Even completely nonrepresentational arts therefore have a cognitive character, and subjective orientations to social reality can be inferred from them.

Easily swayed modes of art

Some arts lend themselves more easily to deliberate manipulation in the interest of consciously distorting the ways in which they reflect reality—whether in style or in content. Fiction, the theatre, painting, and sculpture are more vulnerable to such manipulation than music or lyrical poetry. Representational styles lend themselves more easily to manipulation than highly symbolic or completely nonrepresentational ones. Indeed, one reason for moving toward the latter styles, in modern societies, is the desire of artists to escape manipulation. The mass arts are particularly susceptible to imposed distortions. It is deliberate manipulation, rather than merely the use of art by social agencies and groups, that would appear to give rise to distortions in the way works of art reflect reality in its interrelated social and emotional aspects.

SOCIAL DYNAMICS OF ARTISTIC CREATIVITY

If art performs unintended and unrecognized functions, the creation of artistic values should be affected by the degree to which art is needed, in particular social settings or by particular individuals, to perform these functions. That is, for artistic creativity to become possible, there must be many psychic needs that are neither met by existing social arrangements nor can be consciously identified and purposefully dealt with by means of social policy.

Conditions favourable to artistic creativity arise when rapid changes in either the organization of society or in the emotions of its members produce a sharply sensed disjunction between personal emotion and objective social structure. But while the need for art increases during such periods, the possibility of creating it also depends on the availability of resources for creating art. The supply of such resources tends to be diminished in the phase of most intense action of periods of radical change. This phase occurs during rapid economic accumulation, technological transformation, the high points of religious (or secular ideological) reformations, and struggles for imperial consolidation, national liberation, or change of political system. Artistic creativity is enhanced when an increased social need for art coincides with an ample supply of resources for artistic production—thus before and after, but not during, the most intense phase of any cycle of technological, political, ideological, or communal change in the society. Activistic social movements are unlikely to be artistically creative, but they increase the need for art—after they have succeeded or failed to transform society.

Creativity and social resources

Conscious social policy can affect artistic creativity by supplying, or failing to supply, social resources for artistic production commensurate with the existing social need for art. Not all such resources, however, can be controlled at will—the supply of cultural symbolism, for example, depends on its existence in any social system in a form appealing to the imagination of artists. It does not depend on the immediate policies of governments, churches, or parties, although these institutions can, to a high degree, determine the allocation of economic resources to artistic activities. But while a deficiency of such resources may prevent the creation of works of art for which the potential exists in a society, economic resources cannot generate artistic values if the social need for art is not sufficiently intense, as during periods of general cultural quiescence and social complacency, when existing conditions are taken for granted by most members of a society (as in 18th-century Italy); if the creative people are too widely scattered in a large population or too isolated from each other by ethnic, ideological, class, or disciplinary barriers to attain a necessary density of interaction; if an appropriate social organization for producing art, particularly important for the large-scale arts and the film, is not established; or if the general symbolic design of a civilization is not sufficiently developed or is already too "completed" to sustain efforts at producing great artistic designs.

In modern societies, perhaps especially in the large nations, two threats to artistic creativity have emerged in the possible bureaucratic overorganization of the artistic enterprise, leaving too little space for the unintended and unrecognized in art, and the early popularization by the mass media of new artistic movements before they have had time to mature their contributions. Once artistic movements become widely popular, they tend to drop what they have been doing, since the modern cult of originality discourages a continued exploration of what others have become familiar with. Hence, what a style is potentially capable of may never be developed.

SOCIAL ROLE OF THE ARTIST

Variables in defining the artist's role

The two main variables in defining the social role of the artist are that role's degree of specialization and the extent to which it is conceived as involving manual labour or some sort of "spiritual expression."

By their nature, all the literary arts involve less visible manual labour than the traditional visual arts, and the sculptor or painter has been generally more easily and closely assimilated to the traditions of skilled manual craftsmanship. The poet has tended to be associated with the realm of religious ceremony and, later on, with record keeping within or outside the various church organizations. Thus, some of the prestige of literacy in traditional civilizations has attached to the poet's role. Yet it was originally literacy in the service of a god as well as an illiterate aristocracy, and there is still some tendency to expect writers, more than other artists, both to entertain the classes that have become their sponsors in place of the aristocracy and to serve a moral purpose. It is they, of all the artists, who are expected to be "the moral conscience" of society. But besides this clerical tradition, there is also the more anarchic one of vagabond poets, who live "beyond good and evil."

In music, he who composes a work is frequently—at least since the German composer Beethoven—perceived as partaking of a more "spiritual" role, and he who performs it, particularly in groups of performers, has been frequently seen as more of a "craftsman." In the theatre, the actor is a physical labourer insofar as he uses his own body, but, in performing his role, his body assumes and acts out "spiritual substances" originally alien, and possibly greatly superior, to it. This is one reason for the ambiguities in the social treatment of the actor, who, like the singer and the dancer, has been the most exalted and the most despised of artists.

Within the limits shaped by the nature of the particular artistic medium, the social role of artists has been affected by developments in the organization of society and its value system. The main evolutionary trend has been toward increasing professionalization of artistic roles. In the preliterate societies of a more egalitarian character, all adult members of a society (or all members of one sex and a few of the other) may be engaged in activities of producing objects or performances that, in addition to their consciously intended functions, have some kind of aesthetic aspect

(superfluous from a purely utilitarian point of view). But some individuals are recognized by their peers as more competent carvers, potters, or dancers, and their products or performances, though not superior in a utilitarian way, are more highly esteemed; their authors receive a superior compensation in prestige and, less frequently, in material valuables. Yet even the best of tribal artists are spare-time specialists who devote most of their time to meeting obligations incumbent on all members of their society of their own sex and age.

Full-time specialization

It is only with the development of a hierarchic type of social organization that, for male artists, full-time specialization becomes possible, if their society is both interested in art and sufficiently prosperous to sustain a demand for full-time specialists. Women tend to remain semispecialists or "folk" artists, and, the more professionalized the artistic enterprise, the lesser their part in it. In the primitive states or hierarchically organized chiefdoms is formed the main pattern that has governed artistic enterprise in classical civilizations: the division between "high" and "low" art, the specialization of male artists, formalization of their training, patronage by clients, and general subordination of art to the politically ruling class and the religious establishments.

This pattern has occasionally been modified in mature preindustrial civilizations. The basic hierarchic pattern, however, has been radically challenged only in consequence of the Industrial Revolution and the rise of a civilization associated with it, in which aspirations to universal participation appear. This transition is still going on. But "high" art has already become detached from its almost exclusive dependence on the ruling elite. It is now visible, through the mass media, to almost everyone and is intensely experienced by a recently developed loose coalition of artists and intellectuals. This coalition is the primary audience, even for art that has been explicitly produced for other social classes. An example is the Mexican murals of the 1920s that were intended to revolutionize the proletariat but are appreciated mainly by a bourgeoisified intelligentsia. With respect to art, the socially dominant class tends to lose confidence in its own taste and to become "culturally" subordinate to this coalition. From a symbol of privilege, art changes into a symbol of emotional aliveness; hence, the boundary between "high" and "low" art becomes blurred. But the artistic enterprise itself remains highly stratified in accordance with the prestige granted the type of activity the artist engages in and his personal achievement in it. Women once again return to greater participation in the fully specialized production of the "creative" arts. As for the performing arts, women have either always possessed or regained at an earlier time fully specialized roles in them. This is especially true of the dance, except in those cases where religious asceticism or some comparable influence has eliminated them from the performing arts as well.

"Cult of the genius"

Historically, the most significant single illustration of the ways in which the value system of a society affects artistic roles is the Western "cult of the genius." It developed under the influence of the great achievements and the striking personalities of artists like Michelangelo and Leonardo da Vinci in the Renaissance—a civilization that generally placed a very high value on achievement and expression of individual personality. Comparably great achievements in less "individualistic" cultures have not produced any similarly powerful notions of genius. The cult of the genius had the long-term effects of assimilating all the creative, and at a later date even the performing, arts to the "spiritual expression" of poetry; legitimating the demand for creative autonomy on the part of the men whose genius has been perceived as transcending the dimensions of established custom; and, on the basis of the above two changes, transforming those arts previously treated as crafts (and organized in guilds, etc.) into a peculiar kind of free profession.

All professions are corporate bodies of practitioners in a skilled activity who are devoted to an ethical purpose superior to the mere pursuit of wealth, power, or pleasure for themselves. The professions set their own standards, determine admissions to them, and judge their own members accused of transgressions against the profession's ethical code. Modern artists constitute a peculiar profession in that they do all this without necessarily being organized into an association that encompasses all of its members (as the medieval guilds did) and without necessarily being held accountable to public authorities for its activities. The ethos of the artistic profession, as it has emerged in the West since the Renaissance, is opposed to both general organization and any kind of public accountability of artists. Even with the evaporation of the cult of genius, in the 20th century, the conception of the artist's role as that of a free professional has remained as the ideal expectation. There are evident tendencies, from the latter part of the 19th century on, toward a partial collectivization of the role of the professional artist—whether in artistic movements in which several artists interact in producing individual innovations that have something in common (as did the French Impressionist painters of the 1870s) or in artistic workshops in which several artists cooperate in producing the same, or the same type of, works of art (as in the Bauhaus school of design in Germany, 1919–33). Yet it is a collectivization of a free profession, not a return to the guild structures, that *some* artists of industrial societies are willing voluntarily to accept.

Even as a liberal profession, the artistic enterprise remains professionalized to a lesser extent than other recognized professions. The artists are *morally* a profession, but the degree to which they *economically* depend on it as their main source of income varies tremendously. Most professional artists in modern societies do not derive the bulk of their income from the direct exercise of their profession, particularly where they have to sell their products under conditions of a free market.

Professional incomes

Estimates made at the beginning of the 1970s suggest that no more than one-half of 1 percent of the professional painters and sculptors in Paris and in West Germany receive a regular and sufficient income from the sale of their works. Some arts have retained, even in the self-perceptions of the artists, the character of vocations rather than professions: it is still awkward to describe someone as a professional poet.

It is conceivable that, in the more affluent and leisurely societies predicted for the future on the basis of technological progress, well-trained amateurs might again, as in the past in China, begin producing art of respectable quality. The nature of the aesthetic function sets limits on the extent to which a professionalization of the artistic enterprise can be sustained without aesthetic loss.

ARTISTIC CULTURES

Artistic cultures are the various basic situations in which art is produced. Each such culture involves a distinct type of social organization of artistic activities that is associated with a distinguishable attitude of artists toward their work. Artistic cultures arise from artists' relationships with other artists, their publics, their means of earning a living, and agencies affecting artistic activity; from their involvements with cultural systems not specifically artistic; and from artists' technologies, shared emotions, and ideologies. Changes in any of these variables modify artistic cultures and give rise to new variants of them.

Once the artistically undifferentiated unity of the tribal society has come to an end, several variants of "traditional" and "modern" types of artistic cultures can be distinguished.

Traditional artistic cultures. At least seven highly distinct types of artistic cultures can be identified in traditional (preindustrial) societies.

The folk culture. Artists are nonspecialized members of their community of residence and closely involved with all of its activities. Their art deals with typical experiences, concerns, or needs of ordinary members of that community. The real-life references are, however, linked with elements of highly utopian imagination (folktales, in which an attitude critical of the existing reality is frequently expressed), abstract stylization (geometric ornamentation of utensils), or symbolizing lyricism (folk songs). The artist is one of the people, but living more powerfully in the imagination than the others. Artists create by following or

elaborating on traditional patterns. Much art is produced spontaneously for oneself or in friendship or for communal enjoyment, rather than for pay.

The artisan culture. The artist is a specialist—possibly a member of a specialized collectivity such as a guild or a workshop set up by a state or a church—who works on order and for pay (or on command) only. He develops the pride of good craftsmanship and habits of regularity and reliability in his work. He does whatever style or content is required by his client without asking about his intentions. Whatever individual needs he expresses in his work, his conscious ideology is that he is a skilled worker who respects high standards of craftsmanship and produces to earn a living. He belongs, not permanently to his community, but temporarily to his client, wherever the latter may be located, and perhaps to his clan or guild of fellow artisans who share his craft ethic. To the extent that his client is his local community, however, he becomes a "civic" artist, whose highest purpose is, on command, to celebrate his community, as in the cities of the Mediterranean civilizations. (In modern society, this culture survives most nearly, perhaps, among orchestral musicians.)

Specialized craftsmen

The clerical culture. The artist is very much a craftsman, but his craftsmanship is subordinated to a highly valued tradition of a literate civilization, which he is committed to serve and to defend by his work. By his association with this tradition, he gains in prestige, but he also acquires a moral responsibility which the pure craftsman is not bound by. He belongs to a moral community, not necessarily localized, that provides him with criteria for judging which works of art are worth making. In this can be discerned the potential beginnings of an artistic consciousness and of a critical attitude of the artist toward society and its artistic ideologies. Most of the time, the clerical artist (for example, the medieval manuscript illuminator) is submissive to the discipline of the moral community of which he is a part, but, in the conception of the artist as a civilized servant of a higher moral purpose, there is, at least, a potentiality of criticism of the organization to which he belongs.

The ecstatic culture. This is a general term suggested for situations in which some type of artistic creation—though it is usually more than merely artistic—occurs in the midst of a mediumistic trance performed for religious or magical purposes, during an orgiastic happening, as a consequence of a mystical experience, or even as an element of prophecy. Art, performing or literary, arises from emotionally intense experiences, "divine madnesses," which are perceived to have high symbolic significance but over which the individual (or, sometimes, the group) that has them imposes no critical controls. Traditions that generate ecstatic art creation, such as the Dionysian ritual (performed by followers of the cult of the god Dionysus) in ancient Greece or the poetry of Hebrew prophecy, usually arise in loosely structured societies (as do the mediumistic priest-doctors known as shamans) or in times of social crisis (as mystics and prophets are likely to do). The shaman tends to be conservative with respect to his society, the mystic apathetic, the prophet radically innovative. Numerically, these types constitute a small part of the artistic cultures of the well-ordered literate civilizations and are least important in the visual arts dependent on organized patronage. But they have been sources of revitalizing impulses for established traditions. In the ecstatic culture, the artist is a nonspecialist who opens up higher, or more truthful, realms of existence and transcends the norms of everyday life. Theories of "artistic inspiration" and of the artist as martyr (or scapegoat) arise from this culture and continue to be generated in it.

Sources of revitalization

The courtly culture. The artist, who may be an aristocratic poet or a craftsman of ordinary social standing, is directly dependent upon a secular royal court or a household of the high nobility (or, by extension, plutocracy). To some extent, he is even socially involved with this household as a vassal, royal favourite, or court artist. He therefore not only produces for but also spiritually identifies with the high aristocracy for which he works, without actually being a member of it. The artist is a man ennobled by his art; his art must therefore be permeated with "noble" attitudes: heroic exaltation, fashionable late-medieval despair, or refined Rococo or Rājput sensuousness. While this art, by glorifying the establishment, reinforces its privileges, it may at the same time educate its—and other people's—emotions in novel ways.

The gentlemanly culture. When men of high social standing and independent economic means become active in the production of art, a tradition of gentlemanly art may arise (as is exemplified in much of the calligraphy and landscape painting in historic China). Art tends to be contemplative, subjective, aestheticizing—but not consciously "ennobling," since artists already possess sufficient "nobility" (as courtly artists do not). The artist is a man of independence who observes and contemplates without desire (an attitude that has been influential in modern aesthetic philosophies, frequently produced by such men).

The vagabond culture. All the traditional types of artistic cultures presuppose a firm social location of the artist in relation to a community, except where the artist is a vagabond adventurer who roams the countryside producing or performing works of art—comical tricks, storytelling, lyrical songs (for example, the medieval minstrels or the 15th-century French poet François Villon). In origin, he may be an uprooted folk artist, who belongs only to temporary communities. He is neither an apprentice journeyman nor a gentleman (who may also travel widely) but an impecunious amateur master unconcerned with material gain. (Masterful amateurism in the arts has two sources: the gentleman and the vagabond.) In contrast to the ecstatic artist, the vagabond does not assume that his works possess any higher symbolic significance. They are plays of fancy that amuse the fancier and whoever cares to join him. Individual vagabond artists have probably always existed, but vagabond cultures can arise only in the fringe areas of cities (for example, a transitional zone between industrial and residential quarters) or where nomadic tribes like the gypsies travel over a settled countryside. Travelling theatrical groups exemplify both conditions. Whatever the effect of vagabond artists on society, they, like the ecstatic artists, tend to have a liberating effect on the arts in which they engage.

Modern artistic cultures. Several types of artistic cultures, in some cases anticipated in earlier periods, have come to a full development in industrial societies. Some older types have been transformed and new ones evolved. In the advanced industrial societies, artistic cultures tend to become fluid—composed of overlapping and fluctuating elements, components of a mixed orientation rather than rigidly separated alternatives. Nevertheless, several types can be distinguished.

Artistic cultures in industrial societies

The genius culture. The product of a conjunction of the artisan, clerical, and courtly cultures, the cult of genius emerged in the early 16th century during the High Renaissance and became more fully developed in the age of Romanticism at the beginning of the 19th century. The artist conceives himself as the "unacknowledged legislator of the world" (in the words of the English poet Percy Bysshe Shelley), an autonomous, godlike creator of new orders of reality obedient only to his perceptions and the categories of his mind. He is superior to the specializations of the sciences and the crafts: he creates the unifying symbols of a developing civilization. He is self-confident, a proclaimer of new values rather than a critic of the established ones. Historically, perhaps the greatest significance of the genius has been in the legitimation he has provided for the developing conception of the professional artist.

The professional culture. The most essential characteristic of the artistic professional is that he himself chooses what to produce or perform. But, unlike the gentleman artist, the professional produces in order to sell. His art is the activity of a specialist competent in the techniques of expressing subjective perceptions. In the visual and the literary arts, the professional usually works in his own studio, at times of his own choice. In the performing arts—which are less "professionalized"—he is dependent on facilities provided for him by entrepreneurs and impresarios. He may have an agent who represents his economic interests in any case. But the essential point, in the artistic profession, is not the ownership of the means of production; it

is the artist's sense that he can, and indeed must, depend upon his personal aesthetic sense. This is the predominant image of the writer in modern Western societies.

A sociologically important distinction can be made between private and public professionals. The first have originated in the gentlemanly culture, the second in the salon-coffeehouse-journal circuit of the 18th and 19th centuries, in which the enlightened aristocracy and the increasingly popular press joined forces in supporting a socially critical collectivity of middle-class artists, journalists, and hangers-on. In this public culture, it was most important to be up to date and to respond to current issues by one's wit. As this culture began to disintegrate, the distinction between private and public professionals was partly replaced by the distinction that exists between bohemian and radical avant-gardists.

The applied-arts culture. The old artisan tradition has continued in the modern applied arts (for example, industrial design, advertising). The artist is, however, not usually in a guild but, rather, is employed in a white-collar type of job by large business firms or governmental organizations, sometimes by "artistic firms," which, in their management, are not very different from other business firms. But while the applied artist is employed as a craftsman, his attitudes have been affected by the culture of professional art. He may therefore be a craftsman with aspirations to professionalism and may tend to perceive himself as failing to realize his artistic aspirations, regardless of financial success and social recognition. Applied artists with the self-conception of professionals provide one of the main supports for movements of radical criticism of contemporary society and culture.

White-collar artists

A development toward the applied arts may be occurring at present in such an established profession as architecture: the increasing costs over which he has little control, the various restraints over his activity (zoning codes, labour union regulations, and the like), and group pressures to which he must respond are tending to transform the architect into a high-level organizational craftsman. As their opportunities for being artistic professionals decline, architects tend to conceive of themselves increasingly in the image of social engineers—a trend noticeable also in the other arts.

The mass-arts culture. The notion of the mass arts implies that the artist communicates with his public through the mediation of some type of mechanical or electronic machinery; that he does not kno even the kinds of people of whom his public will eventually consist and may be unfamiliar with the characteristic experiences ofmost of his actual public. Furthermore, the responses that he will get from his public will probably also be, in most cases, mediated either in a mechanical manner (statistics of sales) or by professional intermediaries (newspaper critics). Film, television, mass-periodical fiction, and recorded music represent the clearest cases of mass art. Insofar as the mass media deal with any art, they tend to assimilate it to a greater or lesser extent to the mass arts. Because of the conditions under which he operates, the mass artist lacks a sense of general standards—other than the purely technical ones of craftsmanship—by which to orient his artistic activity. Nor, since he is dependent on large, impersonal audiences, can he trust his own convictions to guide him. His typical solution is to cultivate an "image" that has proved to be popular with a mass audience and that, once established, is "forced" on him by his public. He becomes a victim of his own image, in the sense that he is trapped within its confines. He is an uncertain specialist in the symbolic manipulation of diffuse audiences.

Cultivation of a popular image

Avant-garde cultures. Another characteristically modern artistic culture is definable by two aspects: a principled sense of alienation from significant aspects of existing reality and a conscious commitment to overcoming the deficiencies of existing reality by artistic means of a completely novel character. One of the distinguishable types of this artistic culture is the bohemian avant-garde, which tends to be alienated from all rational and utilitarian aspects of social organization and cultural tradition and aims to create a new kind of exaggeratedly irrational art and, perhaps even more important, an irrational style of life. (Typical of this group is the French poet and critic Charles Baudelaire.) Its behaviour centres on the ambivalent and the self-consciously paradoxical: a tradition of pursuit of the new, the cultivation of a pleasureless hedonism, the development of systems for the liberation of spontaneity. Another type of this culture is the radical (politically committed) avant-garde, which regards itself as alienated from "oppressive" and "exploitative" political and economic institutions and tries to create a new kind of art intended to undermine faith in these institutions and to provide a basis for their abolition or reconstruction (an example is the German poet and playwright Bertolt Brecht). These two avant-garde cultures overlap, and, over time, one may change into the other. The alienated type is more influential among artists when faith in ideological utopias declines generally in their society. Perhaps because of its tendency to limit itself to consciously recognized and intended purposes, the aesthetic achievements of the radical avant-garde (the "prophets") have been less significant, so far, than those of the alienated avant-garde (the "mystics").

More recently, a third type has emerged: the anti-artistic avant-garde, which is alienated from the very notion of art and its practice as heretofore conceived, even (or perhaps especially) in the other versions of the avant-garde culture. Art itself is perceived as "oppressive" and "exploitative," and the obligation of the artist (and of the art critic) is seen as the promotion of the "end of art"—its diffusion into a "life" of childlike impulsivity. The very coherence of art is felt to be an imposition of an arbitrary system on the immediacy of "aesthetic experiences," which are to be pursued with a self-conscious repudiation of any deliberate control.

All the avant-garde cultures can be seen as diverse outgrowths of the culture of genius, with which the peculiarly modern developments in the arts began. But the bohemian and radical version of the avant-garde culture are also reactions, by artists and aesthetically sensitized intellectuals, to the development of the modern industrial society—a society highly rationalized and, in spite of its ideology celebrating the "common people," continuingly stratified. The anti-artistic avant-garde appears to be a "shamanistic" reaction against the professionalization of art, and, like the mediumistic trances of traditional cultures, its efforts may prove to be a socially innocuous (or conserving) ritual that is perhaps psychologically refreshing.

Total-command cultures. Hitler's Germany and Stalin's Soviet Union still provide the best examples of what happens when a modern artistic enterprise is subordinated to a political ideology that has the monopoly of organized power in a society. Artists are forced back into clerical roles—with the difference that the moral responsibility for upholding an ideology is imposed on them without its necessarily corresponding with their own convictions; and, even when it does correspond with their ideological views, it conflicts with the conception of the professional role that they, as modern artists, regard as their primary orientation within the sphere of art. An artistic enterprise can be effectively politicized, without losing most of its aesthetic qualities, only if artists are content with assuming the role of "clerics" illustrating, with appropriate technical means, the manuscripts of a political ideology, yet do so without limiting their imagination to service in its mission.

Politicalization of art

The scientific-technological culture. Partly under the influence of imaginative developments in science and technology and partly out of disillusionment with the avant-garde cultures of art, an artistic culture in which the artist is assimilated to the role of a scientific researcher has arisen. The artist is a strategist of concepts, a deviser of technical systems for others to build, and a manager who invents ways for workers to relate technological or natural processes to each other. His difference from the scientist is that he seeks not new understandings but new perceptions; his difference from the technological experimenter is that the systems constructed are expected to have no utility. They are frequently evanescent or self-destructive. This type of artistic culture requires large financial outlays and tends to result mainly in entertaining decorative effects. It is neutral with respect to the particular institutions of a

society but tends to glorify the technological and economic achievements that enable it to sustain such works of art. Thus it is the equivalent, in a technological civilization, to the monumental art by which absolute monarchs had been glorified in the agricultural empires, such as ancient Egypt.

Movement cultures. Artistic situations that arise in spontaneous movements of a general cultural, rather than specifically artistic or primarily political, character (such as youth movements) may generate an immediacy and intimacy of the relationship between producers and consumers of art that is rare anywhere else in modern societies. The effect on the arts is to simplify aesthetic structures and integrate them more closely with the experiences of the many. The arts most frequently involved are poetry, music, and some form of theatre. The movements themselves are not highly productive of enduring artistic achievements. Their main significance is in opening up large audiences to a sensitive responsiveness to art or to particular kinds of art—a sensitivity anesthetized by the mass arts.

Amateur art making. With the increase in leisure time, amateur art making has become increasingly popular in advanced industrial societies, as had amateur musical performances in the European upper and middle classes in the 16th and 17th centuries. This phenomenon contributes to the blurring of the traditional distinctions between producers and consumers of art and to the involvement of larger numbers of people in a more active manner with the artistic enterprise, strengthening the social anchoring of art in the modern society.

ECONOMIC EVALUATION OF THE ARTS

Factors in the allocation of wealth to the arts

The economic evaluation of the arts in general is indicated either by the absolute amount of economic resources put into their production, acquisition, distribution, and consumption or by the relative share of the total income of an individual, group, or society that these allocations represent.

While the absolute size of expenditures may be determined more by the level of wealth than by the degree of interest in art, it yet has a significant effect on the artistic enterprise, perhaps especially on art collection and prices of artworks. When the absolute amount of funds available for the arts is very large, it produces the phenomenon of "cultural imperialism": it enables the powerful, whatever the degree of their interest in art or their own creative attainments, to dominate and overwhelm the cultural activities of the financially less well endowed, unless the latter remain isolated or protect themselves by consciously designed cultural policies.

The percentage of the income devoted to art, on the other hand, bears a closer relationship to interest in art, and it could therefore be expected to be more closely associated with the level of artistic creativity—which does not necessarily require large economic resources to sustain it. Some of the finest works of primitive art, for example, have been produced in places where life is economically precarious, such as the swampy areas of New Guinea. The percentage of wealth devoted to the arts tends to be greatest in economically comfortable societies that have passed their peak of economic expansion—Italy after the middle of the 14th century and Spain in the late 16th and 17th centuries, to cite two instances.

The economic valuation of art is also affected by the degree of its association with activities that are regarded as possessing great importance in a particular society. Actual importance of an activity can be measured, in economic terms, by the size of economic allocations for its pursuit. The closer art is integrated with the activities (typically, the economic, political, religious, and military) getting the major share of available funds, the larger tends to be its own share. In the advanced industrial societies, a major financial basis of contemporary arts is their integration with the growing leisure industry (as in the Broadway theatre or the arts in the mass media) and with education (as in the widespread employment of artists in schools and universities). It is partly for this reason that modern states tend to allocate larger sums for the support of the performing than of the object-making arts and for the dissemination than for the production of art objects.

To the extent that art itself is a prestigious activity, individuals, groups, cities, or whole societies may compete for supporting its most important living practitioners or acquiring the most famous works. This motive has probably been present, to some degree, in all elite-oriented societies, but its significance tends to increase in civilizations that have many centres rather than a single "imperial" one—in Renaissance Italy, in Germany before its unification, and in the modern world in general. Competition of would-be patrons for the prestige of living artists increases not only the economic allocations to art but also the freedom of "elite" artists. Modern state support for the performing arts, in particular, rests to a high degree on competition for internationally recognized cultural reputations.

The value of art as an investment of wealth, either for ensuring its preservation or for increasing it, grows in periods of social and political instability and currency fluctuation. Small-scale artworks made either of precious materials or by prestigious old masters become the safest investment in times of trouble. This has been an important motive for the collection of art by the rich in the Renaissance as well as in the 20th century. It is characteristic of this system that a high economic valuation of art does not necessarily correlate with the income received by living artists, since the object of speculative interest is the reputation of a master, who is frequently dead, rather than the aesthetic merits of the work itself. It is a system in which artists, instead of being supported by the rich in the pursuit of aesthetic values, support the rich in their pursuit of further enrichment, since an artist may endure a lifetime of poverty to produce works that then become tokens of steadily increasing worth. As a consequence, the system encourages capital accumulation rather than artistic creativity. It has little effect on the performing arts or literature but is currently a major source of discontent among practitioners of the visual arts.

Relation of achievement to value

Increases in the social status of artists and in the prices their works fetch on the market have, in the past, tended to follow increases in the creative attainments of an artistic tradition. Thus, in England, an increase in artistic creativity in the early part of the 18th century (as exemplified by the paintings and engravings of William Hogarth) was not followed by a significant increase in the social and economic status of some artists until the 1780s. During the late 19th century, however, when contemporary English art was admired above that of any other period or society and highly rewarded, it produced only minor achievements. Clearly, economic allocations do not guarantee artistic efflorescences, and inferior art may be ascribed high economic value. Great artistic achievements, nevertheless, appear to increase the economic value of art, though perhaps not immediately. While the economic value of particular objects of art depends in part on their historical significance, uniqueness, and fashion, the economic valuation of art itself is, in the long run, not altogether unrelated to the presence in a society of large numbers of aesthetically meritorious works.

SYSTEMS OF FINANCING ARTISTIC ACTIVITIES

The economic support of artistic activities can be provided by the artist himself, who derives his income from sources other than the rewards he receives for producing art, or it can be furnished by nonartists supplying the economic means for the artist to survive while he is making works of art.

Self-financing of artistic activities has always been, and still is, considerable. The gentlemanly type of artistic culture is wholly supported by artists of independent means. The writing of lyrical poetry has very generally been a leisure-time activity of persons deriving the bulk of their income from other sources. In modern times, self-financing is present when the artist survives from an activity that does not require him to produce works of art, or when his income from producing works of art is significantly lower than what he could earn by being employed in a legitimate alternative occupation accessible to him. The dance is heavily self-financed in the second sense.

Sponsorship of artistic activities may be said to occur when the artist draws upon the economic support of individuals or organizations that do not necessarily expect to get anything for themselves in return for it. They support the artist because they are committed to art for its own sake or to a particular artist. In this specific sense, artistic activities may be sponsored by government agencies supporting art as a "social service" for the people, by private foundations, by individual sponsors, by relatives who support the artist while he is economically unsuccessful, or by friends and acquaintances whom an artist "sponges on." Sponsorship may thus be, in some cases, unintentional.

Financing outside of the market system

Self-financing and sponsorship are economically "abnormal" systems of financing the production of goods and services that potentially possess an economic value to people other than their producers. These modes of financing are found principally in the "cultural" activities, above all religion and art (but also in "ideological" politics), and they occur because artists are not treated as if they were engaged only in the production of economic values. They produce economic values, but there is also a "religious" aspect in their activity that requires to be supported for its own sake. If the arts were to be supported entirely by a market economy, this aspect might well disappear: art would then become purely a commodity.

The arrangements of economic support that have traditionally carried the main burden of sustaining artistic activities are the three types of normal economic systems into which the arts may be integrated like any other productive activity: the exchange, command, and market economies.

The exchange economy. An artist in the exchange economy produces for a customer familiar to him who, in turn, directly or through a series of similar exchanges, provides him with a desired good or service of equivalent value. The transaction between the producer and consumer is, ideally, a ritualized exchange of gifts, and, in addition to providing the goods that each needs, it reinforces their social relationship. This type of economy prevails in the egalitarian type of preliterate societies and in the folk type of artistic cultures.

In more advanced societies, this becomes a minor part of the economic support structure of the arts; it tends to be limited to the private circle of artists and their friends. The economic rewards to artists within this system are small, fairly continuous, and not greatly differentiated in magnitude.

The command economy. The artist in the command economy depends on a consumer—an individual or organization—of superior status who has extracted a great deal of wealth from the producers in the economy and can distribute it, perhaps within the boundaries established by tradition, as he sees fit. This economic system is dominant in a range of societies extending from the Kwakiutl Indians of the northwest coast of North America through the classical absolutist monarchies to modern totalitarian states. Usually, however, it does not completely displace elements of other economic systems.

Patron systems

In this system, the artist depends for his support on relatively few patrons, who are far more powerful and socially resourceful than he is. He may be well rewarded by them, but he must produce works that will be regarded as possessing high value—usually both aesthetic and economic—by his patron. He is paid well if he produces economically valuable aesthetic goods and services. (A significant exception to this principle has been built into the situation of women in the performing arts: in male-dominated societies, they tended to be well paid if they performed sexual as well as artistic services.) Before the modern age, all artists who produced expensive works of art were supported by economies of this type. Even today, architecture—to the extent that it is an art—and the large-scale "technological" arts are supported in this manner.

The market economy. In the market economy, many artists compete for the favour of numerous customers not greatly differing in their purchasing power and not personally known to artists. Since artists are not familiar with their customers, their products become acts either of self-expression or of appeal to an image, an abstract stereotype, of the "average customer." In reality, market economies in the arts only approach this condition to varying degrees: some publics are known to their artists, and there are always elements of the command economy limiting the operations of market economies. There are always economic or political elites or powerful business and political organizations, at whose command, to varying degrees, artists must be in order to achieve "success" even in market economies. And in the modern system of art trade, the mass media have assumed the character of a command element in a market system. To be able to compete for numerous potential buyers, artists must first ingratiate themselves into the favour of a relatively few taste makers in positions of power within the system of mass communications. The more centralized this system is, the more of a command character it acquires.

In a pure case of market economy, the artist's freedom is great, but his rewards are highly unpredictable and uneven; his uncertainty and sense of alienation from his public is a natural consequence. In market economies, the arts are particularly dependent on support through the economically "abnormal" systems of self-financing and sponsorship. These systems typically provide the economic basis for radically new departures in the arts in market economies. They also subsidize the established arts when their costs are too high to be supported by the market (symphony orchestras, for example).

Competition for the "best"

When there are several basically independent command economies in adjoining areas, all interested in art, they are likely to compete with each other for possessing the "best" living artists (as well as the works of the most prestigious dead masters). This competition is apt to favour artistic creativity in a way that pure market economies, concerned with the production of the most usable rather than the most perfect, seem to be less capable of. First of all, pure market economies are hard put to concentrate sufficiently large economic resources to finance expensive artistic achievements (except where they can be immediately "consumed" by large numbers of people, such as in filmmaking). Beyond that, the expectation, typical of command economies, that the artist will produce the most perfect object he is capable of producing may actually increase his motivation (and consequently his capacity) for doing so. Market economies do not hold forth any such expectation. No longer economically supported, expectations of the artist's commitment to excellence in a market economy come to depend mainly on the professional integrity of the artist.

It seems probable that a given allocation of funds to artistic activities can be more directly translated into high artistic achievements in command economies than in market economies—provided, however, that the command economy, in addition to insisting that an artist produce the most perfect, also permits him to do what he is best at. It is in the latter respect that the command economies of modern totalitarian states have fallen short: they have prescribed for their artists a manner of working in which the latter could not do their best work. Mass-media "command economies" may also work to that effect.

Market economies are potentially more capable than command economies of integrating art into the private life of ordinary members of society, and not only into the life-style of the elite. But market economies seem less capable, by themselves, of generating high artistic achievements (as well as producing artistic public environments). There is, therefore, likely to be an increasing demand for art in market economies but a reduction in the quality of what is being produced. There may be other, quite powerful stimuli to artistic creativity in societies with market-type economies, including the creative entrepreneurs who succeed in joining the market mechanism to an artistic purpose. But the impersonal workings of the economic system of the market itself appear to interfere with the accumulation of the "creative capital" while encouraging its dissemination.

THE ART MARKET

The art market is a complex system, with roots in preliterate societies and historical civilizations, by means of

which artistic activities are organized as profit-making enterprises. Expectably, this system is most fully developed in capitalist economies, but, since economic enterprises generally have to take profitability into account, even in state-managed economies, the art market has some degree of existence wherever art is treated as an economic good that is offered, for a price, on a market.

The operations of the art market are determined both by the general dynamics of market systems and by the peculiar character of art as a commodity.

Sellers' options

People buy art for various reasons. But the organizations that specialize in selling art (or also in producing it for sale, as theatrical companies) are built in most cases on one of two assumptions about the motivations that propel people into spending their money on art. There is an elite market that sells reputation and a popular market that sells entertainment.

Selling entertainment. The popular market provides art for customers who seek primarily entertainment. It must, therefore, adjust to the prevailing popularly defined conceptions of what is entertainment or shape such conceptions to correspond with the product it is capable of providing. It conducts market research into audience preferences, but it also manipulates these preferences by spreading the impression, through its publicity agents and the mass media of communication, that entertainment is what the popular art market dispenses: not only its finished productions but the whole behaviour of the people involved with them is "entertaining." The artist must "sell" not only his art but also his behaviour outside of the art system. The implicit aim is absorbing everything that appears to the mass audience to be entertainment into the ambience of the popular art market—by transmuting the entertaining aspects of life into salable works of art, by associating entertaining people who are not artists with the popular art market, and so on.

Conversely, whatever cannot be regarded as entertaining tends to be forced out of the works of art or artistic performances handled by this system. It puts a premium on "slickness," the glittering or sensational packaging, and discounts both high seriousness and unique sensibility, unless they can be seen as "entertaining." Its influence spreads even beyond the arts it deals with, affecting the character of art that people highly exposed, from their early years, to this system produce and consume even when they operate outside of it.

Selling reputation. If the popular market sells entertainment (an important means of overcoming sensory deprivation), the elite market sells reputation (an important resource for legitimating high social status or aspiration). It is, therefore, involved only with customers concerned with reputation. Such customers may be individuals or organizations. While the "masses" are mainly involved with the popular market, organizations, including states, tend to be more closely linked with the elite market. This market attunes itself to the prevailing hierarchies of artistic reputation, but it also reinforces these hierarchies by spreading the news of the high prices (or prizes) fetched by the works of art handled by it, by encouraging competition among buyers (or supporters) of unique art objects, and by subsidizing the work of scholars and critics that is likely to draw the attention of purchasing elite groups to their goods, increasing their cultural reputation.

Transforming cultural significance into economic value

The elite market has a monopoly of the "measuring rods" used to transform cultural significance into economic value. For the buyers, these scales of measurement indicate the degree of their "cultivation." If impresarios in the popular market have the role of tastemakers (as well as organizers of a complex set of productive activities in the performing arts), dealers in the elite market operate as cultivation makers, cultivation being economically definable as the rate at which cultural reputations are paid for. Art sellers for the elite do not create the reputations, but they prosper by increasing the price of acquiring cultivation.

On the living artist, the elite market has the effect of encouraging him to produce a reputation that, highly paid for, identifies the purchaser as a person of cultivation. The artist's works must be the opposite of what the popular market deals with—they must be *not* entertaining but, rather, culturally "significant" (in accordance with current conceptions of cultural significance). In the 20th century, cultural significance is regarded as virtually equivalent to being innovative.

Interaction of popular and elite markets. Markets that are oriented to elites but deal in mechanically reproduced multiples, such as the publishing of "serious" (as contrasted to "popular") books, represent a special case in between elite and popular markets. The objects they deal in are not inherently "entertaining," and yet the cultural significance an individual object is seen to possess does not lend itself, when the object is multiplied many times over, to being transmuted into inflated prices on the market for reputations. Publishing costs have become so high that any published book, unless subsidized, must enter a popular or semipopular market. The transformation of book publishing into an industry oriented to a fully popular market has been encouraged by the paperback revolution and the profitability of motion-picture rights for both publishers and writers. A peculiarity of the book market is that it has developed semicaptive publics, in the form of book clubs, which the other arts have not been as successful in organizing.

The motion-picture industry is in the popular market. So are all the "spectacular" arts, such as the opera and the ballet, insofar as they are economically dependent on a market—even though their markets have been, by tradition, smaller and more selective than those of the cinema. The smaller and more select a market of the popular type is, the more important seems to be the influence of professional critics on it. And the more influential the critics, the more closely does a popular market approximate an elite market.

Relation of selling to quality

As the history of the opera and the ballet indicate, it is not to be assumed that a market concerned more with selling entertainment than with selling reputation does not permit the production of high artistic values. Even though a public is presumed to be spending its money for entertainment, entertainment is not necessarily what the artists think they are selling—one reason for the alienation of the artist from his public. The great impresario in the popular market is the one (such as Sergey Diaghilev) who arranges for artists to sell their best art to a public convinced it is getting glorious entertainment. Pioneering in new forms of entertainment may both change notions of entertainment and produce objects that will enter the elite market. It is not primarily by the objects produced but, rather, by the manner of their economic operation that one may distinguish between the popular and the elite market in the arts.

Popular and elite markets are not wholly separate. Prices of paintings by old masters in the elite market tend to be increased by their pleasing appearance and are usually reduced when they depict inelegant scenes of suffering. The elite market operates most purely, in terms of reputations, when art museums are involved. Conversely, any "respectable" popular market uses (and manufactures) reputations to increase the economic value of its product. Only anonymous entertainment—by "nameless" artists—can constitute a purely popular market.

The popular market, by treating art as entertainment, imposes on consumers' contact with art the expectation of fleeting experiences and of the absence of any cultural significance transcending them. The elite market imposes the expectation of enduring values and of the presence of the highest kind of significance. In both cases, the market exercises an influence over the definitions of art held by the art consumer (and even by the artist, for that matter, to the extent he is governed by his market situation rather than other elements of the artistic role). The effectiveness of the influence depends on the degree to which market considerations alone govern artistic experiences and activities.

In both the popular and the elite markets, contrary to the usual dynamics of the marketplace, increases in the number of artists and in the amount of their production, relative to the size of the demand for their work, do not necessarily reduce the prices of works of art. In modern societies, these trends produce a focussing of attention on a few (who receive very high rewards) and the neglect of

the rest of art producers, many of whom may be only slightly inferior (or not inferior at all) to the "stars."

Publicity in the mass media, which both the popular and the elite markets manipulate and benefit from, reinforces this tendency by promoting celebrity cults. The objects of these cults may be intrinsically meritorious or not, but the cult, in either case, destroys proportionality between merit and reward. Mass media tend to promote in the arts a phenomenon similar to the medieval cult of sacred relics. Current profitability of investments in art objects rests, perhaps mainly, on a cult of this sort.

Artists' unions

Trade unions of artists—a recent phenomenon—are important mainly in the popular art market. They establish lower limits for the incomes of working artists but also increase the cost of full-scale artistic productions, particularly in the performing arts. In effect, they seek to eliminate the need for the performing artists to finance their art by accepting an income less than commensurate with what they could get from an alternative occupation. But the increases in costs of artistic performances render the performing arts more dependent on sponsorship systems. Neither the European theatre nor symphony orchestras anywhere could survive in a perfect market system. While increasing the economic security of employed artists, artistic trade unions tend also to promote the unemployment or the underemployment of the performing artists and the growth of a "serious amateur" system (for example, the off-off-Broadway theatre and other experimental fringe theatres), which, in addition to developing a style alternative to that of the professional theatre, may also nurture fresh talent and feed it into the professional system. Artists working for the elite market have been less concerned with establishing trade unions. They either survive as celebrities—or fail to survive.

Marginal phenomena. In addition to the two major types of art markets, one can distinguish two currently marginal phenomena: traditional (that is, noneconomic) elements in the art market, such as ideological commitments to particular types of art or group commitments to particular artists, and the various "little" markets, at present increasingly numerous in Western societies, in which art is sold not exclusively, or not mainly, for profit and which may indeed be partly financed by proprietors or participants from income not derived from the sale of art or through their own unpaid services (little magazines, artists' cooperatives, the U.S. "underground" cinema at its beginnings). Little markets may be individual or cooperative, expert or amateurish, and built on the most varying motives. While their cultural role may be important, particularly in nurturing difficult artistic innovations, the little markets collectively cover only a minor part of the art market. If they become economically successful, they tend to be incorporated into one of the two major systems of art trade.

Incorporation of traditional elements into art markets

The "traditional" elements are usually incorporated into one of the three types of art markets and limit the intrinsic logic of their operation. In the elite markets, nation states impose restrictions on the export of historically significant works of art; all respectable book publishers print a certain number of worthy books they expect to be altogether unprofitable. In the popular markets, traditional criteria establish limits of "proper" and "improper" entertainment (performances of avant-garde musical works, for example, are consistently less profitable, both in Europe and the United States, than those featuring more traditional ones). On the little markets, which tend to lack economic staying power, commitments to an identifiable tradition exert a stabilizing effect.

REMUNERATION OF ARTISTS AND PROTECTION OF THEIR RIGHTS

Methods of remunerating artists depend on whether they are involuntarily bound (serf artists) or, with their own consent, are attached (palace artists) to a consumer or organization that uses their works, or are free to sell them on the market.

Bound artists are provided with subsistence by their lords; attached artists receive regular salaries (which may be high in the case of palace favourites and academicians) and additional bonuses for particular successful works of art. Freedom to sell on a market may be unlimited, as in modern societies, or it may be conditional on membership in a socially recognized group of producers (the medieval guild system). To the extent that it approaches a perfect market (numerous artists of approximately equal artistic reputation competing for numerous customers of approximately equal purchasing power), the economic transaction takes the form of a sale; in command economics it is more usually a commission. In guild systems, the association of producers may itself transact the sale or at least regulate the conditions of sale, by its members, of their products. In a dealer system, which was known in classical antiquity but (in the visual arts) has developed most fully since the 17th century (in the Netherlands) and the 19th (in France), a group of commercial intermediaries—dealers and artist's agents—has stepped in between individual artists and individual purchasers. Their task is to introduce the one to the other and conduct the economic transaction in accordance with the practices of good business. (Some art dealers and their literary equivalents—writers' agents—have operated also as patrons and even sponsors of young artists.)

Methods of payment

In market systems, methods of remunerating artists depend on the traditional cultural reputation of the art they practice, the degree to which the artists' rights of authorship are legally recognized in an economically consequential manner, and the technological requirements for transacting the sale of an artwork.

When the traditional cultural reputation of an art is high, the artists who engage in it, while not necessarily highly rewarded in economic terms, tend to acquire a moral right to a socially recognized authorship of their works. Their names are affixed to the works they have produced; these are no longer products of anonymous craftsmen or works signed by the supervisors of art workshops, as were lacquerworks from the imperial workshops of China.

The conception of individual authorship began developing in Greece around 700 BC and in China more than 1,000 years later. It was known in the medieval West and India, but the notion of individual authorship remained undeveloped in the Byzantine civilization. This conception is directly dependent on the development of a professional or gentlemanly or genius type of artistic culture. Indirectly, it is supported by a strong stress on individualistic values in a cultural tradition, but, as in eastern Asia, the notion of authorship could arise even in the absence of any strong commitment to individualism.

The notion of a legally protected intellectual property of an art creator in his work, which he retains even after he has sold his work to a user, presupposes a conception of authorship but extends beyond it. Historically, it is a much later development. It began in the West in the late Middle Ages and has been encouraged by technological inventions facilitating the reproduction of works of art (before book printing, possession of a manuscript implied the right to reproduce it); the growth of a competitive market in such reproductions, which needed to be regulated (the most prominent motive in the early phases of the copyright law was the desire to protect the economic interests of the book publishers, rather than the intellectual rights of the authors); and the image of the artist as a man of genius who may sell, like other producers, the works of his labour but retains a right to the spiritual substance, uniquely his own, that he has invested in them. What the copyright and unfair-competition laws protect are tangible contributions to a product that remain identifiably individual. The artist has a legal right only to what he has worked out; the law protects the labour of elaboration, not the idea or intuition. Yet, once the notion of a legally protected intellectual ownership has taken root, it tends to become self-perpetuating, even in the absence of the conditions that have favoured its development.

Implication of intellectual ownership

The retention of a moral right in the artistic product provides the rationale for an author's sharing in the income that the buyer derives from the use of his work and from the appreciation of its value over time. The logic of this system has been worked out most fully in the publishing and the mass-communications industries (radio, film,

and television). In the visual arts, efforts are being made to extend the economic implications of the recognition of the author's moral right to his work by developing a model contract for the sale of a work of art that will not only guarantee to the artist control over the reproduction of his work but also provide for his share in any profits from the resale of his work. U.S. copyright law provides that ownership of a copyright is distinct from ownership of any material object in which the work is embodied, thus giving the artist control over its reproduction, and that the transfer of a material object, in the absence of an agreement in writing to the contrary, does not transfer ownership of the copyright of any work embodied in the object.

The contemporary copyright system guarantees to the writers covered by it that their incomes will be proportional to the current economic value of their works (the total income from their sale). Visual artists' income, on the other hand, is unrelated to the current economic value of their works once these have been sold (except indirectly, as the prices of the works they will sell in the future are affected by appreciation or depreciation of the works sold earlier). Thus, there is a disproportionality in the effects of the law on the relationship between the artist and his product in literature and the visual arts.

In the performing arts, unless the performance is recorded, the performing artist cannot be guaranteed any direct economic benefits from being recognized as the author of his performance, because the performance "vanishes" after it is completed. It is only what is being performed (the text, choreography, musical composition), and not the performance itself, that can be protected by custom or the law. There is thus a difference between the legal rights of performing and "producing" artists in their art, and this difference may well affect their overall social status and self-esteem.

Legal rights in the mass arts

The development of the mass-arts industry has made it possible for performing artists to enjoy the benefits of legal protection of their continued "intellectual possession" of their performances, if they have been made to be recorded and can be reproduced. Recordings of live performances have remained unprotected by copyright law and frequently are "pirated," or reproduced without permission. In 1971 an international convention prohibited piracy of musical recordings, and member countries then began developing legislation on the matter. The U.S. copyright law was amended in 1972 to prohibit the unauthorized duplication of sound recordings. If only in the mass-arts industry, the situation of performing artists has been made similar to that of such "producing" artists as writers and composers, who receive royalties on their past work. The performers in a film continue to receive a set rate of payment from each showing of that film and residuals from reruns on television. While in this way they have a type of continuing property right in their performances, as specified in their contract, performers have no moral right of authorship that would permit them to control the manner in which their work is presented to the public (as writers of books do). This right is usually vested in the film director, and even then, only if it is specifically assured to him in his contract.

Whether the legal protection of artistic property has in fact encouraged the production of high-quality art appears debatable: the system does not distinguish between high- and low-quality products. The operation of the legal-protection systems in the arts must be judged by the degree to which it helps artists both to earn an income and to derive self-respect from their work. It must, however, also be judged by inquiring whether its provisions result in larger benefits to the artists collectively or to the sellers of their works collectively.

ART COLLECTING

Art collecting and the building of architectural ensembles, such as churches and palaces, which can be regarded as immobile "art collections," have probably provided the most important source of income for the major visual artists, particularly in command economies. Once artworks enter a collection, however, they tend to stay there for long periods of time, potentially reducing the space and the demand for the works of later artists. There is thus a certain ambiguity in the attitude of artists toward art collections and museums, particularly in societies that possess well-stocked ones, such as Italy or, by now, the United States. Museums that tend to "enshrine" artists are in this respect worse, from the living artist's point of view, than private collections, which get reshuffled every once in a while—a process encouraged, at all times, by defeats in war, social upheavals, economic troubles, and, in modern societies, by high inheritance taxes. In contemporary societies, collecting by nonartistic organizations—governmental agencies, business firms, universities—seems likely to become increasingly important. This would mean a greater incorporation of art into real-life (as contrasted with museum) environments. As an encouragement for living artists, collecting seems to be most effective when any particular collection is intended to be temporary: that is, when it is established, exhibited, and then dispersed to make room for a different one.

Relations between museums and living artists

Purposeful collecting of works of art is most eagerly pursued in periods of great affluence and a reduction of the creative drive—in the Hellenistic age (323–30 BC), in ancient Rome, and in 18th-century Europe. The relationship between art collecting and political power has been more ambiguous. On the one hand, the powerful have liked to surround themselves with the grandeur emanated by an "overpowering" collection of art. On the other hand, states have compensated for their loss of power by cultivating an image of their cultural grandeur and refinement; art collections serve this purpose, too. But art collecting is also an appropriate response to great outbursts of creativity that have occurred in the past. And it establishes reservoirs from which later artistic efflorescences will, in part, be fed: art collections, if open to the public, provide a place for young artists to study their craft and for people in general to develop an interest in the arts.

Collecting of the art of the past may be encouraged by a lack, or loss, of faith in the creative accomplishments of living artists; by the proved prestige of the works of the past (important when the newly rich collector has neither taste nor understanding to judge by himself); and, particularly in contemporary society, by the profitability of economic investments in famous works of art. It is estimated that art prices in general have multiplied more than 10 times since the early 1950s.

The rise in art prices

Private art collecting on a smaller scale has been developing in the middle classes since the beginning of the modern age. Currently it shades off into "temporary collections" of multiples from the popular market, such as musical records or posters used to decorate student rooms. Attitudes derived from this practice are spreading even into sophisticated responses to art (*e.g.*, conceptions of the social function of museums).

SOCIAL CONTROL OF ART

Little art, except perhaps in the "intimate" genres such as lyrical poetry, can have been created, until the latter part of the 19th century, without some degree of social control—that is, influence exerted by nonartists—over its creation. The influence has varied greatly in degree, in the groups and agencies exerting control, and in the means used. Six basic types of social control of art can be distinguished: (1) suppressive censorship by agencies in total control of channels of possible expression (the "medieval system"); (2) product specification by a relatively few customers, each of whom is not in total control of channels of artistic expression (the "Renaissance system"); (3) administrative restriction of the artist's access to his potential audience without a complete withdrawal of his opportunity to communicate and without destruction of his works ("enlightened censorship"); (4) organizational incorporation of artists into either nonartistic institutions, such as churches or business firms, artistic groups under the authority of nonartistic agencies, such as the academy under the French king Louis XIV (reigned 1643–1715), or state artists' unions (the "organizational system"); (5) expression of preferences of taste by large numbers of individual art consumers not differing greatly in the power to

command and to purchase (the "democratic system"); and (6) criticism by experts specializing in sustained analysis and evaluation of individual works of art (the "intellectual system").

Censorship compared with criticism. Of the six modes of control, censorship is the most severe and least sensitive to the aesthetic merits of works of art. Art is on principle judged in terms extrinsic to artistic values and subordinated to explicitly utilitarian considerations of political, religious, or "moral" character. Criticism, at the other extreme, at least uses the means of imagination, rather than of power, to control products of the imagination and mediates between artists and groups of users of art (however small) whose standards particular critics articulate. Thus, criticism helps to relate artists to the subcultures in which their work is most responsively consumed and to reveal the character of these subcultures to their own members. At best, criticism expands the artist's experience in areas relevant to his understanding of how art operates in the minds of people and increases the consciousness of the community to which the critic "speaks." At the same time, the artist can avoid being influenced by critical interpretations (at the risk of not learning what they reveal) by the simple decision not to read the reviews. Scientific research into empirically demonstrable effects of particular kinds of art on particular types of personalities, under specified conditions, can be treated as a newly evolved element of artistic criticism.

Benefits derived from criticism

Censorship of the arts has been typically justified by "clerical" or "civic" conceptions of art—the view that art has a moral obligation to defend a cultural tradition higher than art itself or a civic obligation to celebrate the community that supports the artist. The first view is usually espoused by various ideological elites, but the second can be quite popular and widely supported, even in a democracy. The clerical conception of art also generally implies the assumption that art has a great power to corrupt or to save. It is, indeed, this tendency to overestimate the power of art that leads to demands for censorship. The civic attitude toward art presupposes merely a tendency to suppress anything that offends local self-esteem.

Whatever the conception of art, censorship becomes "necessary" only when an established clerical or civic view of art is powerfully challenged by other artistic cultures—the ecstatic, the vagabond, the genius, the professional, or the avant-garde. It is perhaps from such challenges to civic and clerical assumptions by ecstatic (Dionysian) or emerging professional types of artists that the first explicit philosophical defense of artistic censorship, by the philosopher Plato, emerged in classical Athens. A conflict between two clerical cultures, the Catholic and the Protestant, together with the invention of the printing press (which made literature more "dangerous"), led to the great development of formal, organized religious and political censorship in 16th-century Europe. Censorship declined, especially in the Anglo-Saxon countries, in the 18th century, with the establishment of religious tolerance and in conjunction with the displacement of clerical by professional conceptions of the writer.

Targets of censorship. Censorship has been mainly directed against the ideological (stated or implied views) or depictive (represented scenes) content of the arts, thus primarily against the arts in which content is more important—literature, the theatre, and the visual arts (including the film and, to a lesser degree, photography). But whole types of art have been outlawed. The ancient Spartans expunged music and dance, as well as poetry, on the grounds that they might promote effeminacy and license in a population that had to be hardened for heroic militarism. Early Christianity suppressed the theatre and fictional literature of the Greco-Roman civilization. Muslims, Calvinists, and in some periods the Byzantine iconoclasts outlawed religious visual art.

Suppression of content

While, in general, secular states have been concerned only with the content of art, ideological organizations have been sensitive, especially in 20th-century totalitarian systems, to the attitudes and values suggested also in style. It was primarily by stylistic characteristics that "degenerate" art was defined in Nazi Germany. Totalitarian movements have not only prohibited some styles but have also prescribed others for their artists to work in (for example, Socialist Realism in the Soviet Union between the 1930s and the 1950s).

The 17th-century absolutist state also perceived the value implications of artistic styles, but (like the medieval Catholic Church in its approach to the visual arts) it relied more on the techniques of product specification—patronage, by a royal court, and promotion, by an official art academy—rather than on prohibition, enforced by police power, against working in particular styles and having works in these styles exposed, in some manner, to artists' customers. As a mode of control, product specification is more congenial to artistic creativity than is censorship; and 17th-century France sustained creative attainments of a high order, even in the visual arts in which the court specified its demands most insistently. But it achieved still more in dramatic literature, over which the court had a less direct influence.

In the 20th century, the large economic organizations that have come to dominate the mass arts have acquired a capacity to exercise a private kind of censorship by depriving artists of their means of work. In the "image" industries, the reason for such blacklisting, as in the U.S. film industry after World War II, has tended to be the political image of the artist rather than either the style or the content of the work he was proposing or had done in the past. In the democratic societies, private watchdog and pressure groups are likely to be more effective in imposing their demands for censorship on the mass-arts industries or on the business firms whose advertising sustains them financially than in influencing governmental agencies. The newer mass media are, because of their dependence on public licensing (television stations), more vulnerable to governmental pressures than the traditional arts.

Self-censorship

Self-censorship by artistic enterprises (for example, the movie rating system of the U.S. film industry) has developed largely in response to the influences private pressure groups have brought to bear on the culture industries, by threatening their mass sales. A different variety of self-censorship is practiced by art museums when they ban types of art with contents that are not congenial to the economic or political interests of the business leaders who usually control their boards of trustees, or that seem capable of causing libel suits, or that appear to conflict too sharply with the traditional conceptions of what art museums should exhibit.

In neither of these two situations does "censorship" completely prevent public exposition of the works censored or terminate the public career of the artist concerned. These cases thus do not represent true, suppressive censorship but, rather, a system of limiting the public's opportunities for viewing certain types of art. Prohibition of the sale of certain types of art to juveniles also represents restrictive censorship. In respect to the arts, the Western democracies have by now virtually abandoned suppressive censorship in favour of the restrictive.

In the U.S.S.R., there have been some trends after the death of Joseph Stalin (1953) toward a transformation of the system of suppressive censorship into a de facto system of administrative restriction: allowing for privileged exhibitions of modernistic art in scientific institutes, showings of avant-garde foreign films to select circles (partly specialists and partly political elite), publication in small editions or in recondite journals. But a somewhat relaxed suppressive censorship is still in operation. China has a completely suppressive system.

In contrast to the censorship practiced by the ideological organizations, style is usually of no concern to the economic interest and private pressure groups endeavouring to subject art, or what is presented as art, to censorship either by pressuring the mass-arts industries and large artistic enterprises or by demanding action by government agencies (in the United States, most frequently on grounds of "obscenity").

CONDITIONS FOR SOCIAL CONTROL

In general, the arts seem to be most susceptible to social control when artists are dependent on a relatively few im-

portant consumers or on a great mass audience of a fairly homogeneous social character, and when their works are either very expensive "uniques" or highly profitable "multiples." The susceptibility of artists to social control thus depends, to some extent, on the kinds of art they choose to produce. This susceptibility tends to decline when there is a differentiated, heterogeneous public with political institutions permitting a free expression of individual choices and an inexpensive access to a wide range of works of art. If access to art is provided at public expense, however, artists again become susceptible to social control—this time by the intermediaries staffing the cultural organizations, such as museums and state publishing houses, that determine which artists and which works will be exposed to the public.

Factors in the susceptibility to restraints

Art is least susceptible to social control when it is sponsored and financially supported by other artists, who acquire their income in some manner other than through the sale of their works (for example, through teaching). The institutionalization of the professional conception of art is perhaps, in the long run, the most reliable defense, from within the artistic enterprise, against deliberate manipulation of the arts by social agencies. But a purely professional tradition of art is so inoffensive as to appear dehumanized, and artists themselves may seek to overthrow it, opening themselves up to deliberate manipulation by those (frequently political movements) they align with to achieve this goal.

IMPLICATIONS OF SOCIAL CONTROL

The issue of censorship and other kinds of deliberate manipulation of art has ultimately to be dealt with in terms of moral and political assumptions about human nature and the proper character of the artistic enterprise, not primarily in terms of the effects of these policies on artistic creativity. The effects of social restraints, or their absence, on artistic creativity seem, in any case, somewhat ambiguous. In general, no direct relationship seems to exist between the degree of personal freedom enjoyed by the artist and the aesthetic quality of his work. The tolerance and cultivation of the art patrons in Italy during the 17th and 18th centuries did not prevent a decline in creativity. On the other hand, a general cultural repressiveness does not always preclude an artistic efflorescence: the period of the most intense activity of the Spanish Inquisition in the 16th and 17th centuries coincided with El Siglo de Oro (The Golden Age), one of Spain's most important periods of artistic creativity.

Freedom, however, may have become a necessary—though not sufficient—condition for artistic creativity in industrial societies. Artists, particularly those who have been affected by artistic developments in the West since the Renaissance, have come to expect creative freedom and cannot help but feel illegitimately constricted in its absence. Their personalities are no longer sufficiently congruent with an authoritarian structure of external controls, as the personalities of medieval artists may have been, to be able to produce aesthetically significant work when subjected to such controls. By abolishing artistic freedom, both the Soviet Union and Nazi Germany destroyed flourishing artistic movements and proved unable to generate aesthetically valid substitutes for them. But a relaxation of earlier rigid controls may be more conducive to artistic creativity than the maximum of freedom, because under restrictive controls an unspent tension accumulates, which is then available to be released in an explosion of creative activity when pressures are relaxed. Under complete external freedom, such energy may never accumulate; unless the artists possess an extraordinary degree of self-discipline, their energies tend to be immediately used up through a variety of outlets.

Effect of easing controls

What seems to be specifically fatal to artistic creativity is not social control of the arts but an imposition, whether by outsiders or indeed by the artists themselves, of a completely intentional conception of the artistic enterprise: that is, its successful limitation to the expression of any set of recognized and intended functions.

If there is to be any social control over the arts, which many artists would dispute, its least objectionable form would seem to be one limited to a combination of expressions of preference by large numbers of interested art users with analysis and reasoned evaluation by art critics and scholars. The two modes of control in conjunction balance each other's biases and increase the range of options available to artists. These controls can be dismissed from consideration only if it is assumed that art is totally irrelevant even to the people most interested in it (other than the artists themselves). Art would then have to be treated as either a private affair of the artists or as pure entertainment without any cultural significance.

SOCIAL RELATIONSHIPS

In both preliterate societies and historic civilizations, the arts have frequently, but not always, had a close relationship with religion. In the past, this relationship has tended to be less direct when art objects were being produced by women, such as the pottery of the Pueblo Indians of the southwestern United States. Women, who have rarely been the religious specialists of their societies, have mostly produced an art of decorated utilitarian artifacts or of personal intimate expression, in either case not characterized by high cultural symbolism. Men have been more frequently disposed, or constrained, to justify their aesthetic interests by relating them to a metaphysical purpose. In effect, this means that men have had more reasons, in most preliterate and historical societies, for making art than women did.

Interaction of art and religion. An explanation for the frequently close relationship between art and religion may be found in the areas in which they are similar or overlap. In both art and religion, there is much concern with the basic needs of the imagination and with a valid perception of subtle qualities of experience, and there is no binding requirement that they provide references to demonstrable fact for their insights (as there is in science). The similarity means that the arts and religion might, in part, operate as functional substitutes for each other: the more successfully one system functions, the less need for the other, and the less successfully, the more need. This may be one reason for the modern growth of interest in the arts, especially among the intelligentsia and the alienated of the middle class.

Religious and artistic similarities

Religion, however, tends to be more dependent than the arts on those aspects of culture that are communal, involved with abstract ideas, and concerned with standards of conduct. It must provide guidelines for action that whole communities (as well as individual persons) could live with. While the arts can perform such "religious" functions, they can also survive quite well without performing them: inherently, the arts have a less direct, less intentional, less "responsible" relationship to social action than does religion. In their specifically aesthetic essence, the arts are more private, more individualizing, less binding than religion. There is more play than obligation in the arts. The opposite is true of religion. Popular interest in the arts tends to decline, and interest in religion to increase, during life-threatening historical crises, such as the Black Death of 1348–50 in Europe or destructive wars. Interest in art frequently increases after such periods. Religion is more of a crisis-management phenomenon, art that of postcrisis integration.

The impact of art on its consumers is likely to be magnified by its association with religion or even, in the absence of any formal association with organized churches, by its perception in terms appropriate to religious experiences. If close alliance between the arts and religion means that they are "used" by religion, the arts become collective liturgies, providing sensuous elements to increase the hold of a religious doctrine over the more private aspects of human experience. If, on the other hand, the arts "interact" with religion, if artists can have an influence on the development of religious orientations (instead of merely being "guided" by them), religion might acquire some of the characteristics of an art—become less systematized, more private, less dogmatic in its claims, more individualizing in the experiences it permits. The medieval Catholic Church, by and large, "used" the arts, while the Asian religions—other than Confucianism—tended to "interact" with the

arts. In classical Greece, the arts became so emancipated from religion that they can be said to have "used" it, with more enduring benefits for the arts than for religion.

Separation of religion from art. A sharp separation of religion from the arts, such as tended to occur during the Protestant Reformation, promotes the rise of a rationalized, "disenchanted" religious world view and an anarchically romantic tradition of the arts. Like the Reformers, religions generally have been willing to accept some kinds of music and literature and have varied mainly in their attitudes toward the visual and the "bodily" arts—the dance and the theatre. Attitudes toward the religious significance of the materially existent underlie this variation. But religions have also varied in the degree to which they made use of, or interacted with, the arts. "Aesthetically deprived" religions lose one sort of appeal to potential adherents and, perhaps especially in times of widespread discontent and cultural crisis, do not compete well for popular support with religions that possess more powerful aesthetic (or mythological) resources. Soviet explanations of the survival of religion in the U.S.S.R. stress increasingly its aesthetic aspect—a problem for Marxist ideologists, who are beginning to perceive a lack of this element in their own system of faith.

Effect of separating art from religion

A religion or a secular ideology without the arts loses one type of symbolic resource for its perpetuation and a possible source of modification of its doctrine. In addition, the arts represent one possible way of testing human significance of various aspects of the message of a religion or secular ideology. Do they survive, when artistically treated, without the support of enforcement by the established authorities of a church (or party)? It has been suggested that some ideological and religious concepts failed to be retained in the popular consciousness because they never received an effective artistic elaboration.

For the arts, a complete separation from religion or secular ideology means, on the one hand, a release from the obligation to serve a tradition regarded as superior to art itself and from its various efforts at social control—including censorship—over artistic expression. On the other hand, the separation of art from religion or ideology means: (1) the loss of a symbolic resource that either can be used directly in art creation or that stimulates artistic creativity indirectly by "disturbing" the imagination of artists; (2) the loss of one type of opportunity to create art that is both appreciatively used by large numbers of people and regarded as significant by them; and (3) the loss of a type of patronage that has historically been more continuously interested in the arts (other than literature), in times good and bad, than any other and that has tended to exhibit more concern even for the aesthetic merits of art than did the two most important types of patrons—the high bourgeoisie and the secular state—that followed the virtual demise of church patronage.

Attitudes toward patronage

A preoccupation with patronage, however, has not always governed the behaviour of artists. Even though painters could expect to lose their most important kind of patronage if the iconoclastic Protestant Reformation succeeded, a surprising number of them supported the Reformation. Historically, literature has been least dependent on religious patronage; music (because of its almost ubiquitous linkage with ritual), perhaps most.

AESTHETIC INFLUENCES

A strong influence of doctrinal religion over the artistic enterprise disposes it toward a "clerical" conception of its task: the upholding of a cultural tradition regarded as higher than art itself. It also tends to eliminate women from roles, except auxiliary, in the artistic enterprise—as in organized religion. But the *mystical* streak of religious experience has frequently influenced the artistic enterprise in a generally liberating direction, producing an ecstatic type of artistic culture. Any kind of direct religious influence on the arts tends to give them a more consciously symbolic character, but the doctrinal religions tend to impose on them an authoritarian rigidity (for example, the Romanesque style current in Europe from the 11th to the 12th centuries), whereas the mystical religions produce tendencies toward a more fluid style that could be interpreted as more egalitarian (Buddhist wall paintings at Ajantā in Mahārāshtra State, India). Mystical religions may also open the arts more equally to both sexes, as it did for such women mystic writers of the late Middle Ages as the English Julian of Norwich and the Italian St. Catherine of Siena.

Feeling-oriented religions like Buddhism tend to produce preferences, in the personalities influenced by such religions, for sensuous styles of art; belief-oriented religions like Calvinism favour austerity in art styles.

Confucianism has also tended to favour austerity, but the Chinese Confucian gentlemen painters drew most of their artistic inspiration from Taoism and variants of Buddhism. Perceptual tendencies may persist even after the religious tradition that has shaped them is no longer consciously adhered to.

Changes in the religious system have, in strongly religious periods, provided impulses for subsequent artistic changes. Periods of increased artistic creativity have frequently followed those of intense religious struggle—conversions of nations, great heretical movements, successes or failures of popular reformations, even iconoclasm. But this is not an effect peculiar to religious changes: other types of intense social action have also preceded artistic efflorescences. A more specifically religious influence on artistic creativity is suggested in the historic tendency for artistic efflorescences to occur after periods of great religious creativity. This suggests that the arts benefit from a partial secularization of intensely religious traditions or that religious efflorescences shape symbolic or emotional resources that are most usable for artistic expression when they have aged ("mellowed") somewhat but have not been completely discarded from the living experience of a people. Expirations of previously potent religious traditions, as of Buddhism in India, have coincided with declines in artistic creativity.

Inferior influence of secular "religions"

The modern secular equivalents of religion have generally been quite inferior to it in their effects on the arts. This may be partly because most of the "secular religions" of 19th-century origin have been rationalistic, purposive ideologies with little mythological content to disturb and stimulate the artistic imagination. In fact, artists have been most affected by the drama of events (or "theatre") brought into being by secular religions, rather than by their "mythology."

A great many modern artists have been willing to let themselves be influenced by the secular ideologies or by the more ancient religious traditions of their own civilization or by religions of cultures alien and exotic to them. The latter two influences have generally been aesthetically more auspicious than the former, but most artists do not experience even these influences with sufficient intensity to be "disturbed" by them and tend to exploit, almost at random, their more superficial characteristics.

It would be difficult to demonstrate that developments in the sphere of art have had any independent effects on religion. Throughout much of history, art has not been sufficiently independent from religion to be perceived as a "cause" of religious developments. More likely, it may have given focus, a tangible concreteness, a dramatic shape, a memorable melody, to abstract religious notions or shapeless feeling states and provided means for celebrating and transmitting them—rather than initiating the experiences and the notions themselves.

The impact of fantasy

It seems likely that the cultural attitudes of the growing number of alienated intelligentsia of the West are greatly affected by its artistic culture. The conception of the "revolution" as espoused by the New Left has, in many ways, been a "surrealistic" one ("all power to the imagination," etc.). Fascism, too, has been considered by some observers to be an aesthetic phenomenon almost as much as a political one. It is, in any case, through the shaping of fantasy dispositions—congenial modes of perception—that art can influence the development of general value orientations, which in turn dispose people to favour particular religious or political ideologies, when choice between alternatives is possible.

Yet the shaping of fantasy dispositions is not a foolproof method of manipulation, particularly in modern societies:

it may provoke a subjective rejection of overly insistent attempts to influence—an anarchic response to the rigid authoritarianism of the "classical" styles or, conversely, a demand for new dogmas and rituals in response to more ambiguity, suggested by the arts, than an individual can live with in his own life.

INFLUENCE OF TECHNOLOGY ON ART

To varying degrees, the arts depend on technological evolution for the very techniques used to create works of art—literature least, music considerably, and architecture most of all. The cinema has been made possible only by recent technological developments.

The arts also depend on technology for the dissemination of the creative product. Book printing has made possible the development of a mass reading public, which in turn has facilitated the rise of new literary genres, such as the novel. Modern techniques rendered objects of visual art mechanically reproducible, hence perceived and treated as less "unique" than they had been in the past. The film is a peculiar art in that the technique of production and the technique of dissemination immediately imply each other, both having been produced by the same technological development.

Beyond these direct technological influences on the arts, there is an indirect one, mediated through the effects that particular technologies have on the imagination of artists or of larger populations.

Effect on content. Generally, the arts do not mirror, in any direct manner, the technological developments occurring in the society in which they have been produced. In their contents, the visual arts are more likely to touch upon the consequences of technological developments that have become intimately familiar (for example, the artifacts represented in still-life painting) but are not necessarily its objectively most important result; or perhaps upon the more vividly visible agents of technological development (such as draft animals, machines, or electronic processes) rather than on the whole system of technology or the basic principles underlying it.

In other words, the arts focus on those aspects of technology that "grasp the senses." Thus, they frequently dwell on technological innovations in a very early stage of their development, then let them drop as subjects for artistic concern after they have become fully developed, attained a dominant role in the economy, and are taken for granted emotionally.

Art as a celebration of the absent

But it is also what technological change has eliminated or does not permit to exist that can grasp the artists' imagination; they will then be concerned with what they miss in a particular technological system and will depict, perhaps, the opposite of what they see as existing in it. It has been found that in preliterate societies where the house shape is circular, straight lines are preferred in art style and that, conversely, where the house type is rectangular, art styles tend toward the curved line. This finding suggests either that art supplies what is most lacking in the technological environment or that one art must provide what another art does not.

Technological developments may both suggest new contents for art and encourage the elimination of old ones. Thus, the invention of photography has virtually eliminated the need for realistic portraiture.

Effect on style. In the development of styles of art, technological factors appear to be more important than they are in the choice of subject matter. Certain styles have been made possible only by particular technological developments (electronic music). In other cases, technological developments (new types of paints) have converged with new scientific theories (in optics) and perhaps psychological changes (the decline of middle class democratic militance) to produce a particular style, such as Impressionism in French painting.

In still other cases, a style may be regarded as a symbolic projection of psychological attitudes inherently linked with a technological process. It has been argued that tendencies toward more geometric styles of visual art have emerged in connection with both of the major technological transformations of society—the agricultural and the industrial revolutions, while periods that have preceded these transformations have favoured more "realistic" styles (*e.g.*, Paleolithic cave paintings and the figurative arts of urban pre-industrial civilizations). A possible explanation of the linkage of the technological revolutions with more abstract styles of art may be the alienation from nature or a sense of mastery over it suggested by the geometric styles. Through them man imposes "his own" type of order on nature, where no such geometricism can be found. This attitude is also implied by the great technological transformations through which man has objectively increased his control over the forces of nature.

This process could have started in the imagination of artists inventing the notion of mastery over nature as a dream; in the attitudes of the people or of the religious, economic, or political leaders, reflected in both technological and artistic changes; or in the process of technological transformation, which, by its success, led to a widespread sense of mastery, projected into artistic style.

Industrial mastery and style

In the industrial transition the apparent sequence has been change in popular or elite attitudes, encouraged by the Protestant Reformation; speeding up of technological changes; and the rise of geometric styles of visual art. Music has "lagged behind" the technological transition even more than have the visual arts, but literature may, in certain respects, have anticipated it.

Thus, it could be argued that the poetry of courtly love and of religious affective mysticism of the late Middle Ages suggested an attitude of mastery over nature (over sexual impulses, limits of time and space). Even if this interpretation were accepted, however, it could not be concluded that literature is necessarily a more sensitive indicator of underlying psychological changes than the other arts. It may be in periods in which the other arts are heavily dependent upon organized patronage and a guild organization, and literature is comparatively free of such encumbrances, that it registers psychological changes more sensitively and at an earlier date. Technology, in this perspective, appears not as the "ultimate" determinant of artistic expression but as an aspect of basic psychological and cultural change—an aspect that became crucial at a certain point.

The interest of artists in science appears to have been greatest on the threshold of the Industrial Revolution, in the 17th century, and toward its end—when technological developments, particularly in the field of electronics, have altered the material basis of advanced societies in the second half of the 20th century. In between, artists, especially writers, tended to be repelled by technology and its effects on people and the environment (and by the science they associated with these effects).

In the latter part of the 20th century, the attitude of repulsion still tends to hold with respect to the machine technology, but a distinction is made between it and the electronic (and cybernetic) technology, which is viewed more optimistically. Electronic technology, by the miracles of its circuitry that exhibits an almost human responsiveness, is expected to overcome the chasm between the "mechanical" and the "spiritual" that the Industrial Revolution had deepened.

Changes in the character of science during the 20th century, its recognition of the principles of relativity and indeterminacy and of the importance of the researcher's subjectivity, have contributed much to the revived appeal of science to artists and to their high expectations with respect to the technology based on this kind of science. In the 20th century, psychoanalysis and several perspectives in the social sciences have also been influential in the arts.

The response of the arts to science

In the long run, it is likely that the artists' response to the sciences and technologies of advanced industrial societies will follow the usual logic of the workings of the artistic imagination. That is, scientific technology should influence the arts through the artists' choices between affirming it and modelling their own work after it or repudiating it and concentrating on what it cannot recognize or suppresses or, if left to itself, would tend to destroy. The power, mathematical clarity, and systematic nature of the sciences and technologies may produce an artistic response in the form of sensuous, amorphous styles, car-

rying suggestions of impotence, of mysticism, of unique experiences.

There is no reason to anticipate that, in the long run, artists will be overwhelmed by science and assimilated into its mode of operation. Indeed, the more science and technology develop, the more they may be subjectively taken for granted, the more imagination may be captured by what science and technology cannot encompass, and the more an art that has not been overwhelmed will be needed. Artists tend to be overwhelmed by a great increase in the technical possibilities of expression only when they have no reason of their own—a mode of perception, a value commitment—to express anything in particular.

Effect on creativity. An essentially descriptive science may influence artistic creativity favourably when it fuses with an older tradition of an artistic craft, as anatomical and optical research vitalized painting in the Italian Renaissance. But a highly abstract science, removed in its formulations from directly perceivable realities, seems to be of most benefit to artists when they permit their imaginations to be stimulated by the general atmosphere of scientific discovery, of the revision of concepts about the nature of reality, or of new perceptions, rather than when artists have set out to apply scientific principles consciously to their work. Even an incorrect interpretation of scientific models by artists can be artistically productive (and perhaps scientifically suggestive).

Other aspects of the relationship. To what extent the arts can influence the sciences is still uncertain. If "visual thinking" precedes conceptual thought, or if there is a basic background element common to both visual and conceptual thought, changes in perceptions that art either reflects or stimulates may generate responses in conceptual theorizing and, in this manner, affect developments in science as well as in philosophy. But it is on social and psychological theory and on the perceived shape of history that developments in the arts may be expected to have the strongest effect.

Effect of art on the social sciences

The association of artists with technology, especially in the visual arts and architecture, has been far closer and more durable than it has been with science, and the requirements and achievements of artists have often led to technological discoveries. Literary artists, on the other hand, have, particularly in the periods of great scientific discovery, such as the 17th century, been influenced more by science than by technology.

It might seem that the new technologies of the mass media of communication, by permitting more pervasive dissemination of particular works of art throughout society, are likely to increase the impact of art on people's personalities and modes of existence. But the mass media may also have the opposite effect: by making of art a commonplace occurrence that is taken for granted but does not generate an enduring emotional response, by assimilating it more completely to "entertainment," the mass media may decrease the effects of works of art on people and, indeed, on the artists themselves and on their aesthetic experiences.

At the same time, the media may be increasing the psychological impact of nonartistic events, particularly those that resist being assimilated to entertainment, such as wars in a foreign country, which were easier to take for granted before the electronic media became fully developed. This has led to tendencies to substitute, especially in the visual arts, the structures of events or of social systems for aesthetic structures (for example, an art exhibition consisting of photographs of slum buildings and landlords' names).

While it is debatable whether the mass media have enhanced the impact of art, they have certainly increased the social status of the performing artists (and tend to transform all artists whom they capture into performers).

AESTHETIC EDUCATION

The broad and narrow views of aesthetic education

In the broad sense, aesthetic education refers to everything that art is, or may be, used for in the education of nonartists. In the narrower sense, aesthetic education is the developing of a sensitivity to aesthetic qualities and works of art and of an understanding of the criteria by which some works of art are regarded, by artists and art critics, as possessing more highly valued aesthetic qualities than others. Aesthetic education could hardly avoid, but it does not need to be limited to, what is implied in the term's narrower meaning. Approaches to aesthetic education vary mainly in how they conceive of its "broader" responsibilities.

As soon as art is created and exposed to others, it always educates in some way. A study could well be made of the implicit philosophies of aesthetic education, inherent in the manner in which art is used and particularly in the manner in which the growing individual is exposed to it, of the preliterate societies in which the role of an art educator—as distinguished from the practicing artists—does not exist.

The formal development of aesthetic education, like the formalization of all education, has occurred in the classical civilizations. In the Western tradition, four basic conceptions of aesthetic education for the nonartist have been sociologically most important, although in practice they frequently overlap.

Didactic. The didactic theory regards art as a means for shaping a particular type of personality (the view of the Greek philosopher Plato). Modern notions of proletarian, black, or feminine "consciousness raising" by means of art, insofar as they presuppose that art is to be used for developing an attitude specified in advance, represent variants of the didactic approach to aesthetic education. But even when the goal is developing a particular type of aesthetic taste (as, for example, a taste for the "classical" styles or for "modern" art), a philosophy of education is didactic, albeit in a more subtle way. The didactic theory implies that if art does not do the task assigned to it, it has either no place in education or a subordinate one; and it may lead to justifying censorship of the arts, "to protect the innocent from corruption." This view of aesthetic education is most congruent with the "clerical" conception of art.

Therapeutic. The therapeutic theory views art as the supplier, for individuals or groups, of experiences that everyday life in society fails to provide but that are assumed to be necessary for a "whole" and "healthy" existence. In a highly rationalized society, art supplies or reveals the irrational, supplementing or confronting reason and duty with spontaneity and sensuousness. Or, in a more sophisticated conception, art reconnects reason with sensuousness, which had first to be separated from each other, for man's self-awareness to advance itself.

Since this philosophy of art as therapy developed in an early phase of the Industrial Revolution, it usually conceives the deficiencies of society in terms of what was becoming suppressed then—needs for emotional expression—or what was promised ideologically but has not been adequately delivered in reality even in the most modernized societies, such as equality and participation. But what appear to be increasingly missed in the advanced industrial societies are credible designs for the symbolic coherence of life; a sense for experiences and objects that resist being used up and are immune to planned obsolescence. The therapeutic function of aesthetic education may be changing accordingly.

Changes in the therapeutic function since the Industrial Revolution

The objective of a "therapeutic" aesthetic education is to use art to bring out the missing elements and either to promote their integration with the existing state of affairs or to overthrow the latter. The task may be conceived as "sociotherapeutic" or "psychotherapeutic"—overcoming the deficiencies of a whole civilization or those of an individual personality. In contrast to the didactic approach, only the initial problem and the general direction of the effort but not the final solution—a specified type of personality or society—are presupposed. Artworks can be judged, in a general way, by the degree to which they fulfill the therapeutic task set to art by the character of the society or the stage of the evolution of civilization contemporaneous with it. But no artwork can be in principle excluded from aesthetic education, since any work may be "therapeutic" for someone.

Developmental. The developmental theory aims at making it possible for anyone to choose from the realm of art whatever he needs to develop his unique potentialities.

The evolution of society has not given rise to any particular type of need that art should be meeting in a given social context. The task of aesthetic education is, perhaps, to increase the range of exposure to art so that every individual will choose on the basis of a more complete knowledge, more intelligently, how to develop his own self aesthetically.

This is a liberating educational philosophy, particularly in a tradition dominated by didactic approaches. But it has three unintended effects: (1) Since it puts the emphasis on what a person gets from a work of art for himself, it destroys the reason for trying to understand the work of art in itself and the ways it has functioned for other people, including its author. (2) While this approach permits individual experiences and descriptions of such experiences, it eliminates the basis for a more generalized critical evaluation of works of art. (3) By its emphasis solely on the individual's aesthetic experience, it eliminates the need for art. What have the works of art got that "aesthetic experiences" on a crowded street or in one's dreamworld do not give?

The developmental approach applied to children

The developmental approach is most applicable to the aesthetic education of children, where it is provided with a sense of direction by the psychologists' notion of the stages of intellectual and perceptual development of the personality, each offering distinctive possibilities and limits to aesthetic education. In the aesthetic education of adults, however, by implying that neither art nor the society nor history are particularly important as compared with a person's "experiences," this approach defines itself as a luxury object, likely to appeal to a social group that is prosperous but impotent to shape its destiny; it needs to be balanced by various other kinds of aesthetic education.

Culture-critical education. The purpose of another form of aesthetic education is criticism of culture, in the broad, anthropological sense, through analysis of concrete works of art and of their functions in the life histories of individuals and in the historical existence of societies (and particular groups within them). This culture-critical approach differs from the therapeutic in that it is based less on the personal feelings and cast of mind of the educator and anchored more in the empirical study of culture history and of people's relationships to the symbolic expression of their experiences. It considers art in the context of all human experience, not just in the narrow social setting of an individual perceiver of art (as the developmental theory necessarily does). It is not concerned with recapitulating human experiences in the technical detail of historical monographs. Instead, it is involved with the recognition of patterns in which experiences are "fitted together"—externally influenced, intuitively structured, and their qualities and meanings interpreted to affect later experiences.

Aesthetic education as culture criticism is concerned with producing an ability to make aesthetic judgments that are founded on the knowledge of the ways in which whole civilizations, as well as everything involved with them, have been working. An analytical understanding of parts and of particular relationships is encouraged to grow into a capacity to perceive and evaluate the overall connectedness of a civilization.

Aesthetic education as culture analysis is a method that can be adequately used only with adults; but it can be placed at the very centre of the education of adults, including college students. Its purpose is not to produce art critics but, rather, persons more competent of judging civilizations, including their own, on the basis of an increasing understanding of what they have done, or failed to do, to human sensibilities.

Supplementary approaches. The culture-critical approach to aesthetic education is predominantly analytical. It needs to be supplemented, for most people, by practice in the making of works of art. The writing and performing of dramatic works permit an exploration of the potentialities of social interaction and its limits, which a person has either had no experience with in his own life and yet senses as potentially significant or which he has experienced in an "incomplete" manner and whose full logic and his own design for interpreting it he wishes to work out. Other kinds of literature seem to have a similar role in aesthetic education. Music and dance clarify the dynamic rhythms in terms of which personal experiences and the character of civilizations are perceived, while the visual arts crystallize images of the emotional qualities these experiences and civilizations are sensed to possess. The film provides a potential basis for integrating the functions that literature, music, and the visual arts have in aesthetic education, but the separate experience of these arts provides a clearer understanding of such functions.

Architecture, with its commitment to relating dispositions of the imagination to the practical exigencies of life—and therefore to social policy—in a viable overall design of aesthetic merit, deserves, perhaps, a central place in aesthetic education that has not been recognized in any large-scale educational system. It may have a particularly important role in the education of those who refuse to recognize the linkages of their imaginations with the requirements of actuality.

The importance of aesthetic education beyond purely artistic considerations

It would appear to be arbitrary and self-defeating to limit aesthetic education entirely to considering the work of art in its isolation. It also must give some attention to the ways in which societies organize, or fail to organize, themselves to build aesthetically adequate whole environments (physical, social, and even "spiritual"). It must demonstrate how individuals can demand and, by their own political actions and spending patterns, support the building of such environments. A complete program of aesthetic education should include a consideration of the costs of building and tearing down (or maintaining) aesthetically inadequate environments as compared with the costs of building environments capable of adequately performing a great many cultural functions for the people inhabiting them. And it must teach the techniques of social action for insisting effectively that the kind of aesthetic environment needed is provided.

The importance of immediate contact with practicing artists in the aesthetic education of nonartists is not primarily in the teaching of the technical skills of making art, which the practicing artists are not necessarily superior to "art educators" in transmitting. What good practicing artists can provide for aesthetic education is the demonstration of how "aesthetic structures" (meaningful connections between disparate elements of experience) arise from skilled labour under the constant judgment of a personal sensibility. Good artists also reveal in workaday practice how the nature of a subject, a design, a genre evolves in the process of elaborating its implications and simultaneously develops intrinsic requirements of its own, a set of "norms," a "logic," that the artist cannot disregard without diminishing his creative attainment. In this way, practicing artists transmit a sobering sense for what is required of creators by the inner logic of their works.

The role of the art critic in aesthetic education is to clarify the criteria by which he judges the aesthetic fitness of works of art and distinguishes between artistic successes and failures. The critic should be distinguished from the art interpreter, who explicates the meanings that a work of art has for him. The aesthetic philosopher's most productive role in aesthetic education would be to compare, in some systematic manner, the criteria of the critics, the meanings of the interpreters, and the intentions and the practices of the artists of various societies. In practice, the aesthetic philosophers mainly study each other. Partly for this reason, a role in aesthetic education has been opening up for the sociologists and psychologists of art, who are concerned with how art actually functions in the life of societies and personalities.

Since different individuals are likely to benefit most from different aspects of aesthetic education, it would seem to be a mistake to have a single model of aesthetic education for everyone.

THE NATURE OF ART PRESERVATION

Art preservation in preliterate societies

Even in preliterate societies, not all art is created for the occasion of its use and then abandoned. First of all, the basic design of the work of art survives in the collective memory of the tribe, or of its more artistically inclined members, and can be reproduced from this recollection, frequently with creative modifications. This principle of

preservation operates most clearly in oral literature, but it gives continuity also to the other arts, both performing and objectifying. And this sort of preservation is not limited to preliterate societies.

In some cultures, art objects may also be preserved either for their utility or for their religious significance. Cave paintings have been continuously "refreshed"—that is, restored by overpainting—over long periods of time by the Australian Aborigines to retain the benefit of their magical effectiveness.

Even before the rise of literate civilizations, art collecting had become a symbolic exhibition of wealth and power. Forms of writing, which made easier the collection and preservation of literature, probably also were developed, in the classical civilizations, for their utility in economic management, the pursuit of religiously significant activities such as astronomy, and the more effective exercise of political power. But once traditions of collecting works of art and of literacy have evolved, they have tended to acquire a degree of autonomy from the purposes they may originally have been associated with. Systems of musical notation have evolved later than literacy and without any significant economic or political motives to necessitate their development; it was encouraged, however, by a religious need to stabilize the liturgical uses of music. Systems of dance notation have been produced for purely artistic purposes. Systems of photographic, sound, and motion recording have been provided by the progress of modern technology. The development of these systems has been propelled more by curiosity than by the anticipation of the great profits they eventually proved capable of producing.

Collections of manuscripts existed in the ancient world, and a system of state and school libraries was established in Rome. But a network of public libraries has evolved only since the middle of the 19th century. The notion of a collection of books has not been as closely associated with the aristocracy as that of a collection of works of visual art: the chief connection of libraries has been with scholarship and its practical applications (in preaching, in administration, and, in modern times, in self-advancement through education).

Art collecting in the industrial age

The nature of art collecting has also changed in the age of industrialization and the democratic revolutions that started in the second half of the 18th century. Art collections, which previously had been possessions of the monarchy, privileged classes, and the church and (except for the visible part of church art) were rarely opened to the public, became, in most cases, public museums.

The traditional association of high art and art collecting with class privilege has led some of the revolutionaries and avant-gardists of the 19th and 20th centuries to conclude that the establishment of a democratic culture requires the destruction of the monuments to an aristocratic culture collected in the museums—and of the museums themselves as repositories of that type of culture.

Not only art museums but also symphony orchestras and even the theatre (in contrast to Elizabethan times, 1568–1603) have, in Western societies, little attraction for the working and lower middle classes. A study made in France in the 1960s found that 1 percent of museum attenders were agricultural labourers, 4 percent industrial workers, 5 percent artisans and tradesmen, the rest white-collar workers and higher social classes. In eastern Europe, where there is more of a tradition of high culture being in alliance or, indeed, in secret emotional conspiracy with the people (an attitude of the folk culture that has been retained), working class attendance at performances or exhibitions of high culture is higher. But even though it is encouraged by government, party, and trade-union agencies (and indirectly by the monotony of much of the officially sponsored "popular" culture), interest in high culture is still stratified by class.

This lack of interest by manual labourers in art poses a contradiction. Works of art, by whomever they have been sponsored or collected, have always been produced by craftsmen who surpassed mere craftsmanship. They represent a conjunction of craftsmanship and sensibility of the men and women who have been in the vital centre of their own times, working and imagining. If there is a monument to the immortality of manual labour, it is a museum of art.

Museums of modern art

Museums devoted to contemporary art have developed only in the 20th century. Instead of preserving what has survived repeated tests of critical judgment over the ages, museums of contemporary art delve into the flux of ongoing artistic developments, at best endeavouring to sort them out into intelligible patterns and to enlighten the public's consciousness of its own times. Frequently, however, such museums have become powerful trend setters for the fashionable, producing an ephemeral new "movement" each year and becoming, in effect, adjuncts of the mass media of communication rather than seekers for the surviving values of art. In the 1960s, there developed demands that contemporary-art museums be conceived of as "houses of controversy"—where anything arousing concern of contemporary artists could be exhibited, whether it is art in the traditional sense or not. There are tendencies toward differentiation between two types of art museums: one engaged in controversy and the other a repository of art-historical collections whose holdings are critically evaluated for their aesthetic merit. There also seems to be an increasing need for museums in which all the arts of a period or a historic group could be shown together, to reveal the overall cultural atmosphere of the period or group and the interconnections of its arts, placed within their sociological context.

For each living generation to preserve artworks implies the desire not to deprive future generations of the kinds of human experiences that the circumstances of the past and present have generated and the future may no longer be able to produce. Possession of what one is not able to produce stretches the mind in ways that, in the absence of these possessions, could not even be imagined. By not preserving works of art, a society would lose much of its awareness of the limits of its own imagination and would tend to treat its present modes of existence and of aspiration as though they were absolute.

Unselective preservation of everything that has claimed to be art would, however, very quickly overload the depositories of culture. A decision to preserve therefore implies a criterion of selection to determine what is worth preserving. Such criteria may be the aesthetic merit of the work of art or its historical significance for a nation, a church, a political movement, or a family lineage. Preserving samples of the "popular culture"—presumably lacking in aesthetic merit but representative of the character of a period—is also justifiable. Since judgments of aesthetic merit are revised and possibly improved over long historical periods, "accidental" preservation of works that have not been judged by their contemporaries to possess aesthetic merit can also become important.

The greater part of the art produced by the preliterate societies and historic civilizations has been lost due to neglect and historical misfortune. Only 12 percent of the tragedies said to have been written by Aeschylus, Sophocles, and Euripides in Greece during the 5th century BC have survived. Of the period that the Chinese traditionally regard as representing the historical peak of their painting—the T'ang (AD 618–907)—hardly any paintings have been preserved. Many works of art have been intentionally destroyed by organizations or individuals (for example, medieval peasants burning classical marble statues for lime to fertilize their fields). Among the major social organizations, the churches have probably done more for the preservation of works of art than any other, but they have also destroyed art in religious wars and campaigns for the eradication of heresy. States have both protected and destroyed. Business organizations, on the other hand, rank among the major destroyers of architectural monuments, particularly in contemporary societies. Only by governmental regulation can the popular market, whether capitalist or Socialist, be restrained from destroying unique works of art to build parking lots.

Destruction of art

SYSTEMS OF DISSEMINATION

At least seven types of systems through which artistic products are disseminated to art consumers may be identified. These systems differ in the manner in which contact

with art is established and the attitude with which it is approached.

1. Ritualistic systems. Art is incorporated into the conduct of special, repetitive occasions and presented as an integral part of them (for example, church services and political rituals in which particular kinds of music are performed). Highly specific types of art receive wide exposure on an occasional basis. An air of the extraordinary, of festivity, is attached to these types of art. The modern opera has inherited some of the trappings of the ritualistic system. Certain avant-garde occasions may approach it. So do art festivals, by breaking with the routine of permanent museum collections or environments, which gives art the character of being taken for granted.

2. Environmental systems. Art is incorporated into the organization of the everyday, stable environment visible to everyone in the normal course of his life. Architectural works, landscaping, public monuments, and private libraries are cases in point. Transmitted in this way, art influences the perceptual expectations of most everyone, but it tends to be taken for granted, to become "invisible." In modern societies, economic organizations and governmental agencies are primarily responsible for the dissemination of art (bad as well as good) through this system. Governments could, indeed, do more for aesthetic education by controlling the environmental system than through their hold over aesthetic education in the schools.

Art in everyday utilitarian objects

3. The utilitarian system. Art is incorporated into the small, usually portable objects of everyday use, from eating utensils to automobile designs. In its effects, this system is comparable to the environmental system. But it is industry, rather than the government, that is in a strategic position to affect art's distribution through the utilitarian system.

4. The art-trade system. Art objects are produced for and sold on the market specialized for goods of their nature. Depending on their cost, they may appeal to a "mass" or a "class" public but, in any case, to those possessed of a purchasing power they can devote to objects of no material utility and who have a tradition of buying such objects. In practice, the system of trade in the objects of fine art tends most frequently to become oriented to economically privileged groups. The market for music records and tapes is, however, creating a tradition of buying art objects even among the underprivileged. The book trade, particularly after the development of the paperback in 19th-century Germany, is oriented to a broader section of the population than is the art trade; the book market tends to divide into an elite-oriented and a popular component, but the two overlap.

5. The mass-entertainment system. Artworks are constantly exposed, at low cost, to large audiences, but not as a stable and integral part of their everyday environment (as in the environmental and utilitarian systems) nor with an attitude of the extraordinary attaching to it (as in the ritualistic system). The ubiquity of exposure, the fleetingness of the occasions of exposure, and the general atmosphere of fun and games surrounding the operations of this system have the effect of obliterating the unique significance of any particular object of art. Everything processed through the mass-entertainment system tends to become like everything else.

6. The educational system. This is the concern not only of schools but also of modern museums and libraries. Works of art are collected and disseminated with an educational intent—in schools potentially to everyone in the right age groups, in museums and libraries to those who choose to make use of the education offered. Museums, however, have not been fully assimilated to the educational system. In the United States, they are still not regarded, in their claims for public support (or for an automatic tax exemption by virtue of their status), as quite equivalent to schools. There are also notable tendencies in the latter half of the 20th century to incorporate museums into the mass-entertainment system. On the other hand, a politically controlled system of "mass entertainment," as in totalitarian states, is in fact heavily "educational," in a didactic manner. The street theatre of the "counterculture" is also intended to be educational.

The educational system typically lacks the power inherent in environmental or utilitarian systems and the intensity characteristic of ritualistic systems. Possibly for these reasons, the art disseminated through the educational system tends to acquire connotations of what in the upper middle class of Western societies used to be perceived as "femininity"—a refined irrelevance to the world of affairs. The possession of "real life" power by the large art museums constitutes one of their educational advantages over art teaching in the schools, where the art teacher is still frequently treated as an inferior type of educator.

7. Movement systems. There are indications that a new system of dissemination of art may be emerging in the youth movement and perhaps other popular movements of a basically "cultural" (rather than traditional "political" or "economic") character that have become a feature of advanced industrial societies. Any particular movement system is temporary, but, while it lasts, it generates an intense audience response to certain kinds of art, and the audience may even be drawn into the collective production of works of art. The dividing line between producer and consumer of artworks becomes attenuated. Art can be consumed without waiting for a set occasion, but its consumption always has an aura of rebellious celebration attaching to it. If movement systems are successful, however, they tend to be assimilated into mass-entertainment systems—revitalizing them, but at the same time losing their own vitality.

While all known societies produce (and therefore presumably need) some kind of art, not all their individual members use art or exhibit any kind of response to it, even when it is offered to them at no cost and in an environment that is not intimidating. Therefore, all societies have some system, or a combination of systems, for the dissemination of works of art, but not all of their members are involved with such systems. While art-dissemination systems might aim at a universal exposure to works of art, they cannot be blamed for not generating a universal response to art.

The sources of the responsiveness to art lie partly outside of the whole artistic enterprise, in the structures of individual personalities and in their experiences in confronting social systems and historical processes. In this sense, an individual's contact with art starts not with society's systems of art dissemination but with the character of his existence. (V.Ka.)

Fraudulence in the arts

The most common type of fraudulence in art is forgery—making a work or offering one for sale with the intent to defraud, usually by falsely attributing it to an artist whose works command high prices. Other fraudulent practices include plagiarism, the false presentation of another's work as one's own, and piracy, the unauthorized use of someone else's work, such as the publication of a book without permission of the author; both practices are generally in violation of copyright laws.

Types of fraudulence

Forgery most often occurs with works of painting, sculpture, decorative art, and literature; less often with music. Plagiarism is more difficult to prove as fraud, since the possibility of coincidence must be weighed against evidence of stealing. Piracy is more often a business than an artistic fraud; it frequently occurs in the publication of editions of foreign books in countries that have no copyright agreement with the nation in which the work was copyrighted. A stage production, the reproduction of a painting, the performance of a musical composition, and analogous practices of other kinds of works without authorization and royalty payments also fall into this category.

The fundamental consideration in determining forgery is "intent to deceive." The act of copying a painting or other work of art is in itself not forgery, nor is the creation of a work "in the style" of a recognized painter, composer, or writer or of a particular historical period. Forgery may be the act not of the creator himself but of the dealer who adds a fraudulent signature or in some way alters the appearance of a painting or manuscript. Restoration of a damaged painting or manuscript, however, is not

considered forgery even if the restorer in his work creates a significant part of the total work. Misattributions may result either from honest errors in scholarship—as in the attribution of a work to a well-known artist when the work was in fact done by a painter in his workshop, a pupil, or a later follower—or from a deliberate fraud.

Excluded from the category of literary forgeries is the copy made in good faith for purposes of study. In the matter of autographs, manuscripts in the handwriting of their authors, forgeries must be distinguished from facsimiles, copies made by lithography or other reproductive processes. Some early editions of Byron's work, for example, contained a facsimile of an autograph letter of the poet. If such facsimiles are detached from the volumes that they were intended to illustrate, they may deceive the unwary.

Motivations for fraudulence

The commonest motivation for fraudulence is monetary gain. Fraudulence is most likely to occur when the demand for a certain kind of work coincides with scarcity and thus raises the market prices. Unprincipled dealers have encouraged technically skilled artists to create forgeries, occasionally guiding them to supply the precise demands of collectors or museums. This is by no means a modern phenomenon: in the 1st and 2nd centuries AD, sculptors working in Rome made replicas of Grecian works to satisfy the demands for the greatly admired Grecian sculpture of the preceding five centuries. These copies or adaptations apparently were not offered as contemporary work but as booty from Greece at the extraordinarily high prices paid for such works in imperial Rome. Similar circumstances may account for the "discovery" of a manuscript or autograph by a dead author or composer, although many such finds are quite legitimate and have been authenticated.

The history of the arts reveals instances of persons who have used forgery either to gain recognition of their own craftsmanship or to enjoy deceiving the critics who had rejected their genuine work. A legend told about Michelangelo illustrates this point. At the age of 21, he carved in marble a small sleeping Eros, or Cupid, based on ancient Roman works that he admired. Some time later this carving was sold as an antique to the well-known collector Cardinal Riario, who prized it highly. When Michelangelo stepped forward and claimed the work as his own he won immediate fame as a young man who could rival the work of the greatly venerated ancient sculptors.

Two further motivations behind forgery must be noted: forged documents have been produced from time to time to exalt or malign some religion, political party, or race; and forgeries are sometimes created as a hoax. Some hoaxes are intended to confound or ridicule the experts; others are intended to parody or burlesque an artist or genre.

Three methods of forgery

There are basically three methods of producing a forgery: by an exact copy, by a composite of parts, and by a work done in the style of an artist or period and given a deliberately false attribution. These methods apply most directly to the visual arts but can be discerned in literature and music as well. (Ed.)

LITERARY FORGERY

Financial gain is the most common motive for literary forgery, the one responsible for the numerous forged autographs that appear on the market. The popularity of such authors as the Romantic poets Burns, Shelley, and Byron led to the fabrication of numerous forgeries of their autographs, some of which remain in circulation. These forgeries were usually made by men who had access to only one or two genuine specimens, which they began by tracing. Their forgeries are stiff, exaggeratedly uniform, and lacking in the fluency and spontaneity of genuine autographs.

Instances of literary forgery. Occasionally a forger appears with a certain specious glamour like Constantine Simonides (1824–67), a Greek adventurer who varied his trade in perfectly genuine manuscripts with the sale of strange concoctions of his own. Maj. George de Luna Byron, alias de Gibler, who claimed to be a natural son of Byron by a Spanish countess, successfully produced and disposed of large quantities of forgeries ascribed to his alleged father and to Shelley, Keats, and others. More commonplace is the forgery encountered in the case of the Edinburgh forger A.H. ("Antique") Smith, who was responsible for forgeries of Robert Burns, Sir Walter Scott, Mary Stuart, and other persons from Scottish literature and history—a feat that ultimately earned him 12 months' imprisonment.

Particularly notorious was the case of the Wise forgeries. Thomas James Wise (1859–1937) had the reputation of being one of the most distinguished private book collectors on either side of the Atlantic, and his Ashley Library in London became a place of pilgrimage for scholars from Europe and the U.S. He constantly exposed piracies and forgeries and always denied that he was a dealer. The shock was accordingly the greater in 1934 when John W. Carter and Henry Graham Pollard published *An Enquiry into the Nature of Certain Nineteenth Century Pamphlets,* proving that about 40 or 50 of these, commanding high prices, were forgeries, and that all could be traced to Wise. Subsequent research confirmed the finding of Carter and Pollard and indicted Wise for other and more serious offenses, including the sophistication of many of his own copies of early printed books with leaves stolen from copies in the British Museum.

No forgery to attain recognition is better known than the "Thomas Rowley" poems of Thomas Chatterton (1752–70), which the youthful author attempted to pass off as the work of a medieval cleric. These poems, which caused a scholarly feud for many years, were influential in the Gothic revival. Chatterton, however, enjoys a place in English letters as a creative genius in his own right. The more conventional forger William Henry Ireland (1777–1835) cheerfully manufactured Shakespearean documents until his forged "lost" tragedy *Vortigern and Rowena* was laughed off the stage at the Drury Lane Theatre, London, in 1796. More fortunate was Charles Bertram, who produced an account of Roman Britain by "Richard of Westminster," an imaginary monk. Bertram's dupe, the eccentric antiquary Dr. William Stukeley, identified the monk with the chronicler Richard of Cirencester, known to have resided at Westminster in the 14th century. Bertram's forgery (cunningly published in a volume containing the works of two genuine ancient authors, Gildas and Nennius) had an enormous influence upon historians of Roman Britain, lasting into the 20th century. Equally influential were the Ossianic poems of James Macpherson (1736–96), which influenced the early period of the Romantic movement. To what degree Macpherson's poems are to be regarded as spurious is not certain. Denounced in his own day they were possibly, as he claimed, based upon a genuine oral tradition of Scottish Gaelic poetry; but there can be little doubt that they were carefully edited and interpolated by their collector.

Among the forgers who have tried to make the experts look foolish is George Psalmanazar (1679?–1763). A Frenchman, he went to England where he pretended, with great success, to be a native of Formosa (Taiwan), and published a book about that island, which he had never visited. Another is William Lauder, who attempted to prove Milton guilty of plagiarism by quoting 17th-century poets who wrote in Latin, into whose works he had interpolated Latin translations from *Paradise Lost.* A forgery made as a joke but taken seriously was the "Ern Malley" poems, offered to an Australian magazine in 1944 as the work of a recently dead poet. Actually it was composed by two young soldiers who wished to ridicule certain aspects of contemporary poetry.

The pure fabrication is a kind of forgery that defies classification, often because there is no false attribution and the motives are difficult to ascertain. An example of this is the *Historia regum Britanniae* (*c.* 1135) of Geoffrey of Monmouth (died 1155), a pseudo-historian who compounded stories from Celtic mythology and classical and biblical sources into a fictitious history of ancient Britain. The book became one of the most popular of the Middle Ages and was the basis for some Arthurian legends recounted in medieval romance and epic.

Detection of literary forgeries. The scientific examination of a forged document may demonstrate its spurious character by showing that the parchment, paper, or ink cannot belong to the period to which they pretend. A

skillful forger takes care, however, to secure appropriate materials; and in any case, scientific examination will not avail against the contemporary forger, living in the same age as his victim. Accordingly, other tests must be employed.

Forgeries may be detected by the methods of examination formulated by Jean Mabillon, in his great work *De re diplomatica* (1681), for determining the authenticity of a document by the writing and the style of the terminology. These techniques have developed during three centuries into the modern sciences of paleography and diplomatic, by which various scripts and formulas can be assigned to particular ages and localities, and effective comparison can be made between two examples of handwriting purporting to come from the same pen. Thus it is possible to state that a particular document could not have been written at the date that it bears. In dealing with printed texts, analogous methods are employed.

Nevertheless, a forgery may pretend to be no more than a copy of a genuine original. It then becomes necessary to examine the language and style in which it is written and to look for anachronisms or for statements that conflict with known authorities. This is the method of textual criticism brilliantly employed by Richard Bentley in his *Dissertation upon the Epistles of Phalaris* (1699), which proved that these letters, far from being written by a Sicilian tyrant of the 6th century BC, were, in fact, the work of a Greek sophist of the 2nd century AD.

While the detection of the careful forger may require an expert, forged literary autographs can often be detected by anyone taking the trouble to compare them with an authentic example. Many collectors have been deceived by their own credulity, because they wished to believe that they were getting a good bargain and subconsciously suppressed their critical faculty. A classic case is that of the French forger Vrain-Denis Lucas, who sold a collection of forgeries including a letter of St. Mary Magdalene, written in French on paper made in France. (Ge.B.)

FORGERY IN THE VISUAL ARTS

Any art object—paintings, sculpture, jewelry, ceramics, fine furniture, and decorative pieces of all kinds—can be forged. The difficulty of forging, however, is as important as market price in determining what is forged. Probably fewer than 1 percent of stone sculptures are false because they require so much labour to make and their market is limited, but as many as 10 percent of modern French paintings on the market may be forgeries. The technical difficulties in making a convincing imitation of an ancient Greek vase are so great that forgeries are almost nonexistent. In contrast the forgery level of tiny archaic Greek and Cretan bronze statuettes, which are simple to cast, is possibly as high as 50 percent. A forger is most likely to succeed with a mediocre piece in the middle price range because such a piece probably will never be subjected to definitive examination. Although the price should be low enough to allay suspicion, the object can still yield a fair return for the effort expended by the forger.

Copies

The copy is the easiest forgery to make and is usually the easiest to detect. When a duplicate has appeared the problem is merely to determine which is the original and which is the copy. At least a dozen excellent replicas of Leonardo da Vinci's "Mona Lisa" exist, many of them by his students. Various owners of these copies have at various times claimed that they possess the original. The Louvre is satisfied that it owns the painting by Leonardo because close examination reveals slight changes in the composition underneath the outermost layer of paint, and because this painting has an unbroken record of ownership from the time that the artist painted it.

A monumental sculptural forgery was a copy based on a Greek bronze statuette of a warrior of 470 BC, only five inches high and located in the Antikenabteilung, Berlin. The forgers made an eight-foot-high reproduction of it in terra-cotta and offered it as an Etruscan masterpiece. The resemblance was noted by the experts, who thought it to be an example of an Etruscan artist borrowing a Greek design motif. In 1961, after it had been in the Metropolitan Museum of Art in New York for 40 years, an analysis was made of the black glaze that covered the figure. It was found that the glaze contained as a colouring agent manganese, which never was used for this purpose in ancient times. Finally, Alfredo Adolfo Fioravanti confessed that he was the sole survivor of the three forgers.

Fine examples of pottery and porcelain have always commanded high prices, which have, in turn, encouraged the making of forgeries and reproductions. Since many European factories tried to imitate Italian majolica during the 19th century when it was especially popular, forgeries are common. The work of Urbino, Castel Durante, Faenza, and Gubbio was copied freely, and, to a lesser extent, so were the wares of Orvieto and Florence. Most of these forgeries are not close enough to deceive a reasonably expert eye. Potters used natural deposits the impurities of which, for good or ill, often affected the final result; until recently it has been impossible to procure materials in a pure state. In all but a few isolated instances (some German stoneware reproductions, for example) the forger no longer has access to these original deposits and he has to imitate the effect of the impurities as best he can. Although the best forgeries are often remarkably close to the originals, they are not very numerous.

Composition fraud

In the composite fraud, or *pastiche,* the forger combines copies of various parts of another artist's work to form a new composition and adds a few connecting elements of his own to make it a convincing presentation. This type of forgery is more difficult to detect than the copy. Such a combining of various elements from different pieces can be very deceptive, because a creative artist often borrows from his own work. In fact, the similarity of a figure or an object in a forgery to that in a well-known work of art often adds to the believability of the new creation.

The Dutch forger Hans van Meegeren employed a combined composite and stylistic procedure when he created seven paintings between 1936 and 1942 based on the work of Jan Vermeer. In "Christ at Emmaus" he combined figures, heads, hands, plates, and a wine jar from various early genuine Vermeers; it was hailed as a masterpiece and the earliest known Vermeer. Ironically, Van Meegeren never was detected as a forger. At the end of World War II he was arrested for having sold a painting attributed to Vermeer to one of the enemy, and was accused of being a collaborator. He chose to reveal himself as a forger, which was a lesser offense, and proved his confession by painting another "Vermeer" while in prison.

A variation in composite forgery, quite common with inlaid French furniture, involves the use of parts from damaged but genuine pieces to create a single complete piece that may or may not resemble one of the pieces from which it has been made. These made-up pieces are still considered forgeries. In composites of archaeological material only one part may be ancient, the balance being made up to complete the object. The head of a small terra-cotta figure may be ancient, the body and limbs of modern workmanship. A single ancient element in a composite forgery will help to deceive the buyer.

Stylistic fraud

Most difficult of all to detect is the forgery done in the style of a particular artist or age. If the forger is skillful and is able to absorb the attitudes, conventions, and techniques of the period, he can often create a very successful piece of duplicity.

The work of the Italian Alceo Dossena belongs in this class. He very competently forged works that were acquired by collectors and museums throughout the world. From 1916 to 1928 he produced hundreds of forgeries created as original expressions of archaic Greek, medieval, and Renaissance sculptors.

A newly discovered type of art inevitably brings on a flood of forgeries. At the end of the 19th century, when the first small, attractive Tanagra figurines were found in Greece, the market very shortly was flooded with a myriad of fraudulent Tanagra terra-cotta statuettes. In the mid-20th century, African primitive art became very popular, and woodcarvers from Italy to Scandinavia responded to supply the demand. Later, a very early civilization was discovered in Turkey, and the few genuine Anatolian ceramic pieces that appeared on the market were followed immediately by very competent forgeries apparently made

in the same location as the ancient pieces. The lack of knowledge about genuine pieces made detection extremely difficult.

Detection of forgeries in the visual arts. The key to detecting forgery of unique objects lies in the fact that every object has within itself evidence of the time and the place in which it was made. The two main approaches, stylistic and technical analysis, are complementary and are best used together.

tylistic nalysis

Stylistic analysis is subjective: it rests on the astute eye of the art historian. Each artist has a style, a flair, a verve unique to himself, and this can be recognized. His style will undergo change throughout his career, and this, too, can be stylistically analyzed and documented from his known works. When an unknown work purporting to be by a certain artist is discovered, the art historian attempts to fit it into the overall body of works by this artist. The subject matter, the brushwork, the choice of colours, and the type of composition are all consistent elements in a given artist's production. Any variation immediately arouses suspicion. When the idiosyncrasies of an artist's brushwork are studied, a fraud can sometimes be detected in much the same way a handwriting forgery is proven. In ancient works, particularly in antiquities, the scholar must examine the iconography of a piece. Forgers rarely have the scholarly background to combine iconographic elements correctly, and their errors often betray them.

An object must also be studied for its purpose. Ancient works were made for functional purposes. A forger usually makes an attractive piece often inconsistent with that purpose. As they were used most ancient pieces developed signs of wear. These rubbed and worn areas should appear in logical places on the object.

Documentation is also an important area of investigation. The apparent authenticity of many spurious pieces is bolstered by false documents to attest to the point of origin, former owners, and expert opinions concerning the pieces. A careful examination of these records often detects the forgery.

The hardest deception to detect is usually one that has been made recently. The forgery is a product of the time in which it was made, and the forger is closer to current understanding of the artist or period forged. The forgery, therefore, is often more appealing than a genuine work of art. As a forgery ages, viewpoints and tastes shift, and there is a new basis of understanding. Consequently, a forgery rarely survives more than a generation.

Technical analysis

Technical analysis, an objective approach, rests on an arsenal of equipment and tests. The fundamental principle is the comparison of a suspected work with a genuine work of the same artist or period. The suspected piece must show the same pigments or materials used and comparable age deterioration. Inconsistencies automatically cause the piece to be suspect. Oil paintings dry out and develop a crackle, bronzes oxidize, and ancient glass buried in the ground develops iridescent layers. The microscope is the most useful basic tool: a close examination of the physical condition often will show if the aging is genuine or has been artificially induced. The type of tools used by the artist can be detected from an examination of their telltale traces.

Ultraviolet rays readily reveal additions or alterations to a painting, since the varnish layers and some of the paint layers fluoresce to different colours. Ultraviolet is also used in the examination of marble sculpture. Old marble develops a surface that will fluoresce to a yellow-greenish colour, whereas a modern piece or an old surface recently recut will fluoresce to a bright violet. Infrared rays can penetrate thin paint layers in an oil painting to reveal underpainting that may disclose an earlier painting on the same canvas, or perhaps a signature that has been painted out and covered by a more profitable one. X-rays are used to examine the internal structure of an object. A carved wooden Virgin supposedly of the 15th century but revealing modern machine-made nails deep inside is obviously a fraud. A forger usually works for the surface effect and is not concerned with the internal structures.

Sometimes it is necessary to remove small bits of materials from a work and subject them to various analyses. Chemical analysis is particularly valuable in determining the pigment used because many of the paints available to the modern forger were unknown in earlier times. Today titanium, a 20th-century product, is used to make the white pigment in most oil paints, whereas white lead was the element used in the time of Rembrandt. Many ancient colours were manufactured by grinding natural minerals such as lapis lazuli for blue and malachite for green. Today cheaper synthetic chemicals are used. Some chemical tests, however, require the removal of more ancient material than is desirable. In that event a speck as small as the head of a pin can be analyzed spectrographically. From the burning of a minute sample a photographic record of the spectrum of the light emitted is analyzed to reveal the elements present and their relative percentages.

The dating of an object by the study of radioactive decay of carbon-14 has had little application in the detection of art forgery because of the large quantities of material that must be destroyed. Thermoluminescent dating is based on the slight damage to all matter, including clays, by the faint nuclear radiation present in the earth. Magnetic dating of ceramic objects is based on the slow but perceptible shift of the earth's magnetic field over the centuries.

Considerations of aesthetics and risk. One may logically question the real meaning of the difference between a genuine and a spurious work of art when in many cases it requires such expert study to detect the difference between the two. Or to phrase it another way, what is the difference in value of a work of art that has been on view in a museum for 40 years, after it has been proved to be false? This is a somewhat philosophical point in that the object itself has not changed, only our opinion of it. Its monetary value has been reduced from that of a rare, expensive, original piece to that of an attractive but spurious imitation. Its aesthetic quality has become a real danger, as it is a perversion of the truth. The forgery presents a false understanding of the work of an artist or an ancient culture, one which has been perverted in its modern translation. To appreciate the work of ancient artists their work must be studied alone and not be diverted by forgeries, or one will be inexorably misguided.

Despite all the studies and technical tests available, forgeries will still be made. The 20th-century art forger is far better equipped and much more knowledgeable than his predecessor. The demand for rare works of art has increased, and he will attempt to supply them. In collecting, whether by the private collector or by a museum, there comes a point when, after all the studies and all the tests are conducted, a decision has to be made as to whether or not to purchase a piece in question. The element of risk can be minimized but not eliminated. At this point, the collector should be ready to back his opinion with the purchase price. In order to acquire great pieces, particularly from newly discovered and relatively unknown cultures, it is necessary to take a calculated chance. The collector who has never bought a forgery probably has never bought a great piece of art. (J.V.N.)

BIBLIOGRAPHY

Preparation of the artist and conditions of work in the arts. There is no one book that satisfactorily covers all of the issues discussed in these sections. There are, however, a number of excellent works that deal in detail with one or another phase of the subject. CARL ROEBUCK (ed.), *The Muses at Work: Arts, Crafts, and Professions in Ancient Greece and Rome* (1969), contains a mass of fascinating information. RUDOLF and MARGOT WITTKOWER, *Born Under Saturn: The Character and Conduct of Artists: A Documented History from Antiquity to the French Revolution* (1963), is an invaluable work. NIKOLAUS PEVSNER, *Academies of Art: Past and Present* (1940), is not only the standard book on its subject but also contains many penetrating observations on the profession of the arts generally. QUENTIN BELL, *The Schools of Design* (1963), adds much interesting and useful detail to Pevsner's book and explores some new ground. CESAR GRANA, *Modernity and its Discontents: French Society and the French Man of Letters in the Nineteenth Century* (1967), is a brilliant study that was originally published in 1964 as *Bohemian Versus Bourgeois.* Much information about the profession of the arts may also be found in books on the history of the various arts, such as the volumes in the "Pelican History of Art" series; or JOHN REWALD'S magnificent two books on *The History of Impressionism,* rev. ed. (1961),

and *Post-Impressionism* (1962). DONALD DREW EGBERT, *Social Radicalism and the Arts: Western Europe* (1970), provides a very full treatment of the effect of radical thought on the arts from the French Revolution to the present.

(Si.T.)

Social and economic aspects of the arts. *Anthologies:* On the visual arts in preliterate societies, see CAROL F. JOPLING (ed.), *Art and Aesthetics in Primitive Societies: A Critical Anthology* (1971), an important reader. On the arts in historic societies, MILTON C. ALBRECHT, JAMES H. BARNETT, and MASON GRIFF (eds.), *The Sociology of Art and Literature* (1970), has very extensive coverage and bibliographies. The best overview of the sociology of literature is LEO LOWENTHAL, "Literature and Sociology," in JAMES THORPE (ed.), *Relations of Literary Study: Essays on Interdisciplinary Contributions,* pp. 89–110 (1967).

Art and society: ROBERT ESCARPIT, *Sociology of Literature* (1970); I.C. JARVIE, *Movies and Society* (British title, *Towards a Sociology of the Cinema,* 1970); and ALPHONS SILBERMANN, *Wovon lebt die Musik?* (1957; Eng. trans., *The Sociology of Music,* 1963), provide overall analyses of the social organization of particular arts. CESAR GRANA, *Bohemian versus Bourgeois: French Society and the French Man of Letters in the Nineteenth Century* (1964); FRANCIS HASKELL, *Patrons and Painters: A Study in the Relations Between Italian Art and Society in the Age of the Baroque* (1963); BARRINGTON KAYE, *The Development of the Architectural Profession in Britain: A Sociological Study* (1960); HARRISON C. and CYNTHIA A. WHITE, *Canvases and careers: Institutional Change in the French Painting World* (1965), are studies of changes in the social role of particular types of artists. An account of a social setting in the performing arts (the Broadway theatre) is SAMUEL W. LITTLE and ARTHUR CANTOR, *The Playmakers* (1971).

Basic works on social factors in the style and content of the arts: WALTER ABELL, *The Collective Dream in Art: A Psycho-Historical Theory of Culture Based on Relations Between the Arts, Psychology and the Social Sciences* (1957); PIERRE FRANCASTEL, *La Réalité figurative: éléments structurels de sociologie de l'art* (1965); ARNOLD HAUSER, *Sozialgeschichte der Kunst und Literatur,* 2 vol. (1953; Eng. trans. of vol. 1, *The Social History of Art,* 2 vol., 1957); VYTAUTAS KAVOLIS, *Artistic Expression: A Sociological Analysis* (1968); ALAN LOMAX, *Folk Song Style and Culture* (1968); LEO LOWENTHAL, *Literature and the Image of Man: Sociological Studies of the European Drama and Novel, 1600–1900* (1957); RENATO POGGIOLI, *Teoria dell'arte d'avanguardia* (1962; Eng. trans., *The Theory of the Avant-Garde,* 1968); PITIRIM A. SOROKIN, *Social and Cultural Dynamics,* vol. 1, *Fluctuation of Forms of Art* (1937); and MAX WEBER, *The Rational and Social Foundations of Music,* ed. by D. MARTINDALE, J. RIEDEL, and G. NEUWIRTH (Eng. trans. 1958). For studies of historic cases see LUCIEN GOLDMANN, *Le Dieu caché: étude sur la vision tragique dans les Pensées de Pascal et dans le théâtre de Racine* (1956; Eng. trans., *The Hidden God: A Study of Tragic Vision in the Pensées of Pascal and the Tragedies of Racine,* 1964); PHILIP E. SLATER, *The Glory of Hera: Greek Mythology and the Greek Family* (1968).

Social conditions of artistic creativity have been studied in VYTAUTAS KAVOLIS, *History on Art's Side: Social Dynamics in Artistic Efflorescences* (1972); and A.L. KROEBER, *Configurations of Culture Growth* (1944). Influences of art on society are theorized in HUGH DALZIEL DUNCAN, *Communication and Social Order* (1962); and RADHAKAMAL MUKERJEE, *The Social Function of Art* (1948). VICTOR W. TURNER, *The Ritual Process: Structure and Anti-Structure* (1969), provides a framework for analyzing the social effects of the performing arts. For a survey of some empirical findings, see JOSEPH T. KLAPPER, *The Effects of Mass Communication* (1960).

Art and economics: There is no basic overall inquiry into the role of economics in the development of art. Good historical studies of art collecting are FRANCIS HENRY TAYLOR, *The Taste of Angels: A History of Art Collecting from Rameses to Napoleon* (1948); and NEILS VON HOLST, *Künstler, Sammler, Publikum* (1960; Eng. trans., *Creators, Collectors, Connoisseurs: The Anatomy of Artistic Taste from Antiquity to the Present Day,* 1967). For trends in art prices, see GERALD REITLINGER, *The Economics of Taste,* 3 vol. (1961–70); and GERALDINE KEENE, *Money and Art: A Study Based on the Times-Sotheby Index* (1971). On economic support of contemporary arts in the United States, see WILLIAM J. BAUMOL and WILLIAM G. BOWEN, *Performing Arts: The Economic Dilemma* (1966); and WILLIAM JACKSON LORD, *How Authors Make a Living: An Analysis of Free Lance Writers' Incomes, 1953–1957* (1962). In nine western European nations, see FREDERICK DORIAN, *Commitment to Culture: Art Patronage in Europe, Its Significance for America* (1964). On both the development and the current status of the legal rights of artists, see BRUCE W. BUGBEE, *Genesis of American Patent and Copyright Law* (1967); BORIS I. GOROKHOFF, *Publishing in the U.S.S.R.* (1959); GEORGE HAVEN PUTNAM, *Books and Their Makers During the Middle Ages,* vol. 2 (1897); HOWARD WALLS, *The Copyright Handbook of Fine and Applied Arts* (1963).

Art and politics: There is no reliable survey of relationships between art and politics. Good examples of monographic studies of particular historic cases include: HUGH DALZIEL DUNCAN, *Culture and Democracy: The Struggle for Form in Society and Architecture in Chicago and the Middle West During the Life and Times of Louis H. Sullivan* (1965); DONALD DREW EGBERT, *Social Radicalism and the Arts, Western Europe: A Cultural History from the French Revolution to 1968* (1970); JAMES A. LEITH, *The Idea of Art As Propaganda in France, 1750–1799* (1965); GEORGE LACHMANN MOSSE (ed.), *Nazi Culture: Intellectual, Cultural, and Social Life in the Third Reich* (Eng. trans. 1966); HAROLD SWAYZE, *Political Control of Literature in the USSR, 1946–1959* (1962).

Art and religion: Relationships between religion and the arts are surveyed in GERARDUS VAN DER LEEUW, *Vom Heiligen in der Kunst* (1957; Eng. trans., *Sacred and Profane Beauty: The Holy in Art,* 1963). For sociological studies of religion and the visual arts, see the works of Abell, Francastel, Kavolis, and Sorokin listed above. Of numerous historical studies, G.G. COULTON, *Art and the Reformation,* 2nd ed. (1953); and JEAN GIMPEL, *Les Batisseurs de cathédrales* (1958; Eng. trans., *The Cathedral Builders,* 1961), deal with the effects of religion in organizing the arts, while HELEN GARDNER, *Religion and Literature* (1971); and ANDRE MALRAUX, *La Métamorphose des dieux* (1957; Eng. trans., *The Metamorphosis of the Gods* 1964), with the "spirit" of religion in art.

Art, technology, and science: A historical survey is provided in CYRIL STANLEY SMITH, "Art, Technology, and Science: Notes on their Historical Interaction," *Technology and Culture,* 11:493–549 (1970). On the impact of contemporary science and technology on the visual arts, see JACK BURNHAM, *Beyond Modern Sculpture: The Effects of Science and Technology on the Sculpture of This Century* (1968); and C.H. WADDINGTON, *Behind Appearance: A Study of the Relations Between Painting and the Natural Sciences in This Century* (1969). On the influence of science on poetry, see DOUGLAS BUSH, *Science and English Poetry: A Historical Sketch, 1590–1950* (1950); and MARJORIE NICOLSON, *Science and Imagination* (1956).

Art and education: The most influential recent prescriptions for aesthetic education have been JOHN DEWEY, *Art As Experience* (1934); and HERBERT EDWARD READ, *Education Through Art* (1943). For psychological studies of perception and imagination in art, see RUDOLF ARNHEIM, *Visual Thinking* (1969); and ANTON EHRENZWEIG, *The Hidden Order of Art: A Study in the Psychology of Artistic Imagination* (1967). The utopian and the analytical dimensions of aesthetic philosophy may be illustrated by HERBERT MARCUSE, *Eros and Civilization: A Philosophical Inquiry into Freud* (1962); and RALPH A. SMITH (ed.), *Aesthetic Concepts and Education* (1970). See also *The Journal of Aesthetic Education* (quarterly).

(V.Ka.)

Fraudulence in the arts. *Literary forgery:* J.A. FARRER, *Literary Forgeries* (1907, reprinted 1969), provides a good introduction, which may be supplemented by H.T.F. RHODES, *The Craft of Forgery* (1934); and S. COLE, *Counterfeit* (1955). For individual forgers and forgeries, see E.H.W. MEYERSTEIN, *A Life of Thomas Chatterton* (1930); T.G. EHRSAM, *Major Byron* (1951); A.N.L. MUNBY, *Phillipps Studies,* vol. 4 (1956), on Constantine Simonides; and B.A. MORRISSETTE, *The Great Rimbaud Forgery* (1956). E.J. GOODSPEED, *Modern Apocrypha* (1956), gives an authoritative account of modern forgeries of Christian writings. On medieval forgeries, see the classic essay by T.F. TOUT, "Medieval Forgers and Forgeries," *John Rylands Library Bulletin,* 5:208–234 (1919); and for an example of forged charters, R.W. SOUTHERN, "The Canterbury Forgeries," *English Historical Review,* 73:193–226 (1958). On the detection of forgeries see W.R. HARRISON, *Suspect Documents: Their Scientific Examination* (1958), and *Forgery Detection: A Practical Guide* (1964); and J.V.P. CONWAY, *Evidential Documents* (1959).

Forgery in the visual arts: SEPP SCHULLER, *Fälscher, Händler und Experten* (1959; Eng. trans., *Forgers, Dealers, Experts,* 1960); and HEINRICH SCHMITT (pseudonym FRANK ARNAU), *Kunst der Fälscher, Fälscher der Kunst* (1959; Eng. trans., *Three Thousand Years of Deception in Arts and Antiques,* 1961), are both standard anthologies. DIETRICH VON BOTHMER and JOSEPH V. NOBLE, *An Inquiry into the Forgery of the Etruscan Terracotta Warriors in the Metropolitan Museum of Art* (1961), is the exhaustive technical and art historical study of an important group of clever forgeries. A similar article is JOSEPH V. NOBLE, "The Forgery of Our Greek Bronze Horse," *Bulletin of the Metropolitan Museum of Art,* 26:253–256 (1968). Stories of frauds told from the viewpoint of the forgers are given in LAWRENCE JEPPSON, *The Fabulous Frauds* (1970).

(J.V.N./Ge.B.)

Style in the Arts

Like much of the vocabulary of aesthetics, the word style resists straightforward definition. The word may point to little more than a mode or form of artistic production; or it can designate traits regarded simply as aids in the task of dating, grouping, and attributing works of art; it can imply skill, grace, or some other sort of excellence; it can mean a manner sanctioned by a standard; it refers to a mode, form, manner, tone, theme, subject, or quality—or a combination of such—that is felt to be characteristic enough to evoke a person, a group, a class, a nation, a place, a period, or a civilization; often also the reference is to features that are said to express an outlook, a doctrine, or a program. As a rule, even in the most carefully controlled context, several meanings will be present; and the tidy meanings will tend to bloom, or decay, into the untidy. Thus, although "sonata style," strictly constructed, should point only to a mode of musical production, in fact it will usually suggest a Classical outlook, the European 18th century, and perhaps the compositions of Haydn. Although to a field archaeologist "Late Helladic III" may designate merely a device for classifying pots of the ancient Greek city of Mycenae, in many imaginations the phrase is apt to inspire a vision of an entire culture, or perhaps of the legendary King Agamemnon bleeding in his bath after being murdered by his wife, as related in Aeschylus' tragedy.

The resulting confusion in thinking and talking about art is often deplored. One art historian likens style to a rainbow, a phenomenon of perception governed by the coincidence of certain physical conditions, which vanishes in the attempt to approach it. Another scholar takes the view that an adequate theory of style awaits a deeper knowledge of the principles of form construction and expression and a unified social theory comprising the practical means of life as well as emotional behaviour.

Such comments have not, however, had much effect; the majority of artists, critics, historians, and ordinary appreciators have continued to employ, loosely but confidently, the familiar term. And this persistence is not indefensible. Rainbows, after all, do exist; the chaser who fails to catch one demonstrates nothing except a mistake about their mode of existence. Also, in talking about art one can always cite the principle, first enunciated by Aristotle, that every study has its own degree of certainty and that a well-educated man will not ask for an unsuitable degree.

With that principle serving as a point of departure, the present article undertakes to expand and combine the several meanings of "style" into a discussion concerning the inner nature, the varieties, and the dynamics of the phenomenon as it manifests itself in all of the arts. (By "dynamics" is meant the patterns of movement and change that are the concern of style-conscious historians and biographers.) The whole discussion is presented on the level of a general theoretical introduction; material from the history of art and the history of aesthetics appears only in the form of examples, which are intended to provide clarification and some historical perspective.

The article is divided into the following sections:

The nature of style

General suppositions

It is easy to suppose that present notions about the nature of style are as old as the human ability to perceive differences, and a sampling of reasonably ancient cultural activity can seem to confirm the supposition. Much of the history of Chinese painting, for example, seems unthinkable without something like a modern critical apparatus for sorting out the dynastic periods, the local schools, and the copiers of venerated masters. Must not the ancient Greeks have been thinking of style when they noted the differences between Doric and Ionic orders in architecture? The way in which ancient Athenians contrasted the "rational" sound of the stringed cithara and the "irrational" reed-pipe wail of the aulos suggests the distinction between classical and romantic styles that was made in the 19th century, and the parodies of the tragic dramatist Euripides by the Athenian comic playwright Aristophanes imply modern conceptions of a personal style.

Much of what is now called stylistics seems to have existed already in the long succession of ancient Greek and Latin treatises that dealt with rhetoric. In sum, examples from both East and West can seem to support the assumption that there has always been something in the arts that may be called style.

INVENTION AND DISCOVERY

That "seem," however, needs very heavy emphasis. Throughout the history of art some unexpected facts have resided below the surface of the sort of sampling that has just been cited—facts that are open to more than one interpretation, but certainly not wide open.

Historical background. The admirer of Chinese painting who consults the old texts on the subject will find illuminating accounts of brushwork and much wisdom about the creative process but practically no discussion of style in the full modern sense of the word. In Indian aesthetics since ancient times, the doctrine of *rasa* (Sanskrit:

"essence") has been used in reference to the flavour or sentiment of works of art and to the modes of affective response to them, but not to what is properly called style. In the West, ancient writers on art exhibit a similar and finally rather enigmatic failure ever to focus squarely on the subject. Linguistic evidence, while inconclusive, suggests that in the Greco-Roman world there was no word that meant quite what is now generally meant by "style."

Vitruvius, the Roman authority on architecture, writing sometime during the 1st century BC about the Doric, Ionic, Corinthian, and Tuscan orders, avoided even the Latin word *ordo* ("arrangement") and contented himself with *opus* ("work") and *genus* ("kind"). The Greek traveller Pausanias, writing about the visual arts in the 2nd century AD, used *kataskeuē* ("device" or "method of fitting out") and *ergasia* ("work"). Among writers on literature and rhetoric a parallel tendency is apparent. Speaking of the style of an author, Aristotle was likely to refer simply to *lexis* ("speech" or "word"), and he was also likely to be talking merely about lucidity. The unidentified Greek critic known as Demetrius, writing probably in the Hellenistic era, used *charaktēr* ("quality" or "mark"), but the noun carried overtones from the verb meaning merely to scratch or engrave. The anonymous Latin text called the *Ad Herennium,* dating from around 85 BC, used *figura* ("figure"). A generation later Cicero used an arsenal of terms that included, with greatly varying degrees of precision, *figura, color, habitus* ("condition," "character"), *dictio* ("diction"), *elocutio* ("elocution," in a very general sense), and *genus.*

"Style" and *stilus*

Only in Late Latin does *stilus,* the word for the sharp-pointed instrument for writing, usually on wax, begin to mean also a manner of writing, as "pen" now does in such expressions as "a fluent pen" and "an acid pen"; and even here modern readers must be alert, for the derivation of English "style" from *stilus* does not prove that *stilus* always meant "style." The Latin term was reserved entirely for discussions of writing and speaking and usually for treatises on rhetoric; moreover, it seems to have implied little more than style in the sense of a skill or grace, and of a manner sanctioned by a standard. Apparently an author or orator in the closing years of the Roman Empire, in the 5th century AD, could have a periodic, loose, effective, ineffective, elevated, elegant, plain, high, middle, or low *stilus* but only very exceptionally, if ever, an idiosyncratic *stilus* that expressed a personality. And, again apparently, an architect, painter, sculptor, or musician could not have a *stilus* at all.

No important change in the usage thus established can be detected during the European Middle Ages. Words that suggest a kind, category, or mode of artistic production continued to be used in contexts in which a modern critic or historian might think in terms of a characteristic or expressive manner, or style. In architecture and the other visual arts, *opus* continued to be favoured; the French Gothic style of building was called Opus Francigenum, and English-style embroidery was know on the Continent simply as Opus Anglicanum. In music, the stylistic change apparent at the beginning of the 14th century, especially in France, was referred to as a new art: Ars Nova. The Latin *stilus*—and eventually its derivatives in other languages—was used only for talking about writing and speaking, and normally in the old rhetoricians' sense of a nonpersonal style sanctioned by a standard. When Dante (*Purgatorio,* xxiv) refers to the *dolce stil nuovo* ("sweet new style") that appeared in Italian poetry at the end of the 13th century, he stresses the importance not of the authors' personalities but of making the manner suit the matter and the occasion. The Host in Chaucer's *Canterbury Tales,* of the late 14th century, has this sense in mind when he addresses the Clerk of Oxford, in the prologue to the latter's tale:

> Your termes, your colours, and your figures,
> Kepe hem in stoor till so be ye endyte
> Heigh style, as whan that men to kinges wryte.
> Speketh so pleyn at this tyme, I yow preye,
> That we may understonde what ye seye.

In Renaissance Italy a shift in attitudes is apparent. Giorgio Vasari, for instance, in his widely influential *Lives of the Most Eminent Italian Painters, Sculptors and Architects* . . . (first edition in 1550, enlarged edition in 1568), built up a fairly consistent terminology on the basis of the word *maniera* ("manner"); his *maniera tedesca* ("German manner") refers to Gothic architecture, *buona maniera greca antica* ("good antique Greek manner") to ancient classical architecture, *maniera vecchia* ("old manner") to Byzantine or Byzantine-influenced painting, and *maniera moderna* to Renaissance architecture and painting. The *stilus* of the rhetoricians, *stile* in Italian, was still, however, reserved for literature. Not until around 1600 did musicians use such expressions as *stile moderno* and *stile rappresentativo,* and *stile* in criticism of the visual arts came still later.

In Britain, the equivalent extension of usage does not occur until the 18th century; the *Oxford English Dictionary* gives 1706 for the earliest reference to "style" in painting and 1728 for the earliest application of the term to music. Concerning the earliest English application to architecture there is some disagreement involving connotations, but a case has been made for a passage in Henry Fielding's novel *Tom Jones,* written in 1749: "The Gothic style of building could produce nothing nobler than Mr. Allworthy's house."

Earliest English usage of "style"

After that there was an era of the refinement of labels. A sharpened distinction between the art of ancient Greece and that of Rome began to be made in the 1760s. The division of British medieval architecture into Norman, Early English, Decorated, and Perpendicular styles dates from 1817. The term "roman" (Romanesque), referring to the architecture of western Europe before the Gothic, appeared in French criticism around 1820. Around 1850, "Rococo," after a career in slang, became a serious term for the style of the 18th century. The idea of the Renaissance as a cultural period, and not just an artistic movement, became fully-fledged around 1860. Definitions of the post-Renaissance styles of Mannerism and the Baroque were elaborated in the late 19th century. By the early 20th century, journalists were applying their own coinages to the styles discernible in modern painting and poetry.

Contemporary thought. The interpretation of the centuries-old mass of mostly semantic data on style is difficult and, for many people, exasperating. How is it possible to reconcile seemingly adequate perceptions of style throughout history with the long lack of adequate terms for it? How can the apparently adequate term, *stilus* or a derivative, be reconciled with inadequate perception of what it now connotes? Assuming that Chaucer's pilgrims reached their destination, what did the Host, familiar as he was with the figures and the high style of the rhetoricians, think when he was confronted by the Frenchness of Canterbury Cathedral?

Retrospective interpretations

Some historians have taken the easy course of assuming that when their ancestors used such words as "kind," or "work," or "speech" in certain contexts they somehow actually meant not what such words normally meant but what is now meant by "style." When *stilus* or a derivative was used, according to these historians, it somehow actually meant not what the rhetoricians meant but what is now meant in references to the composer Igor Stravinsky's or the novelist Ernest Hemingway's personal style. More rigorous minds have decided that style, viewed in the perspective of linguistic and general cultural history, is a will-o'-the-wisp. The majority of scholars, however, to judge from published essays, are reluctant to tamper with the evidence or to indulge themselves in a comfortable skepticism. They might therefore agree with the position that will be adopted in this article, which is that style in the arts is to a considerable extent a discovery, and to a large extent an invention, of a surprisingly late date.

Like many another cultural invention-discovery, style had a basis in human behaviour, but the distance between an ordinary ability to recognize things and what is meant by style in the arts is about as great as the distance between ordinary human memory and what is meant by history. Again like many another invention-discovery, style had a long phase during which some important levers and gears were already in place and more or less functioning. But the main job of assembling and powering the apparatus has been done since the end of the Middle Ages. Con-

tributing to this development were Renaissance ideas of the importance of the individual personality, as opposed to medieval collectivism; 18th-century ideas of order and taxonomy in the natural sciences; and 19th-century ideas of biology, history, and, of course, aesthetics, all of which provided analogues for stylistic perceptions. The inventing and discovering are still going on, with much help from the experiments of artists and from such disciplines as archaeology, anthropology, psychology, sociology, and linguistics.

Philosophical questions on the nature of style

To talk of the nature of style in these terms is to raise some difficult philosophical questions, ranging from those posed by ancient thinkers down to those of recent logicians. For although in practical affairs it may be willingly granted that every invention is to some extent also a discovery and that every discovery involves some invention, the fact is that the word "invention" implies one sort of being and the word "discovery" quite another sort. Moreover, by a curious reversal of what happens in many inquiries, the philosophical questions that are thus raised for art critics, historians, and appreciators tend to be less pressing in regard to style as a concept or a collective noun than to style as it is actually experienced. People speculate calmly, if at all, about the nature of their comprehensive stylistic assumptions and become heated about the alleged reality or unreality of the style of the 19th-century French sculptor Rodin or of the *style galant* in 18th-century music or of the style of the Renaissance.

Can anything useful be said in this realm? Certainly any attempt to deal thoroughly with the issues would lead far beyond the scope of this article and deep into problems for which professional philosophers have not yet found accepted solutions. But it seems legitimate to confront antistyle "realists" with the suggestion that a style is no less real, or no more unreal, than a work of art, which is also a kind of invention-discovery, and to add that this degree of certainty is all that a well-tutored man should expect. Works of art can be, among other things, physical objects, imaginary objects, enduring possibilities, realized possibilities, sensible phenomena, insensible phenomena, sheer processes, and even, according to respectable opinion, transcendental entities. They can also have—and this is important to the argument—what has been called an emergent mode of existence and might be called a "do it yourself" mode; a picture, for instance, emerges from the blobs of pigment on a canvas when the viewer steps back and perhaps squints, and a symphony emerges from blobs of sound in the same way. All these modes of existence, and the emergent in particular, can be found among styles. Rodin's style emerges when a sufficient number of his works are contemplated from a certain psychic distance, and it is also a bronze entity in one of his statues of Balzac. The *style galant* emerges from compositions by Bach's sons, and it is also in a single Mozart serenade, which itself exists on paper, in performances, and, above all, as an enduring possibility. The Renaissance style, according to the scholar Arnold Hauser, "is at once more and less than what has actually been expressed in the works of the Renaissance masters. It is something like a musical theme of which only variations are known." In short, it is probably most usefully thought of as having an emergent mode of existence.

All this might be summarized by remarking that the process of invention and discovery that produced the general notion of style over a period of centuries is constantly being recapitulated by individuals, sometimes over a period very brief indeed, for particular styles. From this, one might hastily conclude that practically anyone can invent-discover what will pass unchallenged as a style; and anyone acquainted with modern art scholarship and art publicity must grant that there is a measure of truth in the conclusion. But in the long run, of course, there are certain limits to what can pass as style, just as there are certain limits to what can pass as a work of art; eventually opinion accumulates to the effect that the invention-discovery in question does not work well enough, or is simply not important enough, to qualify for the standard label. The annals of archaeology in Latin America and the eastern Mediterranean are strewn with styles that broke down, often after extensive repairs by their inventors. And, in fact, the world's major accepted styles turn out on examination to have more rigour and clarity in their nature than might be supposed. They have a number of recurring features that can be extracted and combined so as to constitute a working model of style, a sort of metastyle, that can be used for dealing critically with new labels.

PRINCIPAL ASPECTS OF STYLE

These recurring features can be grouped and considered under the headings of value aspects, poietic (or creative) aspects, morphological (or formal) aspects, metaphorical aspects, polar aspects, and measurable aspects. The word "aspect" is preferable to "feature" or "element" or some other possibility, for it must be kept in mind that a style is an invention-discovery with several modes of existence. Moreover, in each instance, what is being talked about is likely to be affected by the different viewpoints of producers, consumers, historians, critics, and other observers.

Value aspects. The first group on the list is logically defective, since all the other aspects have value aspects, but it is important enough to merit some separate preliminary treatment. That styles are regarded as desirable is evident from common usage. Merely to say that an artist, a work, or a period has style is to judge him or it favourably; merely to use "style" in preference to such words as "manner" or "fashion" is to imply value; even to say that a thing is in a poor style is often to suggest that it does not have enough style. And that styles are actually worthwhile is difficult to doubt. They provide the art appreciator with the pleasure of recognizing somebody or something, a pleasure that is certainly among the basic ones of human existence: witness the popularity of handbooks that tell how to distinguish Tang from Sung styles in Chinese art or Louis Quinze from Louis Seize in French art. Styles provide the artist with the pleasure of being recognized, and they do so without the self-display of a signature. They are like codes; in the language of information theory, they help obtain an invariant output from a variable input, and nearly all art history, which is a comparatively recent phenomenon, and much criticism makes use of them. They also have many less evident sorts of value. They function as the signs and, to some extent, as the agents of integration in individual artists, groups of artists, and sometimes whole cultures; they are brakes on alienation. They are appetizing and preservative, like spices; they have saved from oblivion a number of great minds whose ideas have lost their fascination. To cite only a few examples from British critical and historical literature, it is hard to imagine anyone still reading much of Dr. Samuel Johnson, Edward Gibbon, Thomas Babington Macaulay, Thomas Carlyle, or John Ruskin for content alone.

Evidence from common usage

These positive aspects are accompanied, of course, by some negative ones, and the latter have been worrying critics increasingly during recent decades. Since styles stress similarities and work partly as codes, they tend to blur differences and to simplify excessively; they encourage many viewers to see, for example, merely "a Picasso" or "Cubism" where one ought to see the rich uniqueness of a painting. The pleasure of being recognized, and the money that may come with it, can encourage some artists to develop a mere trademark. Successful period styles of the past may foster, as they did among 19th-century European architects, an absurd amount of eclecticism and fancy-dress historicism. Successful current styles can generate an equally absurd amount of imitation among artists. The difficulty of defining specific styles and style in general creates serious problems in art history, which will be discussed in the last section of this article.

Negative aspects

Do these negative aspects outweigh the positive? The majority opinion is clearly that they do not, for there is no prospect of a return to the supposed innocence of the centuries before the massive, intricate, dangerous, useful invention-discovery got up steam.

Creative aspects. One of the complaints, however, is very legitimate: it is that people who talk about style do so too often from the viewpoint of an appreciator. The same complaint can be made about art discussions in general;

a good deal is heard, for instance, about disinterested contemplation, which is fair enough from an appreciator's viewpoint but close to wild calumny from an artist's viewpoint. The hardworking men who built the Parthenon were certainly not disinterested in any usual way, nor were John Milton in writing *Paradise Lost,* Michelangelo in painting the frescoes for the Sistine Chapel, Richard Wagner in composing the opera *Tristan und Isolde,* and Leo Tolstoy in creating *War and Peace,* and what they were doing can scarcely be called contemplation. It was rather what the Greeks called *poiēsis* ("creation," or simply "making").

Style, then, has what can be called poietic aspects; the adjective brushes jargon but lacks the Romantic connotations of "creative" and the matter-of-factness of such an alternative as "productive." The existence of these aspects can be posited etymologically, with the risk inherent in arguing from dead metaphors; *stilus,* as has been noted, originally meant a writing instrument; and such near equivalents for "style" as "manner" (Latin *manuarius,* "of the hand") and "fashion" (Latin *facere,* "to make") also have clearly poietic pedigrees. Moreover, in some contexts "style" still means little more than a mode of artistic production.

But here an attempt to sharpen common usage seems to be called for, since a style in its poietic aspects is not the whole of an act of making. A style is only the part of the act that represents a deviation from a norm and that, as such, is apparent enough to offer the pleasure of recognition. The norm may be provided by a more inclusive style, by a tradition, by material conditions, or by some other frame of reference within which artists work. The norm may also be assumed, more or less arbitrarily, by observers. For example the personal style of the 17th-century Flemish painter Peter Paul Rubens is not, poietically speaking, the whole of his way of applying paint to canvas; his personal style is merely that part of his way that constitutes a recognizable deviation from a norm—the Baroque style—in which he worked. The Baroque, to continue the illustration, is not the whole of the Baroque painters' ways of applying paint to canvas; it refers to merely the part of their ways that constitutes a recognizable deviation from another norm, the general style that prevailed in European painting from roughly the middle of the 15th century to the end of the 19th, though in many histories of art the norm from which the Baroque deviated is instead assumed to be the High Renaissance style as exemplified by Raphael.

The role of norms

One can conclude that style is dependent on originality and the will to exercise it. But much depends also on the norm. In the first place, many norms are imposed, sometimes by accepted authority, sometimes by social pressure, and most often by unawareness of an alternative; the average Western composer, for instance, between Bach in the 18th century and Schoenberg in the 20th, seems to have regarded the tonal-style norm—the organization of tones and chords in a composition in relation to a keynote—as something like a law of nature. In the second place, certain norms contain fewer variables than others and therefore offer fewer opportunities for deviation; a Byzantine mosaicist had no possibility of developing a marked personal style. And, finally, there are the supernorms constituted by each art, by artistic materials, by languages, and by much else; at this level the number of variables may be decisive. Traditional sculpture offers fewer variables than painting and hence has yielded a much smaller number of styles. Granite offers fewer variables than bronze and hence has what might be called a lower yield in terms of style. Specialists in stylistics—the branch of linguistics that studies the variables in a language and their manipulation—have noted that one of the secrets of the 20th-century Welsh poet Dylan Thomas's strongly personal style was his discovery of unsuspected variables in English: thus where the norm had seemed to insist on nouns of temporal, or linear, measurement he could write "All the sun long," "A grief ago," and "farmyards away." In "Spelt from Sibyl's Leaves," the 19th-century English priest Gerard Manley Hopkins showed a comparable talent for finding possibilities of deviation:

> Earnest, earthless, equal, attuneable, vaulty, voluminous . . . stupendous
> Evening strains to be time's vast, womb-of-all, home-of-all, hearse-of-all night.

Formal aspects. Much of the above might have been put under the heading of the morphological—or "formal," if certain connotations are ignored—aspects of style, and much that is traditionally morphological is equally poietic. The form of a work of art can be regarded as the record left behind by the making process; this idea, implicit in ancient rhetoricians' descriptions of prose and poetry (*e.g.,* as laboured), has been prominent in modern criticism since the appearance in the 1950s of such process-emphasizing accomplishments as Action painting, in which the brush strokes and textures may be regarded as a record of the creation of the work; aleatoric music, in which the notes or sounds are selected by chance; and Brutalist architecture, the sort that leaves the concrete raw and plank marked.

The making process, however, involves the whole work, whereas form may be regarded as excluding content and including only shape, volume, space, structure, pattern, organization, texture, rhythm, imagery, emphasis, balance, and the like. This separation is often denounced by careful critics, and a successful work of art, when contemplated in a properly focussed and expanded state of awareness, does indeed present itself with form and content organically tangled. But there is little likelihood that art appreciation suffers when things are untangled for discussion, since most persons are quite capable of distinguishing between the critical analysis of a work and actual aesthetic experience. So it seems safe to accept common usage concerning "form" and then to agree with commentators who have been assuming for centuries that style has purely morphological aspects—without joining them in assuming that it has practically no other aspects.

Separating form and content

Immediately an apparently drastic shift in emphasis occurs, from one of variables and deviations to one of repetitions and conformity. In fact, style can be reasonably, if incompletely, defined as constant formal elements and their combinations, with content excluded except in certain circumstances. But the shift in emphasis is obviously just a result of the play of aspects. A style is always generated by the manipulation of available variables in such a way as to yield a recognizable deviation from a given or an assumed norm. If, however, the deviation is to be recognized as something more than an accident or a solitary impulse, it must be repeated, either identically or in a recognizable variation. It must become understandable, which a unique event—an accident—cannot be.

Thus a style is always, when perceived from what can be called the poietic stance, rather surprising; the Baroque norm and the English-language norm do not lead to any expectation of Rubens' fleshly swirl and Thomas' "A grief ago." But thus also a style is always, when perceived from the morphological stance, rather familiar; what was unexpected in terms of the act of making becomes expected in terms of form. Again the illustration can be continued above the level of personal style: the energized masses found in Baroque painting of the 17th century are constant in their deviation from the balance of High Renaissance paintings of the 16th century. In sum, an artist, or a group of artists, is obliged by the nature of style to move freely into constraint, and heretically into orthodoxy.

Metaphorical aspects. Hence certain styles are commonly said to be "characteristic," or "expressive." An aesthetic deviation, repeated sufficiently, may become converted into a form and yield recognition pleasure. The result may fairly be described as a manifestation of personality. The flame shapes in El Greco's paintings may be thought of as a handwriting, Bach's driving rhythms as a gait, Proust's long sentences as a voice; and usually no harm is done to understanding.

Such metaphors, however, along with the words "characteristic" and "expressive," can become whimsical or misleading in many situations. Moreover, the stylistic aspects in question are best perceived in depth and in the aggregate by noticing that nearly every style is itself a metaphor, functionally speaking. It implies, as Aristotle said a good metaphor does, an intuitive perception of the similarity in

Similarity in dissimilars

dissimilars. It works, as an ordinary simile does, by seizing striking likenesses and neglecting differences; to see that the 18th-century English poet William Cowper wrote in a Miltonic style resembles, as process, seeing that Robert Burns's sweetheart was "like a red red rose." There is a substitution of a part for a whole, as in synecdoche; in many eyes a pointed arch and a flying buttress are enough to evoke the Gothic style, and for many ears a single chord can summon up Beethoven.

Most importantly, in nearly every stylistic context, and not just in those involving personal and "characteristic" styles, there are two sections, like the two sections in the comparing process of metaphor; these can be thought of as the "window" and the "view." In the simplest situations the view through the window of the style is of an individual artist, or of a well-defined group; the 19th-century French poet Stéphane Mallarmé may be sensed through his repetition of the word *azur,* and the contemporaneous group of Impressionist painters that surrounded Claude Monet are recognizable through their deviant, flickering brushwork. The situation, however, is seldom quite that simple. Through Mallarmé's style as well as through Impressionism the view may be of a doctrine or a program; in other situations it may be of a class, a nation, a place, a period. There are styles that bear a functional resemblance to myths, if the latter are thought of as communal metaphors. The view that looms through the symmetrical frontality of the rigidly posed figures in ancient Egyptian sculpture is of an entire culture, a culture that is dramatically different from the one that looms through the asymmetrical twist, the contrapposto, of figures in Italian Renaissance statues.

These remarks have to be qualified, for if common usage is accepted there are styles that seem to have no metaphorical aspects, that are practically opaque. Possible examples are procedural styles in general: those described by the ancient rhetoricians, those associated with the fixed forms of music and poetry, those that are just simplifications, often geometrical, of natural forms.

Polar aspects. Common usage also provides evidence for the existence of certain structural or self-defining tendencies in style that can be grouped under a single heading as polar aspects. These have long been noticed; Hellenistic and Roman literary critics have a lot to say about the flowery, redundant oratorical prose known as Asiatic, which is presented as the diametrical opposite of the plain, economical Attic. The invention or discovery of such polarities did not get seriously under way, however, until the 18th century in western Europe; and most of it has been accomplished, mainly by German thinkers and largely in the visual arts, since the late 1800s.

Among the great number of such contrasting pairs that could be mentioned are haptic–optic (*i.e.,* oriented to the sense of touch as opposed to sight orientation), idealistic–naturalistic, multifarious–unitary, closed–open, linear–painterly, and many more. It will be noticed that the trend is toward all-inclusive world styles, or at least toward constantly recurring stylistic features, and that the emphasis is strongly on morphological aspects. This emphasis, although open to the usual objections to "formalism," has compensated handsomely, in terms of instrumental value, for the sometimes naïve scientism and general overconfidence implicit in the labelling. Indeed, it is not an exaggeration to say that the best visual-art criticism and history published since World War I could not have been written without the help of polar analysis of form.

Applicability of polarization

But pairs like the rather abstruse ones mentioned are not the whole story. Any style, including the familiar established ones, may be polarized; the 19th-century French painter Delacroix's Romantic style may be paired with his older contemporary Ingres' Neoclassical style, the Gothic with the Renaissance, the French with the English, the Christian with the Islāmic, the Eastern with the Western. Each member of each pair is defined in terms of what the other member is not; each is at once the deviant from the other and the norm for the other. Often in such pairs the forms are not seriously analyzed; mere diametrical opposition, as in traditional political parties, is felt to be enough. This peculiarity may be especially evident in connection with a modern style, which is always polar to begin with and which may stay that way until in its turn it begins to cease to be modern. Only then may historians have a good chance to cut through the partisan propaganda, get at the constant forms, and decide if the style is internally consistent enough to merit a label of its own and a place in the parade that began some 40,000 years ago.

Measurable aspects. The historian who undertakes such a task is not likely to approach the constant forms with a yardstick, for in general the measurable aspects of style are not very highly considered by artists and art critics. But such aspects do exist.

The differences between Doric and Ionic orders in classical architecture are not only in the abacuses and volutes that decorate the columns and their capitals but also in proportions and in the number of flutes on a column. The 20th-century French architect Le Corbusier performed his subtle manipulation of architectural variables with the help of a system of proportion, which he called the *modulor,* based on the human figure. Sculpture styles were influenced for centuries by the canon (now lost) of Polyclitus, a Greek sculptor of the 5th century BC who believed that "the beautiful comes about little by little, through many numbers." In a portrait in the Mannerist style of mid-16th-century Italy, much of what makes it Mannerist may be a matter of how long the body is. Some of Hogarth's pictures can be analyzed in terms of what he called "the line of beauty," obtainable by winding a "precise serpentine line around the figure of a cone." Painters of the 20th century have revived the interest Renaissance artists had in the proportion (about 8:13) known as the golden section. An important part of the stylistic difference between a movie director of around 1930 and one of around 1970 may be discovered by simply noting the smaller number of camera shots and sequences used by the latter. Styles in poetry can be specified, often with surprising results, by counts of images, rhymes, run-on lines, and metrical variations; and such methods are accurate enough to help date Shakespeare's plays. Prose styles are, for certain modern linguists, a matter of the statistical averaging of the use of certain words, performed with the help of a computer. That musical styles have measurable aspects has been clear, of course, since at least the time of Pythagoras, who discovered in the 6th century BC the relationship of musical intervals to the lengths of strings; and the fact has become freshly clear under the impact of 20th-century science and technology. Computers have become standard equipment for many composers; synthesizers have become generators of styles translatable into mathematics; Beethoven's poietic deviations from a norm have turned out to be quantifiable in somewhat the same way as the unforeseeables studied by information theorists.

Qualifications on quantification

Some qualifications are in order. The analysis of a style is not the same as the experience of a style: the whole of a work of art is certainly not the sum of its measurable parts. While it may be true that under certain conditions quantity turns into quality, it does not follow that every quality can be quantified; and certain conditions—Beethoven's genius, for instance—remain to plague the quantifier. But all this does not alter the fact that styles do have certain measurable aspects. Nor does it excuse the neglect of these aspects by some art appreciators, critics, and historians. Perhaps the remedy both for the shortcomings of the quantity-minded and for the attitude of the quality-minded will eventually be found in interdisciplinary work on aesthetic problems.

The varieties of style

Since it is part of the nature of style to provide recognition pleasure, and since this pleasure is usually accompanied by an irrepressible impulse to name, one can suppose that unlabelled styles are rare. Would that they were not, an archaeologist may say; would that there were an opportunity to classify works of art scientifically and to substitute numbers or New Latin labels for such misnomers as "Gothic" (which is unrelated to the Goths) and "Cubist" (which has little to do with cubes). Actually, however, the downright mistaken or merely derisive la-

bels are neither very numerous nor very misleading; to anybody who knows enough to be interested, "Gothic" is likely to mean something like "Medieval West European III," and "Cubist" something like "genus, partly abstract; species, Picasso–Braque 1907–14." And if the majority of style names are not scientifically descriptive, they do as a rule offer adequate clues not only to the thing being talked about but also to the class, or classes, to which the thing belongs. In other words, the familiar nomenclature, although accumulated apparently haphazardly through the centuries, has taxonomic—more precisely, typological—implications. To speak of Rembrandt's style is not only to refer to certain poietic deviations from a norm and to certain recurring formal elements; it is also to imply the existence of a personal variety of style plus a more general sort that includes the personal variety. To speak of a realistic style, or of one of its several polar opposites, is to imply a different variety and a correspondingly different general sort.

When such implications are grouped, they yield some 13 varieties of style in the arts. (The "some" is inserted here to allow for reasonable differences of opinion as to where the dividing lines should be drawn.) These 13 varieties are clearly of two general sorts, which emerge from two types of "view" beyond the metaphorical "windows" that styles create in a sufficiently knowledgeable imagination. The first general sort, to which Rembrandt's style belongs, affords views that focus on single cultures; the second sort, to which a realistic style may belong, affords views that cut across cultures. It will be noticed that a given style may move from one category to another; more will be said later about this mobility. But most styles are reasonably stable, and for the moment it is convenient to assume that the others have certain recognizable home categories. Even a slightly unstable classification can stiffen discussion.

Two types of "view"

SINGLE-CULTURE VARIETIES

Single-culture styles are usually inhabited, so to speak. They usually evoke, more or less in the foreground of the contemplating imagination, either a person, a school, a social class, an ethnic division, a regional community, a nation, an ecological division, a religious community, or the generations that constitute a period. Rembrandt's style usually evokes the bulb-nosed, sad-eyed person known through dozens of remarkably self-searching self-portraits; the Venetian style usually evokes the 16th-century masters Giorgione, Titian, Tintoretto, and Veronese; each style in African sculpture usually evokes a tribe; and so on down the list. Single-culture styles are therefore frequently said to be "characteristic," or "expressive," of a particular people; and in this context these familiar terms of interpretative art criticism may seem to triumph over the mild objection raised earlier, in the discussion of the general metaphorical aspects of style. In fact, these adjectives, and also the noun "people," raise some awkward problems even here.

Personal styles. No variety of style seems, at first thought, quite as vividly, specifically, indubitably inhabited as the personal variety. Quoting the celebrated and seldom-read *Discours sur le style* (1753), by the Comte de Buffon, and neglecting his qualifying remarks, many appreciators assume confidently that "the style is the man himself." In a somewhat modified form, the same assumption can be found as far back as the 1st century in the Stoic moralizing of the Roman philosopher Seneca. In a somewhat pseudoscientific form it has produced some disturbingly glib psychoanalysis of works of art. In its plebeian form it has led to such suppositions as that the flame shapes in El Greco's paintings are evidence of astigmatism, the long sentences in Proust's novels evidence of asthma, the rhetoric of Liszt's piano pieces evidence of Gypsy blood, and the right angles of Mies van der Rohe's architecture evidence of a subtle totalitarianism. The notion may be said to have reached one of the peaks of its career in 1935 in the earnest excogitation of the literary scholar Caroline Spurgeon; after counting and sorting Shakespeare's images, she concluded that the poet was

Style and the man

> a compactly well-built man, probably on the slight side, extraordinarily well coordinated, lithe and nimble of body, quick and accurate of eye . . . probably fair-skinned and of a fresh color, which in youth came and went easily . . . very sensitive to dirt and evil smells . . . gentle, kindly, honest, brave and true.

She also saw, through the obviously wide-open window of the personal style, a man who at 35 had "probably experienced heartburn as a result of acidity."

It is easy to call this sort of interpretation wrong and not easy to explain exactly why it is wrong. After all, everyone indulges in it to a degree. Spurgeon, of course, went a bit too far; she was neglectful of complexity and of the possible differences between a dramatist and his personages. But her counting and sorting of images was a valuable and influential piece of research; it demonstrated, in an irrefutable way, the existence of some of the deviations and constants that make up Shakespeare's personal style. If the person who is certainly recognizable in this personal style is not quite the Stratford man himself, who is he?

When the question is asked about a large enough number of personal styles, a tentative answer may emerge. The style, it appears, is not the man himself but the artist himself—mostly, at least. Naturally, the artist is to a considerable extent the man; he has much of the latter's native capacities, acquired skills, secret drives, and painful defects. But the artist is a professional role, a programmatic personage, a cultural configuration, a persona; in sum, he is a remarkably, often deliberately, synthetic personality. Further, he is conditioned by much besides the man himself and notably by the work of other artists. To put all this another way, every personal style is in part sheer performance, and every artist as such is in a sense (not a pejorative one) a performer. Mies the man went on living in his relatively old-fashioned Chicago apartment, while Mies the artist was "performing" with gleaming right angles in the tall apartment buildings he designed for Chicago's fashionable Lake Shore Drive; Mozart the man disliked flute music, while Mozart the artist "performed" by composing flute music; Petrarch himself philandered, while Petrarch the poet "performed" as a faithful worshipper from afar of the idealized Laura.

The person and the performer

Hence, in the opinion of many modern critics, the once-popular problem of personal stylistic sincerity is meaningless; to pose it is to mistake art for life. Hence also the distinction that is often made between creative personal styles and performing personal (or group) styles should not be too categorical. Although a dramatic text, a musical score, or a notated ballet may seem to constitute a norm that offers a very small number of variables, in practice a competent actor, musician, dancer, conductor, or director usually manages to produce enough deviations to have an easily recognizable personal style. No opera-record collector is likely to confuse an interpretation by the intensely dramatic soprano Maria Callas with one of the same role by the serenely lyrical soprano Renata Tebaldi; and ballet literature suggests that the ethereal 19th-century ballerina Marie Taglioni was as different from her rival the sensuous Fanny Elssler as the 19th-century Italian operas of Vincenzo Bellini were from those of his contemporary Gaetano Donizetti. Moreover, an interpreter has about the same choice as a creator among the more inclusive styles that can always be recognized simultaneously with a personal one; he can be modern, traditional, Classical, Romantic, Baroque, or whatever. And finally, if he tries to be as faithful as possible to his text, the result will approach another personal "performing" style, that of the artist who composed the work. To succeed in producing *Phèdre* exactly as it was conceived would be to play the player Jean Racine.

These remarks should not be interpreted as a complete denial of the presence of "the man himself" behind a personal style—into the polar opposite of Spurgeon's error. Since the artist is partly conditioned by the man, so, of course, is the style; and stylistic evidence can often be made to match biographical information in an enlightening way. Friends of Mies noticed in his manners and dress a certain fastidiousness and a love of good material that reminded them of his architecture. The reported simplicity, honesty, and modesty of Haydn are an agreeable match for qualities in his musical style. Most readers probably sense an authentic personality, a real voice, behind the

sprung rhythm and breathless rush of the verse of Hopkins. Also, some quite successful methods of stylistic analysis apparently depend on a strict correspondence between the manner and the man himself; a fascinating example is the technique for attributing paintings, developed by the Italian art critic Giovanni Morelli in the 19th century, which assumes that the touch of a particular master can best be detected in unimportant details, such as the ears in a portrait, that were presumably painted without taking much thought.

None of the counter-evidence, however, seriously shakes the argument that the person in a personal style is mostly the artist as such, and some of it does not stand up very well under scrutiny. Fastidiousness, simplicity, and vehemence do not become meaningful in this context until they are given energy and focus by the artist; and the telltale details used for attributing paintings may have about as much aesthetic interest as fingerprints.

School styles. When artists are considered as a stylistic group, or school, all the problems raised by personal styles reappear in new guises in the company of other problems; and one of the more nettling of the latter is the precise meaning of "school." For there is no denying that this part of the apparatus of criticism has got badly out of hand since its invention and discovery in the 18th century (largely by the pioneer Italian archaeologist Luigi Lanzi). Art critics, historians, and especially painting-museum curators have acquired the habit of using the term as an elegant variation for what usually turns out to be merely a country of residence; thus J.M.W. Turner is said to be a painter of the British school and Thomas Eakins of the American school. Sometimes the geographical designation is narrowed, with a commensurate gain in stylistic information; thus Mantegna is said to belong to the North Italian school and Perugino to the Umbrian school. The gain, however, may be in confusion; the division between the so-called Northern and Southern Sung schools of Chinese painting, for example, has been called the most misleading and arbitrary in art history. Sometimes geography is abandoned for a general stylistic designation; thus the 18th-century French painter Jean-Baptiste Chardin is said to belong to the Realist school. Sometimes the stylistic designation is narrowed drastically, and then a gallery visitor may be confronted by a brass plaque attributing a 15th-century Florentine painting to the school of Fra Angelico, for instance; this can mean that documents point to the studio or to a follower of Fra Angelico; or that the picture is an ancient, and therefore respectable, copy; or that it looks rather like a genuine Fra Angelico without his customary quality—the unstated premise being that all pictures by Fra Angelico are first class.

The meaning of "school"

The situation is regrettable not only because one word is being forced to do things other words can do better but also because "school" has its own work to do. Musicologists can profitably use the term to talk about the centres of musical activity that existed in the 12th century at such places as Paris, Compostela, Padua, and Winchester; or about the late-16th-century amateur "academy" known as the Florentine Camerata, in which the monodic style that led to opera was fostered; or about the group of composers of atonal music that surrounded Arnold Schoenberg in 20th-century Vienna. Painting historians must refer to the Tours school of Carolingian miniaturists, the Shen Chou school of 15th-century Chinese ink artists, the Barbizon school of 19th-century French landscape painters. Literary historians must consider such schools of poetry as those of the 16th-century French Pléiade, the English Lake poets of around 1800, the Tokyo (then Edo) haiku masters of the 17th and 18th centuries, the American Imagists of around 1914. Even architectural historians, who tend to think in large units, have to take account of such phenomena as the Burlington group of Palladianists in 18th-century London, the Glasgow School in the 19th century, the slightly later Chicago School, and so on. Each of these examples, and of the hundreds of others that could be cited, involves a well-defined and usually not large geographical area, a relatively short time span, and a relatively small number of artists working in a describable shared style; here are the essential requirements for using the term "school" profitably. Of course, a curator, in announcing that Turner is of the British school, may really intend to commit his museum to the proposition that a nationally shared painting style has been perceptible in Great Britain down through the centuries. Such is not usually the intention, however, and when it is, it should be made explicit.

The confusion is worth dwelling on because in many works of art a school style, defined as the shared style of a relatively small number of artists, is as striking as a personal style. It is a "window" that affords a "view" inhabited by a synthetic personality, a programmatic personage, almost (but not quite) vivid enough to justify thinking in terms of something like an overartist, as some German philosophers have. Moreover, in the interaction between a personal style and a school style, there is a model, manageable for study, of the complex relations that emerge when several styles are found together in a given work. The personal style of John Donne, the 17th-century English poet of the Metaphysical school, is recognizable in the dense, macabre imagery, the sometimes violently wrenched metre, and the self-dramatizing switch of the following lines from his "Nocturnall upon St. Lucies Day, being the shortest day":

Interaction of artist and school

> The world's whole sap is sunke:
> The generalle balme th' hydroptique earth hath drunke,
> Whither, as to the beds-feet, life is shrunke,
> Dead and enterr'd; yet all these seeme to laugh,
> Compar'd with mee who am their epitaph.

At the same time, the style of the Metaphysical school is apparent in the rather conversational tone, the compact syntax, and the use of extravagant poetic conceits. Deviations and a norm solicit attention, and the solicitations will multiply if the recognition process is continued into the maze of styles—the Elizabethan, the Jacobean, the English, the religious, the aristocratic, the Mannerist, the formal, the haptic, etc.—which a sufficiently subtle and patient critic may discover in Donne's poetry. The common reader, in Dr. Johnson's sense, can be excusably irritated at a certain point by the game of hide-and-seek between the different and the same, the self and the other. But, in fact, such simultaneous recognitions are normal in the actual experience of art; and they are no more mysterious than recognizing, with confidence and usually without being able to say exactly why, that a given face is of an individual, a family, a region, a nation, and a race—is at once itself and not itself.

Social styles. On a scale of recognizability, many critics would probably put social styles—those associated with a particular class or section of society—directly after personal and school styles, at least when much of the art of the centuries before the Industrial Revolution is being considered. Even an untrained appreciator can sense courtly styles in the intricate fixed forms of troubadour verse, the stiff etiquette of a Louis XIV portrait, and the languors of medieval Japan's *Tale of Genji,* by Murasaki Shikibu; bourgeois styles in the solid forms of 17th-century Dutch still-lifes, the uncomplicated rhythm and harmony of a Protestant hymn, and the matter-of-fact, 18th-century English prose of Daniel Defoe; and peasant or proletarian styles—often more accurately described as traditions—in songs, carvings, and embroidery. In the 20th century such clear-cut social styles have become less and less noticeable, partly because class distinctions have become much less evident in the technologically more advanced nations. Also, artists have ceased to know for whom they work, the art market having replaced, except on special occasions that are most frequent for architects, the old system of direct commissioning by patrons. Nevertheless, a sophisticated, or merely mischievous, critic can point to contemporary social styles; some evoke the established families, others the new millionaires. A number are related to young people, who since the 1960s have exhibited many of the economic and cultural characteristics of a separate class. Politicians in all countries throughout the 20th century have occasionally attempted, sometimes on a totalitarian scale, to impose on artists one of the old courtly or bourgeois styles; and a few dictators, notably Hitler and Mussolini, have temporarily revived an imperial-court style with the parvenu touch, familiar to art historians in such

Tradition and class distinction

outsized forms as those of Darius's palace at Persepolis in ancient Persia and Napoleon's church (originally temple) of the Madeleine in Paris.

Here two concessions seem called for. The first is that in talking about social styles it is often impossible, even for the purpose of cold analysis, to keep morphological aspects separate from content. The second is that throughout history the artist as such has been remarkably, sometimes depressingly, available for the "expression" or "characterization" of a social stratum other than that of the man himself. Perhaps the artist as such—shaman, bard, craftsman, entertainer, 19th-century demiurge, 20th-century iconoclast—has been rather more of a performer than has already been suggested.

Ethnic styles. Like social styles, ethnic styles have become less noticeable in the modern era. Their formerly high level of recognizability, their complex morphological aspects, and their surprising persistence over long periods have made them, however, favourite subjects for study among both archaeologists and aesthetics. Good examples are plentiful in the Indian arts of North America; the sculpture, painting, and music of black Africa; the pottery of the ancient Middle East and east Asia; and the surviving decorated objects, mostly metalwork, of the so-called barbarian peoples who moved across Asia and Europe between roughly the 6th century BC and the time of Charlemagne, at the end of the 8th century AD. In some instances scholars have been able to trace the borrowing of motifs; the Germanic animal styles of the migration period, for example, show the influence of Roman figurative art and Mediterranean ribbon ornament. But the borrowed motif is invariably transformed by a repeated deviation into one of the morphological constants of the borrowing tribe, and the reasons for this enduring assimilative capacity are not well understood.

Theories of stylistic disposition

The notion of a physically inherited stylistic disposition has long since been discredited, and archetypes in a collective unconscious, as postulated by the 20th-century Swiss psychologist Carl Jung, have been dismissed by the majority of professional historians. Among the more attractive theories are some that point to analogies with the inertia and the assimilative capacity of a language; yet even these seem inadequate before such a fact as that the Eskimo ethnic style has lasted for about two millennia. Another attractive theory links the lack of stylistic change to a general lack of history—or at least to a general lack of awareness of history. But if this theory is plausible in a North American or an African context, it is much less so in the context of the Asian and European migrating peoples, who managed to pass through an immense amount of history without making important changes in the design of their crowns, buckles, and other useful objects.

Regional and national styles. The problem of stylistic stability reappears in the consideration of regional and lagnational styles, which are partly just ethnic styles that have settled down. But they are always more than that. The high—and polar—recognizability of Oriental and Occidental styles cannot be satisfactorily accounted for by references to ancient tribes and by linguistic analogies; nor can the long preoccupation in the Mediterranean basin with human forms; the long preoccupation in northern Europe with animal, zoomorphic, fantastic, symbolic, and abstract forms; the rigidity of Egyptian pictorial and sculptural forms during some 3,000 years; the persistently emotional, romantic, and expressionist tendencies in German music, painting, and literature; the French emphasis on structure in architecture, painting, and poetry; the English linear and decorative tendency that runs through medieval miniature painting, Perpendicular Gothic architecture, the drawings of William Blake, and Art Nouveau. Common sense, of course, is needed in thinking about such styles; the theorizer who sets out to show that all French art is rational and all German art emotional will be in trouble immediately. Also, a certain vagueness is often suitable, for the distinctive features of a regional or national style may be of the sort better described as qualities than as morphological aspects. Recurring differences, however, finally add up to recognizability, and a quality hard to define can be strongly felt.

Climate and landscape were formerly popular as explanations for regional and national styles (for all styles, as a matter of fact); Gothic architecture was supposed to have emerged in northern Europe because of the many forests. Evocations of some kind of permanent national outlook were also frequent; Russian musical styles were thought to be inhabited by a Slav soul. Such ideas are now perhaps too much out of fashion; it is not quite unthinkable that an English stone carver's affection for linear pattern was encouraged by the frequent absence of bright, shadow-casting sunlight in England and that the symmetry of much French architecture and painting corresponds to the average Frenchman's enduring attachment to order. But eventually, of course, an explanation must include an entire regional or national culture and environment and at the same time take note of the fact that an art may have an existence of its own. To neglect this latter possibility would be to repeat on a large scale Spurgeon's confusion of men with writers.

Ecological styles. A subvariety of the regional variety is perhaps distinct enough to be classed separately as the ecological: here, that is, the style in question evokes in fairly specific ways the relationship of human organisms to their environment. Here also there are likely to be strong polarities; typical contrasting pairs are urban–rural, mountain–plain, hunting–farming, inland–seaboard, nomadic–sedentary, capital–provincial (in some ways), and (at least in comic strips and science fiction) earth–space. Such styles are most recognizable in architecture, painting, and the making of such useful objects as furniture, tools, and weapons. But they may be recognized more distantly in the dance, in music, and in poetry: the imagery of the Parisian poet Charles Baudelaire is urban; that of the New England farmer-poet Robert Frost is rural; while Homer's is seaboard, at least in the *Odyssey*.

Influence on appreciation

It is probable, too, that a well-defined ecological pattern affects in subtle respects the appreciation of foreign styles and hence the emergence of new local ones. Psychologists have found that the Zulus of South Africa, for instance, who live in a "circular culture" of windowless round huts and meandering paths, are relatively immune to visual illusions seen by people who have been conditioned by the right-angled, straight-perspective, so-called carpentered world of European and American cities.

Religious styles. Conditioning also undoubtedly affects recognition of religious styles. The problem of hard-to-define "qualities" that contribute to recognizability, mentioned above in connection with regional and national styles, also returns here to vex a conscientious historian. So does the problem of the separation of form and content, plus the false problem of the sincerity of the artist as such—of the artist as mere "performer." Strictly speaking, in terms of forms and their combinations, one must grant that a vast number of profoundly moving works of religious art are not recognizably in religious styles. One can plausibly argue, for example, that there has been no religious style at all in Western painting since about the 15th century; later painters of religious subjects, such as Van Eyck, Raphael, Rubens, and their successors, painted the Virgin much as they painted their wives and mistresses. A similar point can be made about Western post-Renaissance music; Bach's sacred works sound much like his secular ones, and Giuseppe Verdi's magnificent *Requiem* (1874) has sounded to many ears like his operas. Even in the European Middle Ages and in the worlds of the great Eastern religions, the evidence is not always clear; the style that seems to yearn toward God in the 13th-century Gothic cathedral at Chartres, France, served also for town halls and ivory combs.

Style conditioned by content

The point, of course, should not be exaggerated, partly because other styles often make use of borrowed forms and principally because many religious stylistic elements do exist in their own right, even though their secular ancestry can sometimes be traced. In architecture, examples of religious style can be seen in the basilican plan of Christian churches, the cosmic-mountain form of Hindu temples, the rectangular and cruciform plans of mosques, the needle shape of the minaret; in music, the single vocal line and free rhythm of Gregorian chant, the florid

melody of Islāmic chant; in painting, the free brushwork of Zen Buddhists, the abstract arabesques of Islām, the nonillusionistic kinds of pictorial space favoured by medieval Christians. But when the list of such elements is completed, the fact still remains that the average appreciator recognizes a religious style primarily because it is tinged by long association with religious texts, ritual, and iconography. In brief, it has for him a quality that seems to emanate from content.

Period styles. A style belonging to one of the single-culture varieties may be divided, conventionally or arbitrarily, into periods. Beethoven's personal style is usually split into early, middle, and late; the ethnic style of the migrating Germanic peoples may be referred to as Animal I, II, and III.

The adjective period is often reserved, however, for a distinct variety of style: the one in which the metaphorical "windows" afford "views" inhabited by generations whose cultural and other activities appear to constitute definable units of history. Familiar examples are the Gothic period style, the Renaissance, and, in general, all styles that bear the names of rulers or dynasties: the Victorian style, the Carolingian, the Sung. Frequently the period style is itself "periodized," the Gothic is Early and Late, the Renaissance is Early and High. The issues raised by periodization of all sorts will be discussed in the last section of this article (see below *The dynamics of style*). But period styles have their place in this part the classification, for they are clearly, often emphatically, of the single-culture sort.

CROSS-CULTURAL VARIETIES

Cross-cultural styles are usually uninhabited, in contrast to personal styles. They do not, in any event, call up into the foreground of the appreciator's imagination a John Donne, a Metaphysical school of versifiers, a Murasaki Shikibu in an ancient Japanese court, a Germanic tribe on the march, a succession of English stone carvers expressing their Englishness, a Baudelaire in an urban twilight, a medieval monk singing for God, or a group of eminent Victorians.

Uninhabited, unmetaphorical styles

In a sense, then, cross-cultural styles are relatively unmetaphorical; they tend to focus attention on themselves. They have modes of existence that lie outside of history, and they often are not so much aesthetic terms as general-utility adjectives: almost anything, and not just works of art, can be "classical," for instance, or "organic," or "abstract." Cross-cultural styles in art, however, are by no means without implications; they can evoke outlooks, contexts, methods, and professions.

Outlook styles. Any style, of course, can be said to express an outlook or an attitude along with whatever else it expresses. Michelangelo's personal style expresses the outlook of Michelangelo; the Biedermeier style in furniture and decoration expresses the outlook of the German and Austrian middle classes between the Congress of Vienna in 1814–15 and the Revolution of 1848. Certain cross-cultural styles, however, can be said to specialize in general outlooks and attitudes that recur everywhere in all arts in all eras. There is a Classical style—a deviation toward clear, logical, nobly impersonal, carefully proportioned forms—that is recognizable not only in works that are ordinarily labelled "Classical" or "Neoclassical," such as the sculpture of the Parthenon of 5th-century-BC Athens and the plays of Pierre Corneille of 17th-century France, but also in the 13th-century sculpture of Reims Cathedral, the 20th-century Symbolist poetry of Paul Valéry, the 18th-century English portraits of Sir Joshua Reynolds, a Japanese screen, a Chinese vase, and some buildings on Fifth Avenue in New York City. There is a Romantic style—a deviation toward restless, alogical, nobly personal, intensely expressive forms—that is recognizable not only in works of the Romantic period of the early 19th century, such as the music of Frédéric Chopin or a poem by Heinrich Heine, but also in the Italian baroque architecture of Francesco Borromini, the fantasy painting of Paul Klee, Shakespeare's *King Lear,* a Hindu stone relief, a Hellenistic mosaic, and an avant-garde dancer. The same sort of thing can be said about such styles as the realistic, the fantastic, the expressionistic, and the idealistic, each of which implies an attitude toward life and the world. And perhaps the list should be rounded out with the academic style, into which the other outlook styles tend to sink when they are reduced to a set of teachable rules.

Classical and Romantic styles

Since these styles cross every aesthetic, geographical, or temporal frontier, embrace common human attitudes, and exhibit strong polar aspects, they have tempted many critics into an effort to arrange them all into two contrasting groups: the classical and the romantic, the conservative and the liberal, the idealistic (here the classical joins forces with the romantic) and the realistic, the rational and the emotional, and so on. Such arrangements have the advantage of breaking up traditional divisions and freshening awareness along with the disadvantage of not exhausting the evidence.

Contextual styles. At first glance the outlook styles may seem capable of annexing the contextual, which include the traditional genres and kinds of art—such styles as the tragic, the comic, the satiric, the pastoral, the heroic, and the melodramatic. But here the outlooks, when they exist, are at one or two removes from those of real life and the real world.

A tragic style evokes not a death in the family but a work by Sophocles, Shakespeare, or some other playwright; a pastoral style evokes not herdsmen in the Alps but swains in a mythical Arcadia; a heroic couplet evokes not a brave warrior but the practice of John Dryden and others of using units of two rhyming lines of iambic pentameter in their "heroic" drama. Moreover, when the context—the genre, or kind of art—is not literary, the contextual style may evoke nothing that can properly be called an outlook. The operatic style in Verdi's *Requiem,* cited above, is a contextual style; so is the landscape style, with its hints of land, horizon, and sky, that is sometimes recognizable, like a ghost from Dutch paintings of the 17th century, in modern abstract paintings. So are the sculptural styles recognizable in the bizarre 20th-century architecture of the Spaniard Antonio Gaudí, the painterly styles in the Italian baroque sculpture of Gian Lorenzo Bernini and in certain Cubist sculpture of the 20th century, the musical styles of the poetry of Edgar Allan Poe and Alfred Lord Tennyson.

Procedural styles. Procedural cross-cultural styles are similar to but not quite the same as the contextual. For the term "procedural" is here meant to refer to ways of making art, to varieties of *poiēsis,* that may be shared by several arts and, in fact, are often shared by all. Hence, when the point of view is that of appreciators, procedural styles are commonly called "formalistic." Familiar examples of such styles are the mimetic, the representational, the naturalistic, the abstract, the geometric, the organic, the linear, the optical, the haptical (or tactile), and the plastic; some of these can be regarded, not always profitably, as subvarieties of others, and all of them can be regarded as members of polar pairs.

Formalistic members of polar pairs

Traditionally, certain procedural styles are associated with certain arts, but they need not be. Mimetic, or representational, styles, for example, are most often associated with painting and sculpture (*e.g.,* Gilbert Stuart's lifelike 18th-century portraits of George Washington or a Sumerian animal statue of the 3rd millennium BC), but they can also be recognized in the resemblance of Frank Lloyd Wright's mid-20th-century Guggenheim Museum to a snail, in the similarities of Nikolay Rimsky-Korsakov's musical composition *The Flight of the Bumble Bee* to the buzzing of a bee, and in Tennyson's frequently quoted lines in *The Princess:*

> The moan of doves in immemorial elms,
> And murmuring of innumerable bees . . .

which employ onomatopoeia, the use of words that actually sound like what the words represent.

Similarly, linear styles can be recognized not only in such obvious places as a line engraving by the 16th-century German master Albrecht Dürer or in the linear patterns of Plateresque ornament of 16th-century Spain but also in the vocal line of an Italian Renaissance madrigal or in a modern building by the American architectural firm of Skidmore, Owings, & Merrill.

Traditionally also certain procedural styles are associated with certain cross-cultural outlook styles, but again they

need not be. Mimetic styles are accompanied by Classical in the serene 17th-century landscapes of Nicolas Poussin, by Romantic in the turbulent 19th-century seascapes of Turner, by Realistic in Caravaggio's works of 16th-century Italy; by fantastic in the 20th-century Surrealistic compositions of René Magritte, and by Expressionistic in the tortured figures of Matthias Grünewald's Isenheim Altarpiece of the early 16th century. Abstract styles in 20th-century painting are accompanied by Classical in Piet Mondrian's use of pure lines and rectilinear shapes, by Romantic in Wassily Kandinsky's watercolours of his Blaue Reiter period, by Realistic (of a sort) in Frank Stella's striped, "objective" works, by the fantastic in Yves Tanguy's Surrealistic vistas, and by Expressionistic in Willem de Kooning's often violent canvases. The point is worth some emphasis, for even the most careful critic is likely to slip into the error of assuming that the classical, the romantic, and especially the realistic styles call for eternally fixed methods of production.

Professional styles. When the context of a contextual style or the procedure of a procedural style is derived from somewhere outside the arts in a sufficiently recognizable way or form, the result may be called—at the risk of some confusion—a professional style. Every artist, of course, has a professional style insofar as he is indeed an artist. But the critical occasions for talking about an "art" style in a meaningful fashion are rare; they are likely to arise only when an emphatic distinction is needed between two levels of artistry or when, as during the 19th-century English Aesthetic movement, artists and allied nonartists are waging war against alleged Philistines.

Styles from non-artistic professions

On the other hand, artistic styles derived from nonartistic professions have never been rare and have recently been plentiful. An engineering style is easily recognizable in Roman, Gothic, and modern architecture and also in several kinds of modern sculpture and painting. Many novelists, from the beginning of the genre, have borrowed the styles of historians, and many the styles of journalists; directors of fictional films have worked in documentary styles, composers of modern music in the styles of experimental science and electronic-age technology. Occasionally the stylistic trend among 20th-century artists toward outside professions has been noticeable enough to lead a few observers, always promptly challenged by others, to conclude that the traditional arts were running low on unmanipulated variables.

The dynamics of style

This article has considered styles more or less as a panorama in which all the features are contemporaneous with each other and with the appreciator. The approach, to borrow a term from modern linguistics, has been largely "synchronic"; and such an approach has evident advantages. It gives an explicator a chance to classify, to structuralize, without worrying overmuch about the randomness of history. More importantly, it corresponds in a sense with an appreciator's actual experience of works of art, which is, of course, always synchronic, always in the present; experientially speaking, the Gothic style of the 13th century is not a second older than the style of Skidmore, Owings, & Merrill of the 20th. The historical, or diachronic, approach can offer, however, its own set of advantages. Facts and speculation concerning styles as they occur or change over a period of time can have the same sort of value as political or military history, and in addition they can contribute greatly to the recognition pleasure that is basic in an appreciation of styles.

Difficulties of stylistic history

The pleasure, it should be said right away, is normally accompanied by some peculiarly daunting doubts. Like all history, stylistic history is at once a branch of knowledge that tries to explain significant past events, a chronological record of such events, and somehow the past events themselves, in all their intractable uniqueness. But whereas the members of the trinity can usually be kept reasonably distinct in political or military history, they can usually be found in a state of oneness in stylistic history; the beginning of the Renaissance style, for instance, is notoriously a blend of theoretical explanation, selective records, and such "events" as surviving paintings, poems, and buildings. Also, whereas elections and battles have their fixed times and places, many styles do not; romanticism can leave the ahistorical limbo of cross-cultural styles to become the single-culture 19th-century Romanticism and then suddenly materialize, to some historians, as a phase of late Greco-Roman single-culture Classicism. In sum, the emergent mode of existence to which styles are disposed may make impossible a clean distinction between conceptions and historical facts and may expose the most conscientious of art historians to the charge of merely manoeuvring with private abstractions.

Another difficulty, which was glanced at in connection with the measurable in style, can be mentioned again here. It is the familiar one of overspecialized interpretations presented as general theory. Much of the terminology and nearly all of the chronological framework for the history of style are inventions of specialists in the study of the visual arts. Fortunately, the terminology and the framework can be adapted, with the help of some stretched meanings, to the study of music and literature. In what follows, a strong possibility of visual-art bias should be kept in mind during the discussion, which will concern origins, diffusion, and change and duration of art styles.

HISTORICAL ORIGINS

The question of the origins of styles has already been raised. From an appreciator's strictly synchronic point of view, to raise such a question is often to be irrelevant, or to get into a round of tautologies and to confuse the merely relational with the causal. The origins of the courtly style, of the Eskimo style, and of personal styles are by definition courts, Eskimo, and artists, respectively.

The relational and the causal

There is something to be said, however, for the elaborated tautology as a form of analysis and also for the fact that adding the space-time of history makes a difference: it is one thing to point to an emphasis on order in the French national style and another thing to account for the advent of the Gothic style at the abbey of Saint-Denis in 1140. Moreover, art historians do not usually mean anything as determinative as a cause when they refer to the origins of a style; they merely mean a set of conditioning historical factors. And when they disagree they usually do so over the relative importance to be assigned to one of four sorts of factors: the politico-economic, the cultural, the technical, and the artistic.

Politico-economic factors. The most evident sort is the politico-economic. If the ancient Greeks had lost the Persian Wars in the 5th century BC, the columns of the temples on the Acropolis would probably be taller. If the Great Depression in the United States had begun earlier, the Empire State Building and other skyscrapers in New York of the 1920s and 1930s would probably have been shorter. The styles of Chinese blue and white porcelain were perceptibly influenced by export possibilities. Films are an industry as well as an art. The Italian Renaissance cannot be fully understood without references to the personality cults of monarchs, petty tyrants, and usurping mercenary leaders, to the capital accumulated in the bank of the Medici family of Florence and the coffers of the Vatican, to the expansion of international trade, and to the loosening of feudal ties. Such examples, along with more subtle ones, can be multiplied into a solidly factual kind of stylistic history.

Cultural factors. Nothing quite so solid is likely to come out of a consideration of contributing cultural factors; here one is obliged to weigh such imponderables as changes in manners, morals, psychologies, philosophies, and that pervasive sensibility system known as "the spirit of the age" and then to try to match these changes with the arrival times and the formal elements of presumably new styles. Thus it can be argued that the new period style gradually apparent at the close of the Greco-Roman era—the lack of naturalism, for instance, in painting—reflects Christian otherworldliness.

The question of the "spirit of the age"

Similarly, it might be claimed that the exuberantly enumerative prose style of the French satirist François Rabelais has roots in the widespread 16th-century zest for learning and discovery. To take still another example, the

development of the orchestral crescendo in the middle of the 18th century, principally at Mannheim, Germany, has been said to reflect a new kind of human self-assertion, one destined for importance in all the arts with the arrival of 19th-century Romanticism.

Technical factors. The presence of politico-economic or of cultural factors need not, of course, rule out the technical sort, and examples of the latter are plentiful. The stylistic change from 14th-century to 15th-century European painting coincides with a new utilization, although not the invention, of the oil technique. The refinement of volumes and of implied movement in Renaissance bronze sculpture is linked to a new skill in casting techniques acquired in the making of artillery. Chopin's personal style and significant improvements in the piano both date from the 1830s.

Sometimes the technical factors are partly disguised by a change of material; the Doric order in architecture is a wood style in stone. Sometimes, too, the technical factors must be understood in a large sense that embraces the functional, and they may then be accompanied by moralizing; thus, some 20th-century architects have maintained that it is "honesty" to let a building reveal its construction and use. Usually more than one set of technical factors can be discovered; a poet's style that mixes true rhyme with eye rhyme (similarity only of spellings) has origins—rather remote—both in ancient oral mnemonics and in the silent reading fostered by printing.

Artistic factors. Explanations of stylistic origins involving politico-economic, cultural, and technical factors tend to have in common the defect of not providing very exclusive matchings of formal elements with supposed sources. Many times and places have had ambitious princes and accumulated capital equal to those of 16th-century Italy without producing anything like the smoky style, the sfumato, of Leonardo da Vinci's paintings, or the clarity and solemnity of Bramante's architecture. If a lack of naturalism goes well with early Christian otherworldliness, it also seems to go somehow with the reputed this-worldliness of the 20th century. If the perfected piano can be matched with the smooth 19th-century compositions of Chopin, it can also be matched with the staccato 20th-century music of Béla Bartók. To be sure, these are only contributing factors, not specific causes. Even so, the versatility of the factors mentioned usually leads an art historian to add some relatively direct, purely artistic factors and eventually to postulate a group of ancestors and influential cousins for a given style. Thus Leonardo's style is said to be derived partly from his early master Verrocchio's, early Christian from ancient Roman and Syrian, Chopin's from the styles of other Polish pianists and from the Irish-born John Field. Art historians of this persuasion do not, as is sometimes alleged, reject the hypothesis that an artist may think up a style partly on his own; they merely regard it as lying outside the concerns of scholarship.

Matching causes and effects

Here another difficulty in the writing of stylistic history has to be mentioned. In which painting did Leonardo's personal style become his own and not Verrocchio's? At what date did the early Christian style cease to be Roman and Syrian? Questions of this sort, if pressed far enough, can raise the suspicion that a discussion supposedly about the origins of a certain style is actually about the arbitrary periodization of a more inclusive style.

DIFFUSION

Periodization is always, although the fact is not always made clear by periodizers, a spatiotemporal operation: it poses questions of where as well as of when. The Gothic period can be said, if dates are used with convenient recklessness, to end around 1420 in Italy, around 1530 in France, and a generation later in England. Moreover, the "where" may be not only a point in geographical space but also a zone in the imaginary space constituted by any classification of the arts.

The factor of place in diffusion

The Italian Renaissance can be said to begin in painting around 1300 with Giotto; in poetry around 1330 with Petrarch; in architecture around 1420 with Brunelleschi; and in music around 1525 with madrigal composers, or perhaps around 1600 with Monteverdi. In sum, staggered dates, distinct localities, and the separation of the arts combine to involve the student of period styles in a complex problem of assumed or alleged correspondences and sometimes in unexpected variants of the problem of stylistic origins.

Correspondences in real space. Like their colleagues in other branches of cultural history, art historians sometimes are obliged to choose between diffusionism and theories of spontaneous development; the striking similarity between a motif on a Chinese bronze and one on an Aztec temple, for instance, may be attributed to prehistoric migration, to analogous stages of civilization, or to chance. Normally, however, and for good reasons, art historians rely on diffusion from a creative centre in their attempts to explain stylistic correspondences in even widely separated areas; and they rely on it almost exclusively in discussing period styles.

It is fairly easy to trace the movement of Andrea Palladio's style from the buildings he designed in Vicenza, Italy, in the 16th century to Inigo Jones's Banqueting House in London in the 17th century, of opera buffa from Naples to Vienna in the 18th century, or of Cubism from Paris to Prague and New York in the 20th century. A difficulty in this sort of history is that diffusion implies receptivity, and stylistic receptivity implies at least a degree of spontaneous development.

Correspondences in the arts. There are a number of problems in attempts to deal with stylistic correspondences in different arts. One of the most common is the difficulty of making a distinction between a period style and a period of time when the same label is used for both—and perhaps for each in more than one sense. Thus, a historian who refers to Baroque music without explanation may mean that the music in question has certain stylistic affinities with the theatricality of Bernini's sculpture, with the full-blooded movement and illusionism of Pietro da Cortona's painting, with the undulations of Borromini's architecture, and with the cadences of Sir Thomas Browne's prose; instead, or also, he may mean, as many musicologists would, that the music carries the technical eramark of a basso continuo part; again he may mean that the music evokes the manners, public life, science, and general culture of "the Age of the Baroque"; and finally he may mean nothing more than that the piece was composed in the 17th century. If he does mean that the music has stylistic affinities with Baroque sculpture, painting, architecture, and prose, then other difficulties appear. He must manage to translate each art into the others, construct a convincing historical pattern, and explain the correspondences.

Problems and ambiguities of correspondences

Nevertheless, there is considerable agreement among art historians that such correspondences do exist—not in every period but often enough to merit attention. In addition to those of the Baroque, which are frequently cited, one can mention those between musical and visual-art proportions during the Renaissance, those among nearly all the arts during the Gothic period, and those among painting, sculpture, architecture, and music during the 20th century. Persuasive arguments for correspondences between art and other disciplines have been advanced, notably in a comparison of Gothic architecture with scholasticism, the contemporaneous church-oriented philosophy, and by several commentators in comparisons of Leibniz's highly complex but integrated philosophy to Baroque art.

CHANGE AND DURATION

The already mentioned problem of distinguishing between the origins of a distinct style and the periodization of a more inclusive one complicates every discussion of stylistic change and duration. It can be argued that whenever a style moves into a new period or phase it becomes a new style; and a few modern historians have drawn the conclusion that the concept of style should be kept out of art history and be reserved for synchronic structuring and critical analysis. They feel that one can usefully take a cross section of Baroque, for example, and study it horizontally, out of time, but that to talk about the historical antecedents of Baroque, or about phases of Baroque, is to fall into logical contradiction. Such historians, however,

are very much a minority. Most of their colleagues are committed not only to stylistic change and duration but also, if often only implicitly, to various theories about supposedly typical stages in the stylistic historical process.

Cyclical theories. In the preface to his *Lives,* Vasari remarked that the arts, like human beings, "are born, grow up, become old, and die." Four centuries later, the infinitely more knowledgeable and sophisticated art historian Henri Focillon entitled an essay on style *The Life of Forms in Art* (1934). André Malraux, in his *Voices of Silence* (1951), refers to

> these imaginary super-artists we call styles, each of which has an obscure birth, an adventurous life, including both triumphs and surrenders to the lure of the gaudy or the meretricious, a death-agony and a resurrection.

References to the "maturity" or the "decay" of styles are part of nearly everybody's art criticism. In sum, the biological metaphor for style has become commonplace. But it once was a serious theory of the historical process, and as such it perhaps still deserves some denunciation for having been frequently both a misrepresentation of facts and a source of prejudice against estimable works of art—Hellenistic, late Gothic, Mannerist, Baroque, and so on—which were believed to belong to the "decadent" part of "the life cycle" of a style.

In an effort to get closer to the observed course of events, some historians have favoured, in whole or in part, a cyclical theory that pictures each major single-culture style as going through the same irreversible series of cross-cultural styles, which is usually given as archaic, classic, baroque, impressionist, and archaistic. Thus the first, or archaic, phase of the Greco-Roman cycle is supposed to correspond with the first phase of the Gothic and of the Renaissance cycles; and one can speak of third-phase baroque Greek, third-phase baroque Gothic, even third-phase baroque Baroque. The theory has some fascination, and occasionally it works well enough to shed light on the actual historical behaviour of styles. But often it works only because one has been careful to start the cycle in the right place, and far too often it does not work at all.

Dialectical theories. The notion that history is a pendulum, with left inevitably followed by right, occurs to everyone, and it is especially likely to occur to art historians because of the polar aspects of style. Sometimes the pendulum is imagined as swinging forever between cross-cultural outlook styles: classical then romantic, idealistic then realistic. Historians who are inclined to feel that outlook styles are too vague and too laden with value implications may favour the idea of swings between cross-cultural procedural styles: mimetic then abstract, organic then geometric.

The pendulum notion of history

The influential theory devised by the Swiss art historian Heinrich Wölfflin (1864–1945) postulates an elaborate dialectic involving pairs such as linear–painterly and closed–open. A theory devised by the German critic Paul Frankl (1878–1962) combines a movement between styles of being and becoming with a cyclical development through preclassic, classic, and postclassic stages. Another theory, devised by the Austrian scholar Alois Riegl (1858–1905), postulates cycles of evolution within a long swing across the centuries from an early haptic to a later optic phase. All these models of the stylistic historical process break down eventually, but they can provide useful categories for organizing the flux of artistic events.

Sequential theories. Historians who refuse to believe that styles are haunted by destiny may be willing to grant that within short periods and mostly in terms of technique something that looks like a preordained evolution can occur. Thus it is possible to discern in Gothic architecture a predictable movement toward lighter construction, in Early Renaissance painting a movement toward a desired illusionism, in 16th-century European music a movement toward the tonal system that emerged in the following century. But beyond this concession, such historians are likely to rely on simple sequences devoid of any trace of determinism. Their works often consist of a sequence of biographies of artists and a sequence of centuries.

Do such theoretical methods suggest that art historians are engaged in mere chronicling rather than history? Perhaps partly to answer such a question, a theory advanced by the contemporary U.S. historian James S. Ackerman presents stylistic change as a "process" of a sort, although not one that is preordained; art history is envisaged as a series of steps away from the past, but not toward the future. "Each step," Ackerman says, "for the artist who takes it, is final and definitive; he cannot consciously make a transition to a succeeding step, for if he visualizes something he regards as preferable to what he is doing, he presumably will proceed to do it. . . . "

Satiation theories. Why does the artist take a step away from the past? A common explanation is that he, or his audience, has become overly familiar with the prevailing style. The phenomenon is similar to what psychologists call semantic satiation—the loss of meaningfulness in repeated words. According to one theory, stylistic satiation is hastened by comparison; thus when much was changing at the beginning of the 20th century, many painters felt that their old styles were intolerably familiar. According to another theory, satiation is always merely temporary. To revert to part of the biological metaphor, styles are born, but they never really die.

BIBLIOGRAPHY

Theoretical surveys: An excellent critical résumé of early 20th-century speculation, with emphasis on the visual arts, is M. SCHAPIRO, "Style," in A.L. KROEBER (ed.), *Anthropology Today* (1953). J.S. ACKERMAN, "Theory of Style," *Journal of Aesthetics and Art Criticism,* 20:227–237 (1962), stresses relational values and offers a lucid critique of evolutionary notions. R.L. SCRANTON, *Aesthetic Aspects of Ancient Art* (1964), analyzes in sharp detail the structure of style and applies his conclusions to specific works, including literature. HERBERT READ, in the introduction to *The Styles of European Art* (1965), emphasizes the psychological aspects of the subject. RENE WELLEK and AUSTIN WARREN, *Theory of Literature,* 3rd rev. ed. (1966), have much to say about style in passing, from the standpoint of Anglo-American New Criticism. GEORGE KUBLER, *The Shape of Time* (1962), wages brilliant war on conventional art history and presents a formal substitute for the concept of style. BRUCE ALLSOPP, *Style in the Visual Arts* (1956), reviews the history of the concept and stresses the allegedly damaging effects of style on contemporary artistic activity.

Special studies: ERWIN PANOFSKY, *Renaissance and Renascences in Western Art,* 2nd ed., 2 vol. (1965), examines the idea of rebirth in Italy and discusses the significance of medieval renascences; his *Meaning in the Visual Arts* (1955), although focussed on iconography, has extensive reflections on style; his *Gothic Architecture and Scholasticism* (1951) is a study in correspondence. ETIENNE SOURIAU, *La Correspondance des Arts* (1947), presents a basis for comparative aesthetics. WYLIE SYPHER, *Four Stages of Renaissance Style* (1955) and *Rococo to Cubism in Art and Literature* (1960), make good use of stylistic polarities in an analysis of painting, sculpture, architecture, and literature. RUDOLF WITTKOWER, *Architectural Principles in the Age of Humanism,* 3rd ed. rev. (1962), deals in depth with problems of symbolism, optics, and harmony in Renaissance buildings. NIKOLAUS PEVSNER, *The Englishness of English Art* (1956), offers one of the few instances of a serious historian taking the idea of a national style seriously. G.M.A. GRUBE, *The Greek and Roman Critics* (1965), discusses stylistic theories from Homer to the 3rd century AD. WALTER WIORA, *Die vier Weltalter der Musik* (1961; Eng. trans., *The Four Ages of Music,* 1965), presents a panorama of the world's musical styles from ancient times to the 20th century.

Classical studies: FRANZ BOAS, *Primitive Art* (1927, new ed., 1962); H. FOCILLON, *Vie des formes* (1934; Eng. trans., *The Life of Forms in Art,* 2nd ed., 1948); P. FRANKL, *Das System der Kunstwissenschaft* (1938); A. RIEGL, *Stilfragen: Grundlegungen zu einer Geschichte der Ornamentik* (1893; 2nd ed., 1923); HEINRICH WOLFFLIN, *Renaissance und Barock* (1888; Eng. trans., *Renaissance and Baroque,* 1964); *Die klassische Kunst* (1899; Eng. trans., *Classic Art,* 1952, reprinted 1968); and *Kunstgeschichtliche Grundbegriffe* (1915; Eng. trans., *Principles of Art History,* 1932).

Bibliographies: Useful short lists of theoretical works on style may be found in the above-mentioned publications of Schapiro and Scranton. Longer lists, mixed with other material, may be found in those of Panofsky, Sypher, and Wellek and Warren.

(R.McMu.)

Aschelminths

The phylum Aschelminthes (or Nemathelminthes) includes five diverse classes of wormlike animals, mostly of microscopic size: Nematoda (or Nemata), Rotifera, Gastrotricha, Kinorhyncha (or Echinodera), and Nematomorpha. The American zoologist Libbie H. Hyman, in her classic textbooks on the invertebrates, originally included Priapulida as a class of the aschelminths, but Priapulida are usually not now included. Aschelminths have in common a body cavity, the pseudocoel, that arises in the embryo in a way different from that found in more advanced animals and that has no epithelial lining—*i.e.*, it is not a true coelom. Priapulids possess such an epithelial lining and are therefore coelomates. Aschelminths are bilaterally symmetrical, have a tough external covering, the cuticle, and, except for the kinorhynchs, lack segmentation.

For coverage of related topics in the *Macropædia* and *Micropædia,* see the *Propædia,* section 313.

This article is divided into the following sections:

GENERAL FEATURES

Size range and diversity of structure. The five classes of aschelminths are of different sizes and varying importance. The nematodes are by far the largest group, with 13,000 to 14,000 named species and many times that number undescribed. Most of the described species are parasites of human beings, domestic animals, or cultivated plants and are therefore of great importance in medicine and agriculture. Typically, nematodes have a simple wormlike body, elongated, without appendages or segmentation, which moves with a characteristic sinuous movement, though there are exceptions. Parasitic nematodes may be large enough to be seen with the naked eye, with a few 50 centimetres (20 inches) or longer and are often referred to as roundworms. Most are not parasites and are microscopic, between 0.1 and two millimetres (.004 and .08 inch) when fully grown, living in soil or aquatic muds or sands.

The rotifers, with about 2,000 known species, are common microscopic animals in lakes and ponds but also occur in the sea and damp soil. Generally between 0.1 and 0.5 millimetre long, they are usually recognizable under the microscope by the water currents set up by rows of beating hairlike cilia, which are used to collect food, on their heads (the corona).

The gastrotrichs, with at least 450 named species, and of a size range similar to rotifers, creep or swim by cilia but do not possess a corona. They are found in fresh waters and marine mud sand or among plants. The kinorhynchs, with about 150 described species, and less than one millimetre long, are little known but not uncommon in marine sands and muds. The body is segmented and spiny, with a retractable head. The nematomorphs, with several hundred species, are, when juvenile, parasites of insects, spiders, centipedes, or marine shrimps. The adults, which may be as long as 0.5 to one metre, were formerly called horsehair worms because it was believed that they arose from horses' hair that had fallen into the water. Because they can become tangled in knots, they are also sometimes called gordian worms. Nematomorphs are long, thin worms with a brown, leathery body. They swim or crawl with a sinuous movement superficially resembling that of nematodes.

Horsehair worms

Distribution and abundance. Nematodes are the most abundant of all multicellular animals and are found wherever life can be supported. They reach their greatest numbers in estuarine mud flats, up to 20,000,000 per square metre, decreasing in density in marine muds and sands but often reaching one to several million per square metre. Typical densities for terrestrial soils, including forests, grasslands, agricultural land, and even arctic tundra, number several million per square metre. Though basically aquatic, as are all aschelminths, nematodes can be found in deserts and polar regions because of the ability of some to survive drying or freezing conditions in an inactive state (cryptobiosis [see below *Adaptations*]), a capacity shared with some rotifers. Unlike many rotifers, however, nematodes are not adapted to a free-swimming planktonic life.

Rotifers are common in the surface waters of lakes and ponds, often showing short-lived seasonal blooms of several thousand per litre. Others attach themselves to aquatic plants (sessile rotifers) at even greater densities (tens of thousands per litre). Some inhabit the interstices of aquatic sediments; at one lake more than 1,000,000 per litre were found. They also may be found in the soil and in the sea.

Importance. The aschelminths are primarily particle feeders, grazing on bacteria, microfungi, algae, and protozoans, though some are predators on other aschelminths. As such they play important roles in recycling nutrients, which then become available for plant growth, and in promoting the decomposition of dead organic matter. Aschelminths are at the base of many food chains.

The nematodes, however, have a much more direct impact on human welfare as parasites. In agriculture they cause great losses in the production of cereals, root crops, and many other plants. Crop rotations, the application of potentially toxic nematocides, and the development of resistant cultivars are required to control their activities. One nematode, *Bursaphelenchus xylophilus,* in association with beetles, even destroys pine forests. Nematodes can be used in the biological control of some insect pests.

Effects on agriculture

Nematodes cause several of the most serious tropical diseases, such as, for example, filariasis, an infestation of the lymphatics and subcutaneous and deep tissues that causes inflammation and scarring. Filariasis is transmitted by biting insects.

Filariasis

Many nematode parasites live in the alimentary canal and are spread as infective eggs. *Ascaris lumbricoides,* a human intestinal parasite, is common wherever human sewage is used as a fertilizer because its eggs contaminate food. Even in the most hygienic countries, most people suffer at some time from the human pinworm *Enterobius vermicularis,* an innocuous inhabitant of the bowel whose eggs are passed from person to person. Some species for which a human is not the normal host can invade humans and cause diseases. For example, the dog roundworm, *Toxocara canis,* can only mature in the dog's intestine, but the microscopic larvae from dog feces can invade human tissues and occasionally can cause blindness.

NATURAL HISTORY

Reproduction and development. Most aschelminths are bisexual; the male inseminates the female during copu-

lation so that she lays fertilized shelled eggs. In some cases, these eggs may have started to develop before being laid. Parthenogenesis, by which the female's eggs develop without fertilization, is common. In most rotifers (*i.e.,* the order Monogononta) males are smaller and less frequent than females. Monogonot females are of two kinds: amictic or mictic. Amictic females produce amictic eggs, which develop without being fertilized (parthenogenesis). Mictic females lay mictic eggs, which develop into males if unfertilized or into amictic females after a period of dormancy if fertilized. Of the remaining orders of rotifers, males are unknown in the order Bdelloidea, and the order Seisonidea is bisexual.

Amictic and mictic females

In some terrestrial nematodes, individuals of female appearance first produce spermatozoa, which are stored and then used to fertilize eggs produced by the same gonad; such individuals are called protandrous hermaphrodites. Normal functional males may occur much less frequently in the same population. Parthenogenesis also occurs in nematodes, sometimes with environmental factors determining whether the female's eggs will develop parthenogenetically or require fertilization by a male. Some gastrotrichs are parthenogenetic; others are protandrous hermaphrodites. Nematomorphs and kinorhynchs are bisexual.

Generally, aschelminth eggs have relatively little yolk, and their embryonic development may be technically described as holoblastic cleavage (*i.e.,* division of the entire egg into separate though contiguous cells), during which a blastula (a hollow, single-layered ball of cells) is formed, followed by the formation of a gastrula (a hollow, two-layered ball of cells). Gastrulation brings the cells destined to form the adult organs, derived from mesodermal and endodermal layers, and the reproductive cells into the interior. The earliest stages of embryonic development in rotifers show a limited form of spiral cleavage, also found among the flatworms and segmented worms, but such a pattern is much less apparent in other aschelminths.

Determinate development

Development is highly determinate, giving rise to a body with a relatively constant number of cells in highly characteristic positions, little, if any, powers of regeneration of lost parts, and lack of capacity for asexual reproduction. The cuticle, which covers the body, is molted as the body grows in nematodes, nematomorphs, and kinorhynchs. In nematodes there are four molts separating four juvenile (or larval) stages preceding sexual maturity. Sometimes there may be substantial growth after the last molt. In kinorhynchs a larva of three segments adds more segments as it grows.

The terrestrial bacteria-feeding nematode *Caenorhabditis elegans* is of interest to developmental biologists because unique features have made it invaluable for the analysis of the genetics of development and for other aspects. A generation may take as little as 3½ days to develop, and each hermaphrodite lays between 200 and 300 eggs. The sequence of cell divisions and differentiations that transforms the fertilized egg into an adult worm is known in precise detail. A highly predictable sequence of divisions gives about 550 cells by the time the juvenile worm hatches, with all of its organs formed apart from the sexual organs. The place and function of each cell, muscle, or nerve, for example, are determined largely by its place in the invariant sequence of binary cell divisions (*i.e.,* its cell lineage) as well as by chemical messages from other cells. There are 811 cells in the hermaphrodite and 971 in the male, in addition to the eggs and sperm. The juvenile (or larva) sheds and replaces its cuticle four times before becoming a sexually mature adult.

Experimental studies

Hundreds of experimentally induced mutated individuals have been bred as self-fertilizing genetically identical clones, which are invaluable for the study of the genetics (inheritance) of developmental processes. The clones can be stored indefinitely at very low temperatures until required and have been used to study a wide range of biologic processes. Self-fertilization by hermaphrodites facilitates isolating mutated individuals, while crossing females with males makes it possible to map the genes on the five pairs of non-sex chromosomes and on the one pair of sex chromosomes in the female or on the unpaired chromosomes in the male. The DNA of *C. elegans* has some 80,000,000 base pairs, about 4,000 genes. Genetically speaking, it is a very simple animal.

Locomotion. Nematodes have a characteristic sinuous movement in which waves travel along the body, which generally lies on its side, backward waves driving the body forward, forward waves driving it backward. These waves are brought about by successive contractions of longitudinal body wall muscles, in dorsal and ventral blocks, acting out of phase. The muscles increase the hydrostatic pressure in the internal tissues, causing the flexible, but not very extensible, cuticle to bend and producing graceful body curves. The body waves enable nematodes to move efficiently through the fluid-filled interstices of mud, sand, and soil or to crawl in thin films of water, using the resistance offered by surface tension. Nematodes also swim with body waves but not so efficiently, some of their effort being dissipated as turbulence. Nematomorphs move in a way similar to nematodes.

Rotifers and gastrotrichs swim by means of beating cilia (protoplasmic hairs), which in rotifers also generate food-collecting currents. Many rotifers swim continuously close to the water surface, while others loop along over surfaces, alternately attaching the corona of the head and the toes of the tip of the foot to the surface. Some rotifers are sessile, remaining attached to an object, sometimes building a tube by gluing together particles with body secretions so that only the head projects. Gastrotrichs glide by using cilia to propel themselves over surfaces. The kinorhynchs use the repeated eversion and retraction of the spiny front end of the body, coupled with muscle contraction, to pull themselves forward.

Behaviour. Although their nervous systems are simple, aschelminths can respond to a range of environmental stimuli by, for example, attraction, avoidance, feeding, or copulation. They respond to touch, temperature gradients, chemicals, and light. Some species in each class have two or more pairs of simple eyes capable of indicating the direction and intensity of light but unable to form an image.

Many nematode males are attracted to females by chemical attractants (pheromones) released by the female. Plant-feeding nematodes are attracted to the roots of host plants. *C. elegans,* a soil inhabitant, is attracted to a number of dissolved salts (sodium, potassium, magnesium, and chlorine), some amino acids, cyclic nucleic acids, basic solutions (OH^-), pyridine, and the products of its bacterial food. Other chemicals are repellent. Nematodes respond with changes in the pattern of body waves and in the frequency with which their direction of movement is altered.

Adaptations. The ability of some nematodes and rotifers to survive drying or freezing conditions in a state of suspended animation, *i.e.,* cryptobiosis or anabiosis, has enabled them to inhabit the driest deserts and the coldest polar regions, as long as free water occurs occasionally and for long enough periods of time for them to reproduce. This ability also allows both groups to inhabit large areas of the world where the soil is seasonally arid or frozen. The plant-parasitic nematode *Ditylenchus dipsaci* has been revived after 23 years in dried plant material. The animal responds to negative environmental signs by changing the hydration of its proteins and cell membranes, synthesizing antifreezes, such as trehalose or glycerol, and storing energy. It is possible to store *C. elegans,* which does not naturally tolerate freezing, in a liquid-nitrogen refrigerator at very low temperatures and later to revive it after artificially infusing glycerol.

Cryptobiosis

Plant- and animal-parasitic nematodes often have infective stages that remain quiescent for long periods. *A. lumbricoides* eggs may infect a host after remaining for several years in the soil. The eggs of the cereal cyst nematode *Heterodera avenae* may have to experience winter chill before they will hatch in the spring. After long periods of inactivity, the eggs of the potato cyst nematode *Globodera rostochiensis* hatch when stimulated by substances diffusing from the roots of potatoes.

Nematodes from two different orders, Tylenchida and Dorylaimida, have mouthparts forming a hollow needle, made of cuticle, that acts like a hypodermic syringe (Figure 1). Digestive juices from pharyngeal glands can be pumped out through the stylet and plant and animal juices

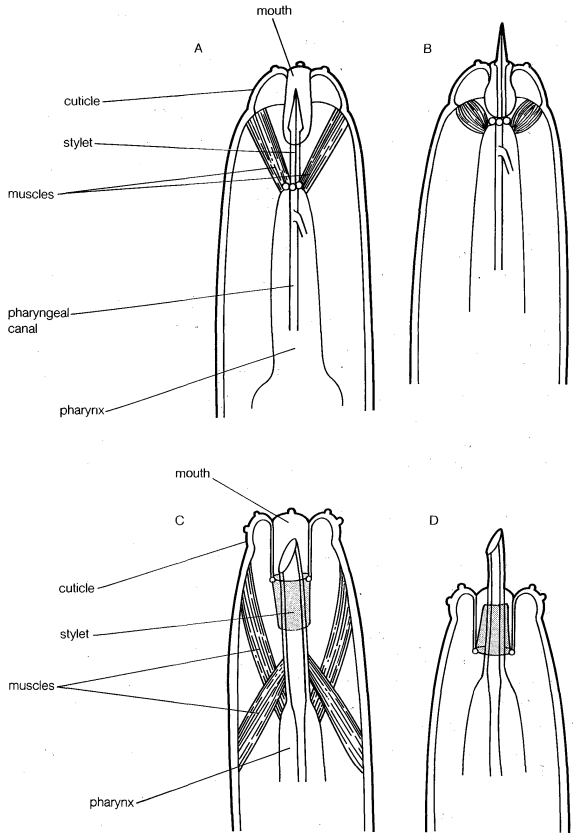


Figure 1: *The heads of stylet-feeding nematodes.* Tylenchid with (A) stylet withdrawn and (B) stylet protruded as in feeding. Dorylaimid with (C) stylet withdrawn and (D) stylet protruded.

pumped back into the pharynx. Many species from these two orders feed on plant roots, puncturing plant cell walls with their stylets, but others are predatory or parasites of invertebrate animals. Some tylenchids invade plant tissues to feed as endoparasites.

Associations. Bacteria form mutually beneficial associations (symbiosis) with nematodes. The insect-pathogenic nematodes *Steinernema* and *Heterorhabditis* carry different species of *Xenorhabdus* in their intestines, which their soil-living infective juveniles inject into any insect they can invade. Bacterial toxins kill the insect, and the nematodes then multiply for several generations, feeding on the bacteria that proliferate in the cadaver. Some stylet-feeding dorylaimids aquire plant-pathogenic viruses, which they transmit to new plant hosts while feeding.

Predatory fungi

Many nematodes feed on fungi, while many soil fungi specifically attack nematodes. Some soil fungi possess sticky traps for nematodes, and others form loops, which close tightly around a nematode that pokes its head into the trap. The loops respond to specific chemicals on the surface of the nematode cuticle. The nematode is then digested. Some fungi only form such traps when nematodes are in the vicinity.

The pine-wilt nematode *B. xylophilus* feeds and multiplies on fungi but can also feed on pine tissues. After a period of reproduction some juveniles enter a stage in which they can invade the wood-boring larvae of certain beetles (Cerambycidae). When they become adults, these flying insects carry the nematodes to new pine trees, spreading the infection and in due course killing the trees.

FORM AND FUNCTION

General form and external features. The great majority of nematodes have slender, elongated cylindrical bodies, without obvious external appendages, that move in graceful curves (see Figure 2). The external cuticle that covers the whole body is often transparent and may be smooth, but it usually has closely spaced grooves, which at high magnification can be seen encircling the body (annulation) and which give the animal a segmented appearance. Some have a thicker cuticle with obvious annulation, or they may have rows of pits, knobs, ridges, bristles, or other ornamentation. The mouth, at the anterior end, usually leads into a cuticle-lined buccal cavity that may be armed with a variety of teeth, jaws, or stylets, depending on the kind of food usually ingested. The anus is near the posterior end, with a short or long, thin post-anal tail, the latter facilitating swimming. Most marine nematodes have caudal glands in the tail that secrete mucus through a pore or spinneret, enabling the worm to maintain a hold on an object and not be washed away. The female sexual opening (vulva) is usually mid-ventral but may be anywhere between the head or anus. The male sexual organs usually open adjacent to the anus. Two cuticular rods, the spicules, used to open the female's vulva, open into the anus. Occasionally, cuticular flaps that surround the male sexual opening also clasp the female during copulation.

Darkening of the cuticle

The adult nematomorph has an elongated cylindrical body resembling a nematode. When the fully grown worm escapes from the body of its host it has a rough, thick cuticle that becomes darkly coloured. The tip of the head is pale and the hind end is lobed where the anal sex organs open. When it hatches from the egg, the head of the larva possesses a proboscis with stylets that enable it to penetrate the body of an appropriate invertebrate host, often an insect, in which it completes its growth.

Gastrotrichs have a short body that swims smoothly by means of cilia, which typically are arranged in bunches on the head and in bands along the ventral surface of the body (Figure 3). The cuticle may be spiny or form plates or scales. Adhesive tubes, the outlet of glands, are a feature of gastrotrichs also found in kinorhynchs and a few unusual nematodes (*e.g., Draconema*). In gastrotrichs there may be many adhesive tubes situated along the sides of the body or merely two at the tips of a forked hind end of the body. Kinorhynchs have a short body of 13 to 14 segments with cuticular body spines (Figure 3). There is an oral cone at the front armed with stylets, and there are further hinged spines around the first segment. The first segment together with the spines and stylets can be completely withdrawn into the second and third segments.

Rotifers show a great diversity of body form (Figure 3). Cilia surround the anterior end of the body and mouth, forming the corona. Cilia may form circumoral rings or two or more ciliated lobes or be arranged in other ways. In some sessile rotifers the corona, which takes the form of a ring of long, thin lobes bordered by stiff bristles, replaces beating cilia and serves to entrap prey. Behind the head, with its corona, is a trunk region leading, in

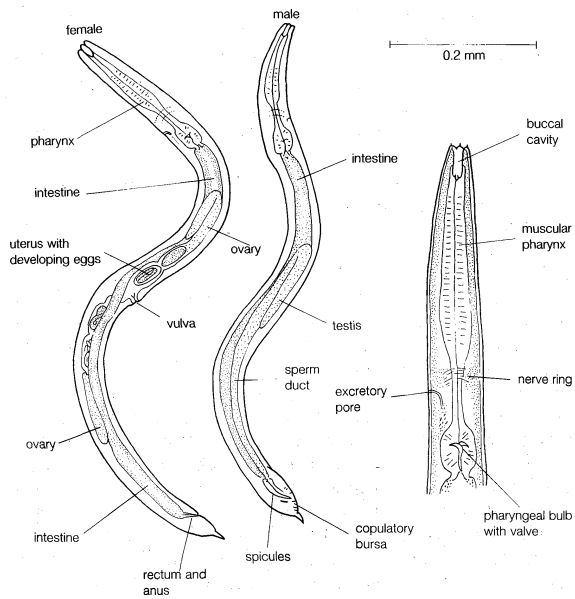


Figure 2: Female and male *Rhabditis oxycerca*, a free-living soil and freshwater bacteria-feeding nematode; enlarged head of female at right.

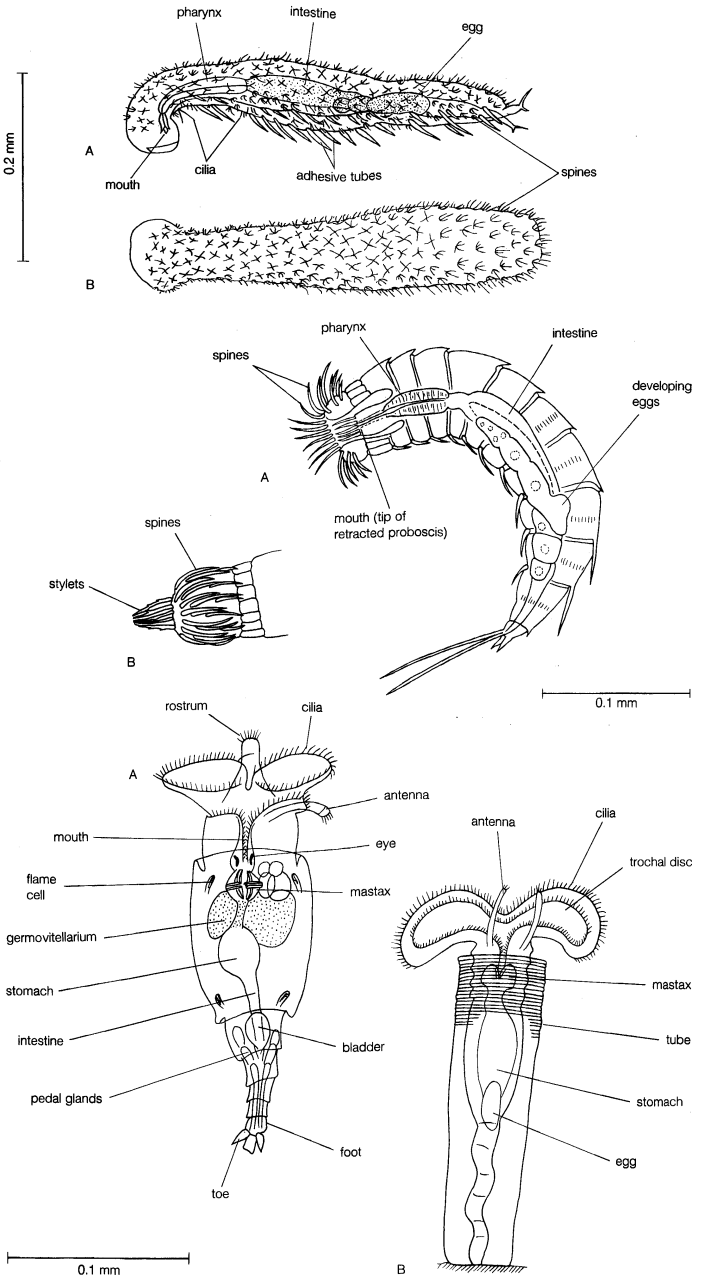


Figure 3: *Representative selection of aschelminths.* (Top) A marine gastrotrich (Macrodasyida) from a sandy beach seen from (A) the side and (B) above. (Centre) A marine kinorhynch from estuarine sand with proboscis (A) retracted and (B) protruded. (Bottom) Rotifers showing (A) a freshwater bdelloid and (B) a sessile flosculurid.

many, to a narrower tail, or foot, often ending in two or more toes. The foot often has a pair of mucus-secreting glands opening on the toes and the cuticle can be annulated, superficially appearing to form segments. Some free-swimming rotifers have a spherical trunk, sometimes with long spines extending from it. Many have the trunk cuticle strengthened to enclose the body in armour (the lorica).

Internal features. The graceful body waves by which nematodes move are brought about by the alternating contraction of dorsal and ventral longitudinal muscles acting on the cuticle and opposed by hydrostatic pressure in the body tissues. This hydraulic skeleton affects all aspects of the nematode. The cuticle must resist stretching in length and circumference and yet remain flexible. The cuticle is a multilayered structure strengthened by an internal system of fibres, struts, or plates. It is secreted by the layer of underlying cells, the hypodermis, before each molt. The longitudinal muscles are obliquely striated. *C. elegans* possesses the same molecular and biochemical machinery in its muscle cells as do higher animals, such as vertebrates, but is arranged somewhat differently. The cross-banding characteristic of striated muscle fibres in vertebrates is replaced by bands making a small angle to the long axis of the cell. The sarcomeres (the structural units of striated muscle that shorten as the muscle contracts) of nematode body wall muscles in adjacent cells are staggered, instead of being arranged side by side. Obliquely striated muscles contract more slowly than cross-striated muscles, but they can maintain tension when stretched to a greater degree, which is important in an animal that must coil and uncoil a long body. The mouth leads into a pharynx (or esophagus), which is a muscular pump opened by intrinsic radial muscles and closed by the elastic cuticular lining and fluid pressure. The pharynx has a triradiate symmetry (Figure 4). It forces food through the straight non-muscular intestine. A short rectum leads from the intestine to the anus.

The cuticle

The alimentary canal of gastrotrichs and kinorhynchs is like that of nematodes. A buccal cavity leads into a muscular pharyngeal pump, which is triangular in cross section. It is followed by a simple intestine, without glands, the rectum, and the anus. In the nematomorphs the gut is greatly reduced, and during their growth as endoparasites food is probably absorbed through the body surface.

The alimentary canal of rotifers is quite different. The funnellike mouth leads into a pharynx armed with a complex cuticular organ, the mastax, or trophi, which serves to masticate food. (The muscular pharynx is called a mastax; hard jaws within the pharynx are called trophi.) It takes many different forms but basically possesses a fulcrum and movable lateral opposable pieces. Food then travels via an esophagus to a saclike stomach, from which it passes to an intestine and finally to a dorsal anus. The rotifer alimentary canal is unlike that of other aschelminths, with, for example, the digestive glands opening into the stomach instead of the pharynx. The mastax is unique to rotifers.

The cuticle is secreted by a hypodermis. Instead of blocks of longitudinal muscles, there are discrete strands of muscle in the body cavity, and there are also muscles around the stomach. There is a spacious body cavity. Excretory organs, known as protonephridia, are found in rotifers, many gastrotrichs, and kinorhynchs; they are not found in nematodes or nematomorphs. Protonephridia are tubes, usually paired, sometimes branching, that end blindly within a cell. The cells, called flame cells from their appearance, set up a water current with waving protoplasmic hairs (cilia or flagella). Their primary function is not excretion but the expulsion of excess water to the exterior through an excretory pore (*i.e.,* osmoregulation). Nematodes and nematomorphs lack flame cells, but nematodes possess one or more ventral cells that contain canals running along both sides of the body. These canals open to the exterior by a ventral pore.

Flame cells

Aschelminths have a simple brain and systems of ganglionated nerves coordinating the muscles and serving the sense organs. In nematodes the brain forms a nerve ring around the pharynx. Sense organs take the form of nerve endings associated with sensitive hairs or papillae concentrated on the head and around the copulatory organs. Sensory ciliated pits probably are chemoreceptors.

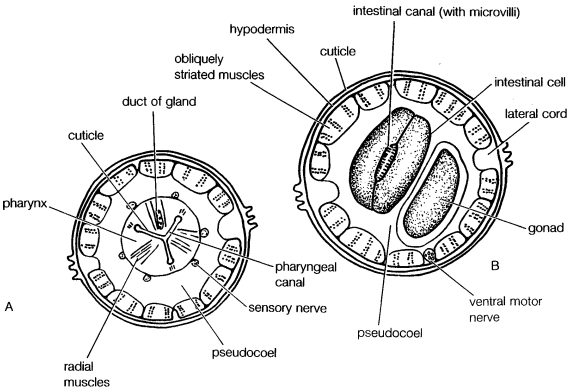


Figure 4: Diagrammatic transverse sections through (A) the pharyngeal and (B) the intestinal regions of a nematode.

In nematodes there are two deep pockets opening near the mouth, each containing a gland and a number of sensory nerve endings, the amphids. The structure of nematode sense organs shows them to be modified cilia, whereas normal locomotory cilia do not occur in nematodes. Aschelminth eyes, one or more pairs, located on the head or within the pharynx, possess a pigment cup, nerve endings, and a simple lens.

Reproductive organs

The reproductive organs are usually paired and tubular. The female organs consist of an ovary, an oviduct, a receptacle for sperm, and a gonopore to the exterior. There may be accessory glands and muscles. The eggs are fertilized within the gonad following copulation and then secrete a shell. The egg may contain yolk, as in nematodes, or enclose separate yolk cells produced in a separate organ, as in rotifers. The testis, which may be single or double, passes sperm down a sperm duct to the male pore. Often this opens together with the anus. There often are accessory copulatory organs forming a kind of penis, as in many rotifers. Nematodes have a pair of cuticular rods, the spicules, to open the female pore, the vulva, which may be located ventrally anywhere between the head and the anus. Nematode sperm, rather than swimming with a protoplasmic hair (flagellum), crawl (amoeboid movement).

EVOLUTION AND PALEONTOLOGY

The only aschelminth fossils known are some nematodes in amber, but these are, in geologic terms, recent fossils, being not more than about 100,000,000 years old. They are similar to, or the same as, living aschelminths. More recent still are parasitic nematodes in fossilized feces.

In the absence of useful fossils, the evolutionary history of the aschelminths can only tentatively and incompletely be reconstructed from their structure and development. There are strong similarities in the pharynx, except in rotifers, and in the adhesive tubes or caudal glands. There are sufficient similarities between gastrotrichs and nematodes, especially in their digestive organs, to suggest a remote common ancestor. One hypothesis suggests that in the Precambrian seas these aschelminths inhabited marine sediments, but, whereas the gastrotrichs preserved the primitive ciliary form of locomotion, the nematodes lost functional cilia (though their sense organs show evidence of derivation from cilia) and adopted sinuous movements powered by muscles to move through the spaces between sand, mud, and soil particles. Both invaded fresh waters, and the nematodes further invaded the soil. The kinorhynchs also resemble both groups in a number of organ systems, but segmentation and the retractable head are unique. Their affinities are much less clear, though they remain inhabitants of marine sediments.

Parasitism

The adoption of parasitism has been accompanied by a great increase in the size of some nematodes and nematomorphs. There are sufficient similarities between nematomorphs and nematodes to make a distant common ancestor likely. The marine nematomorphs are probably the most primitive. Reduction of the alimentary canal is often associated with parasitism. Of the two nematode subclasses, the Adenophorea are the most diverse and, except for the Dorylaimida, primarily marine and not parasitic, while the second subclass, the Secernentea, are primarily freshwater and terrestrial, with many parasitic species. Perhaps the Dorylaimida and Secernentea evolved with the rise of the land flora and fauna, many becoming parasites of plants and animals.

The evolution of the rotifers is not clear. They show some resemblance to the smallest flatworms (Turbellaria). The most primitive are probably the few marine seisonids, with the freshwater forms being more advanced.

Protonephridial organs, present in rotifers, gastrotrichs, and kinorhynchs, are probably lacking in nematodes and nematomorphs because the locomotory system requires higher internal hydrostatic pressure. The characteristic, determinate pattern of cell division in the embryo giving rise to an adult of few cells, even in the larger parasites such as *A. lumbricoides,* perhaps can be explained as a primitive feature on an ancestral aschelminth adapted to life in the small spaces in marine sediments, where miniaturization was an advantage.

CLASSIFICATION

Annotated classification.

PHYLUM ASCHELMINTHES (or Nemathelminthes)

Multicellular; bilaterally symmetrical; triploblastic; mostly microscopic, though some parasitic species many centimetres long; body surface, mouth, and pharynx covered by the cuticle, sometimes with spines, scales, or mouthparts; pseudocoelom; includes a mouth, pharynx, intestine, anus, simple brain, and sensory and motor nerves; some with protonephridial osmoregulatory organs; some possess simple eyes; without respiratory or circulatory systems; reproduction bisexual, parthenogenetic, or hermaphroditic, but not asexual. Development determinate, with small numbers of cells in predetermined, predictable positions in each species.

Class Rotifera

Microscopic free-swimming, crawling, or sedentary animals; feeding and swimming by cilia (beating protoplasmic hairs); a ciliated organ, the corona, of diverse form, on the head; the pharynx, with mastax; body form very diverse, often with an elongated foot.

Order Seisonidea. Marine species on the surface of Crustacea; not parasitic; weakly developed corona; bisexual.

Order Monogononta. Mostly freshwater; free-swimming or sessile; males smaller and less common than females, often seasonal; female with single gonad.

Order Bdelloidea. Mostly freshwater; swimming or crawling; corona in two parts; retractable head; female gonad paired; no males.

Class Gastrotricha

Microscopic; swimming or crawling by means of cilia, often in tufts on the head or in bands on the ventral surface of the body; cuticle, often with spines, scales, or plates; muscular triradiate pharynx, but without corona or mastax.

Order Macrodasyida. Marine; adhesive tubes, often numerous, along the body.

Order Chaetonotida. Mostly freshwater; distinct head and forked hind end on the tips of which caudal glands open; protonephridia.

Class Kinorhyncha (or Echinodera)

Marine; microscopic; body with 13 or 14 segments; first segment (the head) with stylets and spines withdraws into the following segments; movement by protruding and retracting head.

Class Nematoda (or Nemata)

Elongated; cylindrical; crawling or swimming by sinuous movements; no functional cilia or flagella; usually less than 2 mm long; parasitic species may be larger; cuticle smooth, annulated, or otherwise ornamented; triradiate muscular pharynx; no protonephridia.

Subclass Adenophorea

Mostly marine; nonparasitic, except Dorylaimida; usually with caudal glands and a single ventral excretory cell and pore; amphids (lateral cephalic sense organs) posterior to lips.

Order Chromadorida. Mostly marine and nonparasitic; amphids spiral or circular; each ovary usually folded back on oviduct.

Order Monhysterida. Mostly marine and nonparasitic; amphids spiral or circular; single or paired ovaries outstretched.

Order Enoplida. Mostly marine and nonparasitic; amphids pocket-shaped; characteristic stretch receptors (metanemes).

Order Dorylaimida. Mostly soil-inhabiting; many feed by means of hollow stylet, others by teeth; amphids a slit or pocket; includes parasites of invertebrates and vertebrates; large glands (stichosomes) associated with pharynx.

Subclass Secernentea

Mostly terrestrial or parasitic; without caudal glands; excretory organs often intracellular canals running along sides of the body, opening by ventral pore; amphids inconspicuous pores on lips.

Order Rhabditida. Mostly nonparasitic; terrestrial; bacteria-feeding; pharynx expands posteriorly into a muscular bulb with valve.

Order Diplogasterida. Soil-inhabiting; bacteria-feeding or predaceous; pharyngeal bulb; mouth often with teeth.

Order Tylenchida. Many feed on higher plants, others on fungi or predatory, using stylet; others, insect parasites; mostly terrestrial; characteristic hollow, slender mouth stylet; pharynx with bulb, valve, and prominent glands.

Order Ascaridida. Very large intestinal parasites; club-shaped or cylindrical pharynx.

Order Oxyurida. Intestinal parasites; pharynx with posterior valved bulb.

Order Strongylida. Parasitic; often with nonparasitic larvae; males with characteristic copulatory muscular flaps (bursa).

Order Spirurida. Parasites of vertebrates; larval stage in invertebrate host; pharynx with muscular and glandular parts; includes Filarioidea transmitted by insects.

Class Nematomorpha

Develop as parasites of invertebrates; free-living; reproduce sexually as adults; up to 1 m long; elongated; cylindrical; swimming or crawling by sinuous movements.

Order Gordioidea. Freshwater or terrestrial; single ventral hypodermal cord; cell-filled body cavity.

Order Nectonematoidea. Marine; dorsal and ventral hypodermal cords; fluid-filled body cavity.

Critical appraisal. The true evolutionary relationships of the five aschelminth classes is a problem because it is difficult to reconstruct a possible common ancestor and because there are no relevant fossils. Since it is a fundamental principle of zoology that the classification of animals should be based on their evolutionary history, some zoologists treat each of the five classes as a phylum, thereby implying that they are not necessarily more closely related to one another than to some other simple invertebrates. Except for the Nematoda, there is little difficulty in subdividing the classes into orders (or separate phyla into classes).

With the Nematoda it is difficult to give a satisfactory classification. One reason for this is that those most concerned, marine biologists, plant pathologists, and animal parasitologists, have worked in isolation, each putting forward classifications that raise the ranks and increase the subdivisions of the nematode groups on which they work, while combining and reducing the importance of other groups. The leading authorities have proposed irreconcilable hypotheses on the evolutionary history of the nematodes, which underline the different classifications.

BIBLIOGRAPHY. Two comprehensive classical works on the aschelminths are L.H. HYMAN, *The Invertebrates,* vol. 3, *Acanthocephala, Aschelminthes, and Entoprocta, the Pseudocoelomate Bilateria* (1951); and PIERRE P. GRASSÉ (ed.), *Traité de zoologie: anatomie, systématique, biologie,* vol. 4, fascicle 2, *Nèmathelminthes (nèmatodes),* and fascicle 3, *Nèmathelminthes (nèmatodes, gordiacés), rotifères, gastrotriches, kinorhynques* (1965). More recent accounts include PAUL A. MEGLITSCH, *Invertebrate Zoology,* 2nd ed. (1972); and VICKI PEARSE *et al., Living Invertebrates* (1987), especially ch. 12 and 13. Papers on various aspects of rotifers are collected in the published proceedings of the INTERNATIONAL ROTIFER SYMPOSIUM; four meetings had been held by 1987.

There is a much larger literature on the nematodes than other aschelminths because of their importance to humans. Introductory works include NEIL A. CROLL and BERNARD E. MATTHEWS, *Biology of Nematodes* (1977); ARMAND MAGGENTI, *General Nematology* (1981); WARWICK L. NICHOLAS, *The Biology of Free-Living Nematodes,* 2nd ed. (1984); and GEORGE O. POINAR, JR., *The Natural History of Nematodes* (1983). Parasitic forms are discussed in WILLIAM R. NICKLE (ed.), *Plant and Insect Nematodes* (1984); GERALD D. SCHMIDT and LARRY S. ROBERTS, *Foundations of Parasitology,* 3rd ed. (1985); and NORMAN D. LEVINE, *Nematode Parasites of Domestic Animals and Man* (1980). Nematodes in biologic research are the subject of BERT M. ZUCKERMAN (ed.), *Nematodes as Biological Models,* 2 vol. (1980).

(W.L.N.)

Asia

Asia is more a geographical term than a homogeneous continent. The most diverse of all continents, it extends over a latitudinal range of 92° from north to south, has the greatest range of land height of any continent, is subject to climates ranging from Arctic to tropical, and produces the most varied forms of vegetation and animal life in consequence. Similarly, its patterns of human adaptation range from the lives of the nomads of Arabia and Central Asia to those of the crowded cities of the Yangtze Basin and the Gangetic Plain.

Asia is the largest continent, occupying 30 percent of the world's land area, with a mainland area of approximately 17,000,000 square miles (44,000,000 square kilometres). To the east the Pacific Ocean forms its natural boundary; the chains of islands that include the component territories of Japan, Taiwan, the Philippines, and Indonesia are part of Asia. To the west the boundary is generally regarded as running southward along the eastern slope of the Ural Mountains, after which it turns approximately southwestward to the northern shore of the Caspian Sea, from where it again runs generally southwestward to the Kuma River, thence following the Kumo-Manych Depression northwestward to the Sea of Azov; from the Black Sea, the coast of Asia Minor and the Mediterranean coast of the Levant form Asia's western limits, after which the boundary runs south across the Isthmus of Suez and along the coast of the Arabian Peninsula.

This article treats the physical and human geography of Asia, followed by discussion of geographical features of special interest. For discussion of individual countries of the continent, see specific articles by name, *e.g.,* CHINA, INDIA, and JAPAN. Other Asian countries are treated in articles on regions under the titles ARABIA and SOUTHEAST ASIA, MAINLAND. For discussion of major cities of the continent, see specific articles by name, *e.g.,* DELHI, PEKING, and TOKYO–YOKOHAMA METROPOLITAN AREA. For discussion of the history of specific Asian regions, see the articles PALESTINE and ISLĀMIC WORLD, THE. Related topics are discussed in articles on religion (*e.g.,* BUDDHISM, THE BUDDHA AND; HINDUISM; and ISLĀM, MUḤAMMAD AND THE RELIGION OF) and arts and literature (*e.g.,* CHINESE LITERATURE; JAPANESE LITERATURE; and SOUTH ASIAN ARTS). For further references, see also the entries for these topics in the *Index.*

The article is divided into the following sections:

PHYSICAL AND HUMAN GEOGRAPHY

General considerations

Asia is the most populous of the continents, containing more than half the human race. The continent includes the two most populous countries in the world—China and India—in addition to Japan and Indonesia, each of which has a densely packed population; among non-Asian nations these last two countries are surpassed in numbers only by the populations of the Soviet Union (the territory of which also extends into Asia) and the United States. Two other Asian countries with very large populations are Bangladesh and Pakistan. During most of the 20th century the population of Asia grew at a rate faster than the world average. There is a general awareness in Asia, especially in the most populous countries, of the need for some regulation of growth. Nevertheless, it appears that—barring an unexpected breakthrough in the acceptance of birth control—the population of Asia may reach 3,400,000,000 by the year 2000.

Origin of the name Asia

The name Asia is very ancient, and its origin has been variously explained. The Greeks used it to designate the lands situated to the east of their homeland. It is also believed that the name may be derived from the Assyrian word *asu,* meaning "east." Another possible explanation

is that it was originally a local name given to the plains of Ephesus and gradually extended to include Anatolia (contemporary Asia Minor) and the rest of the continent.

Geological development has resulted in a configuration that includes the Arabian massif and the plains of Iraq; the mountain belts of Tūrkey, Iran, and Afghanistan; the great Himalayan mountain chain stretching from Afghanistan to the Burmese peninsula; the Indo-Pakistan subcontinent; Central Asia; the vast Siberian lowlands extending from the Urals to the Pacific; and the great island chains that sweep in arcs from Japan to Indonesia.

As a result of this configuration Asia's population is unevenly distributed. Thus, while there has been a concentration of population on the Arabian plains and river valleys, in the Indo-Pakistan subcontinent, in Central Asia, especially China, and to some extent in the Pacific borderlands and on the islands, there are vast areas with a low density of population.

Siberia, which represents approximately one-third of the land area of Asia, has an estimated population of about 11 persons per square mile (four persons per square kilometre), compared with an average density of 156 persons per square mile in the rest of Asia.

The mountain systems of Central Asia not only have provided the great rivers with water from their melting snows but also have formed a forbidding natural barrier that has influenced the movement of peoples into the area. Migration is possible only through mountain passes. As a result the historic movement of population has been broadly from the arid zones of Central Asia through the mountain passes into the Indo-Pakistan subcontinent, from China through Southeast Asia to modern Indonesia and Malaysia, and from the Arabian Peninsula and from India across the Bay of Bengal into Indonesia and Malaysia. The Japanese people and, to a lesser extent, the Chinese have remained ethnologically more homogeneous than the populations of other Asian countries.

Asia's religious heritage

Asia has been the birthplace of all the great world religions, including Buddhism, Christianity, Hinduism, Islām, Judaism, Sikhism, Taoism, and Zoroastrianism. Of these, only Christianity moved westward; it subsequently exerted little influence, in its religious aspects, on Asia, although many Asian countries have Christian minorities. Buddhism has had a greater impact outside its birthplace in India and is prevalent in various forms in China, Korea, Japan, the Southeast Asian countries, and Sri Lanka. Islām has spread out of Arabia eastward to Afghanistan, Pakistan, and India, and thence to Malaysia and Indonesia, as well as west and south to several areas of Africa. Hinduism, basically a non-proselytizing religion, has been mostly confined to the Indian subcontinent.

Before the Industrial Revolution in the 18th and 19th centuries, most principal technical achievements began in Asia. Three millennia before Christ, Asians knew the arts of cooking and pottery and the use of fire for the smelting of ores. They had already progressed from nomads to cultivators, using irrigation and practicing crop rotation. They had learned to domesticate animals and had invented the wheel, the harness, the saddle, and the chariot. They had begun to use a form of paper and had developed elaborate scripts. Early Asians were familiar with wood carving, stonecutting, and the casting of metals and left monuments of stone and metal that still evoke admiration and astonishment. They used the art of calculation, including the decimal system, and employed means of measurement. Various indigenous systems of medicine were developed, many of which are still in use. In ancient times major Asian empire-states arose, employing intricate systems of laws and regulations and delegating power and authority to institutions of government at various levels.

The impact of the West

In the 15th century Asia's material advances were envied by people in Europe. The fabled prosperity of Asia prompted the movement of adventurous spirits from Western nations to Asia and led to the eventual conquest of many Asian countries. The Western impact on the civilizations of these Asian countries, on their attitudes, and on the development of their political institutions and economic systems will only be referred to briefly here. Even though almost all Asian countries have attained their independence, the era of European conquest and colonization has left an indelible mark on much of Asia. The Asian countries received many benefits by way of the establishment of an infrastructure, a system of posts and telegraphs, railway and road networks, ports and harbours, public health systems, modernized educational systems, and the rudiments of an apparatus of modern Western-style government—civil, judicial, and military—as well as the impetus for publishing newspapers and magazines. On the other hand, there was little progress in industrialization. The effect of the neglect of industrial development during these years of colonization is still noticeable in commercial relations among many Asian countries. Broadly, it may be said that the role of most Asian countries—with the exception of Japan and a few others that did not feel the impact of conquest and colonization—was that of primary producers exporting their products to the metropolitan countries, where they were turned into finished products and reexported back to the Asian countries. As a result, most Asian countries have economies that are imbalanced, and efforts to produce greater economic integration and coordination among Asian countries have had only limited success.

In addition to the uncertain economic situation, the lack of any ethnic, religious, or cultural homogeneity among the Asian peoples has made it difficult for a common Asian political consciousness to arise. In the struggle for independence most of the Asian countries developed a political philosophy favouring nationalism. Furthermore, since achieving independence, many Asian countries seem to have lost the unifying force of emergent nationalism and are groping to find a substitute for it. Even the doctrine of nonalignment, which formerly seemed to be common to several Asian countries, has lost its impetus.

In the economic field Asian countries are caught between two conflicting forces. On the one hand, the need to ensure improved levels of living has led to a desire to move toward orderly and planned economic and social development. On the other hand, many countries there have a long tradition of otherworldliness and austerity. This conflict of forces may produce a new philosophy of economic development in Asia that will reconcile the need for material progress with the imperative of not abandoning spiritual and religious values.

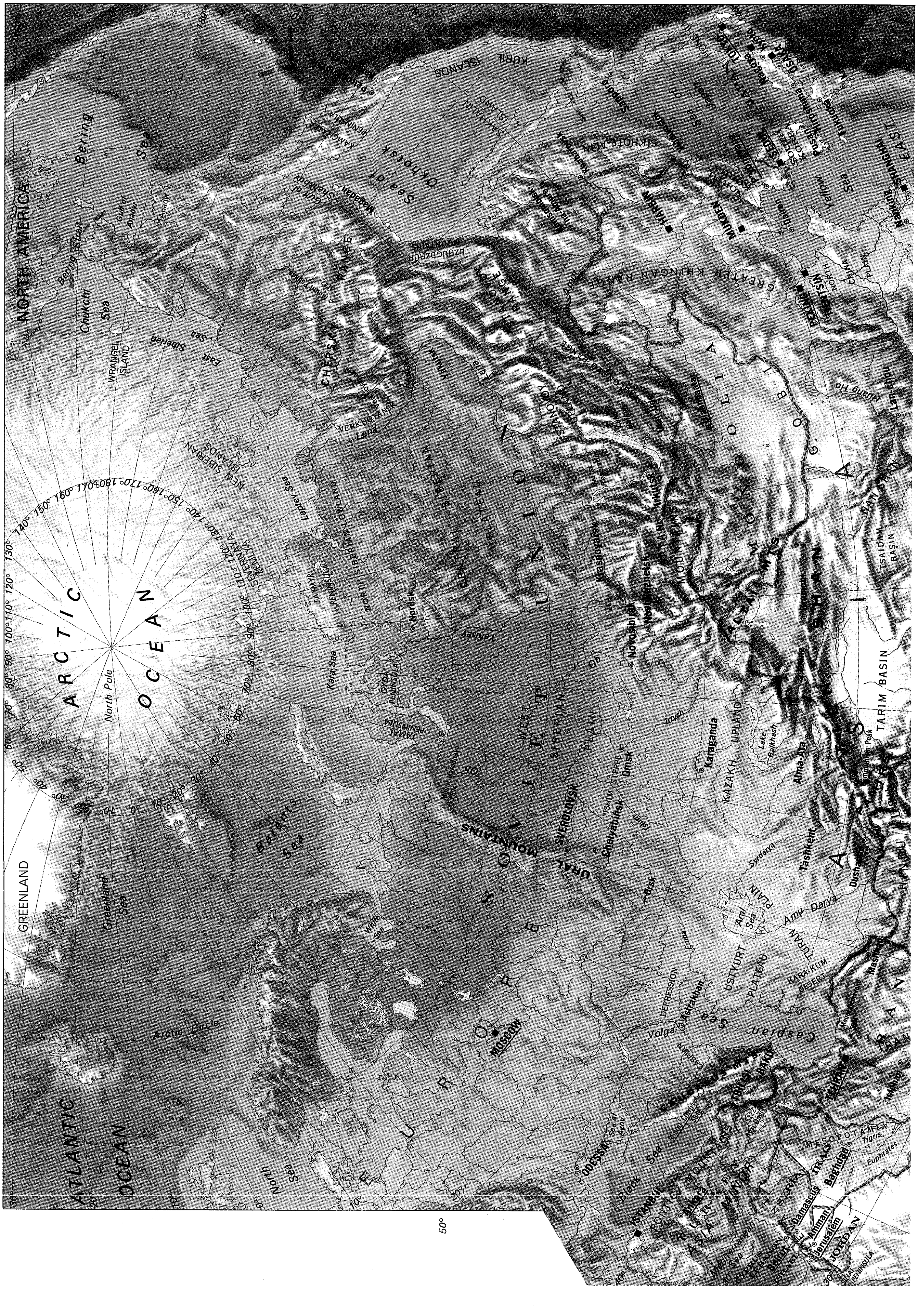
Asia's entry into the nuclear age

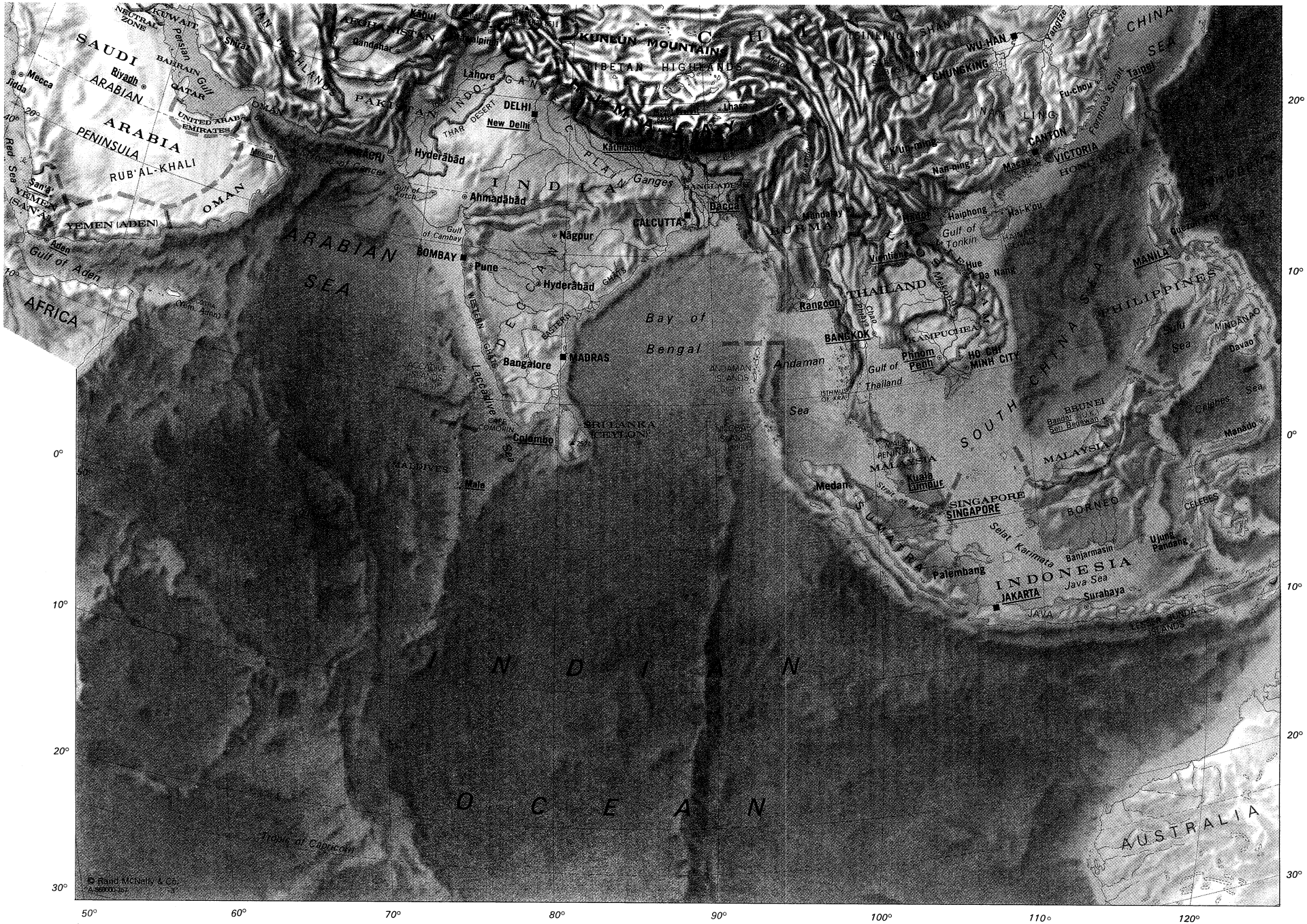
The political forces that are emerging in Asia cannot be ignored. China has begun to play an important role in world affairs; until 1974 it was the only nuclear power in Asia. Japan has become one of the great industrial nations of the world and clearly has the capability to develop into a nuclear power, although the Japanese, in the light of their own experience at the end of World War II, have accepted a constitution renouncing war as a sovereign right of the nation and abjuring the threat or use of force as a means of settling international disputes. India became a nuclear power in 1974. Nuclear proliferation in Asia, coupled with the region's political instability, is a matter of growing international concern.

The modern political history of the continent has resulted in part from the political vacuum created by the withdrawal of the colonial powers. The continuing disturbed political situation in the Middle East illustrates this fact. The difficulties that arose between India and Pakistan after their independence in 1947 and between Indonesia and Malaysia from 1963 to 1968 are clearly part of the same pattern, as are the conflicts that, during the 1960s, 1970s, and 1980s, produced war in the countries of Indochina.

With a new awakening of political consciousness, with a new spirit of regional cooperation, and with a new urge at the grassroots level for the improvement of living standards, there are positive factors at work that could make possible a resurgence of Asia. The high rate of population growth throughout most of Asia constitutes perhaps the greatest negative factor. Several Asian countries have shown, however, that even under existing circumstances great progress can be made in the limitation of population growth. If this trend becomes more widespread, the chance that Asia will become economically stronger and socially better developed will be correspondingly increased.

(C.V.N./Ed.)


ATLANTIC OCEAN
GREENLAND
Greenland Sea
ARCTIC OCEAN
North Pole
Arctic Circle
Barents Sea
Kara Sea
White Sea
North Sea
EUROPEAN SOVIET UNION
MOSCOW
ODESSA
Sea of Azov
Black Sea
ISTANBUL
Ankara
PONTIC MOUNTAINS
ASIA MINOR
TURKEY
CAUCASUS MTS
TBILISI
BAKU
Caspian Sea
CASPIAN DEPRESSION
Astrakhan
Volga
USTYURT PLATEAU
KARA-KUM DESERT
TURAN PLAIN
Aral Sea
Amu Darya
Syr Darya
Tashkent
HINDU KUSH
Mashhad
IRAN
TEHRAN
Isfahan
MESOPOTAMIA
Tigris
Euphrates
IRAQ
Baghdad
SYRIA
Damascus
Beirut
LEBANON
CYPRUS
Mediterranean Sea
ISRAEL
Amman
Jerusalem
JORDAN
SINAI PENINSULA
URAL MOUNTAINS
SVERDLOVSK
Chelyabinsk
Orsk
YAMAL PENINSULA
GYDA PENINSULA
Ob
WEST SIBERIAN PLAIN
ISHIM STEPPE
Omsk
Irtysh
Karaganda
KAZAKH UPLAND
Lake Balkhash
Alma-Ata
TIEN SHAN
TARIM BASIN
Novosibirsk
Novokuznetsk
Krasnoyarsk
Yenisey
Norilsk
TAYMYR PENINSULA
NORTH SIBERIAN LOWLAND
SEVERNAYA ZEMLYA
CENTRAL SIBERIAN PLATEAU
SAYAN MOUNTAINS
Irkutsk
ALTAI MTS
Laptev Sea
NEW SIBERIAN ISLANDS
Lena
VERKHOYANSK RANGE
CHERSKY RANGE
Yakutsk
East Siberian Sea
WRANGEL ISLAND
Chukchi Sea
Bering Strait
NORTH AMERICA
Gulf of Anadyr
Bering Sea
KAMCHATKA PENINSULA
Sea of Okhotsk
Magadan
Gulf of Shelikhov
SAKHALIN ISLAND
KURIL ISLANDS
DZHUGDZHUR MOUNTAINS
STANOVOY RANGE
Khabarovsk
SIKHOTE-ALIN
Vladivostok
Sea of Japan
JAPAN
Sapporo
TOKYO
Nagoya
OSAKA
Hiroshima
Fukuoka
SEOUL
Dairen
Yellow Sea
Harbin
MUKDEN
Amur
GREATER KHINGAN RANGE
PEKING
TIENTSIN
NORTH CHINA PLAIN
MONGOLIA
GOBI
Huang Ho
Lan-chou
NAN SHAN
TSAIDAM BASIN
SHANGHAI
Nanking



ASIA

Size of symbol indicates relative size of town • ⊛ ■

Elevations in metres

Geological history

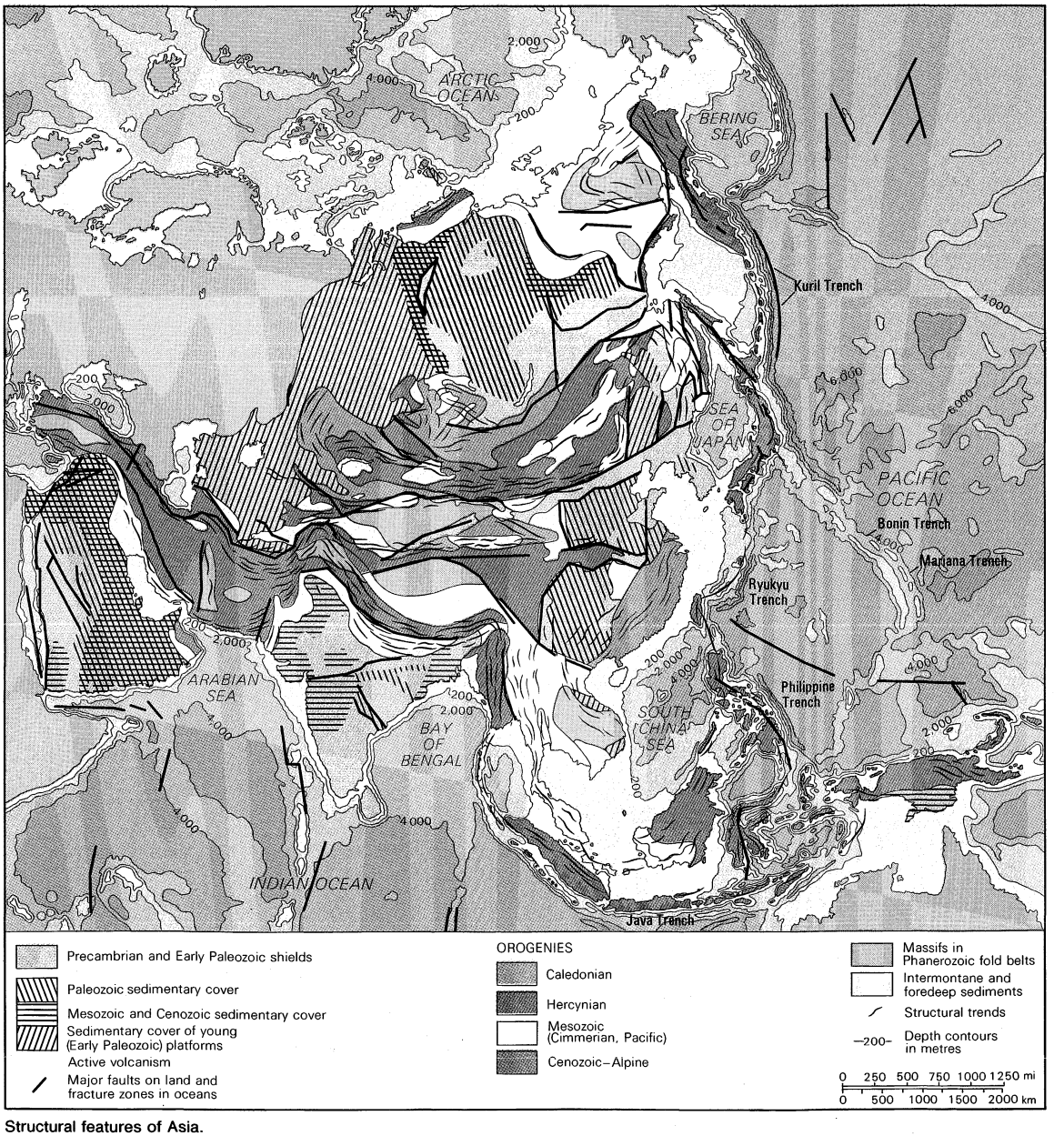
ELEMENTS AND PROCESSES IN THE MAKING OF THE CONTINENT

Platforms, shields, and geosynclines. Asia consists, in part, of several platforms that have been subjected to almost no folding since the Precambrian Era, which lasted from 4,600,000,000 to 570,000,000 years ago. It also consists of vast regions of folding, formed later in various periods or ages, that originated in geosynclines (large and generally linear troughs that gradually subsided over long periods of time and subsequently became filled with thick accumulations of sediments). Contemporary geosyncline systems, still not consolidated by folding, join the shores of the Asian mainland to the Malay Archipelago, underlie bordering seas such as the Sea of Japan and the Sea of Okhotsk, and extend to the island arcs of the Pacific Ocean.

The four continental platforms

The principal continental platforms of Asia are the Siberian Platform in the north, the Chinese Platform in the east, the Indian Platform in the south, and the Arabian Platform in the southwest. Those parts of the platforms where the Precambrian crystalline bedrock is not covered by a sedimentary deposit are called shields; almost all the shield areas were dry land during the last 500,000,000 years and include the Aldan Shield in eastern Siberia, the North Chinese Shield, the Indian Shield in peninsular India, and a large part of the Arabian Shield. In addition, the Russian Platform has played an important role in the geological history of western Asia. Along its western edge the folded structures of the Urals and of Kazakhstan were formed at the end of the Paleozoic Era (which lasted from 570,000,000 to 225,000,000 years ago).

Endogenetic and exogenetic forces. Two geological forces have molded the Asian continent into its present configuration. On the one hand are the endogenetic forces, which represent vertical or lateral forces originating deep within the Earth and that produce fissures in the Earth's crust through which magma, or molten rock, wells up in the form of lava or forms intrusive bodies; when it solidifies, this rock, called basalt, is described as igneous (rock solidified from the molten state). Exogenetic forces, on the other hand, represent those endless processes of erosion and sedimentation that take place on the Earth's surface, where rocks are subjected to weathering and denudation. Associated with these two forces is a principle of equilibrium called isostasy, which represents a theoretical balance maintained between large sections of the Earth's crust, which act as though they were floating on a denser underlying layer. In maintaining this equilibrium, less



Structural features of Asia.

dense material rises vertically while denser material sinks.

The mountain-building process. Mountain building, or orogenesis, is the result of folding or thrusting, which occurs when thousands of feet of sediment accumulate in geosynclines, causing them to sink deeper, thus either activating, or interacting with, vertical or lateral endogenetic forces. In this way the geosynclines themselves, representing zones of structural weakness between crustal blocks or platforms, are folded upward into new mountain systems—a process associated with the upwelling of granitic magma and the deformation of metamorphic rock (rock that has been altered in composition, texture, or internal structure as a result of heat and pressure).

THE TERRITORIAL FORMATION OF ASIA

Blocks and zones of folding

As a result of folding, the intrusion of granite, the swelling of the Earth's crust in response to pressures, the isostatic uplifting (uplifting caused by equal pressure on all sides) of mountains, and the drying up of marine basins that lay in the geosynclinal regions between the continental platforms, ancient blocks of the Earth's crust and weaker zones of folding were united to form the Asian continent. From the formerly inundated geosynclines, several belts of folding—including the Alpine-Himalayan belt—were formed in the Late Paleozoic Era, the Mesozoic Era (from 225,000,000 to 65,000,000 years ago), and the Cenozoic Era (which is dated from 65,000,000 years ago to the present time).

The composition of the continental platforms. *The Siberian Platform.* Greatly deformed schists (metamorphic rocks containing micaceous minerals), gneisses (rocks in which bands rich in granular minerals alternate with bands containing schistose materials), and granites of the Precambrian form the basement of the Siberian Platform. Extending between the river basins of the Yenisey and the Lena, it surfaces within the boundaries of the Aldan Shield, north of the Stanovoy Mountains, and the Anabar Massif (mountainous mass), west of the Lena River and south of the Arctic. In other areas the crystalline bedrock is overlain almost horizontally by sedimentary rock, including marine and salt lagoon deposits from the Cambrian Period (from 570,000,000 to 500,000,000 years ago), marine deposits from the Ordovician Period (from 500,000,000 to 430,000,000 years ago) and the Silurian Period (from 430,000,000 to 395,000,000 years ago), as well as deposits from the Permian Period (from 280,000,000 to 225,000,000 years ago) containing strata of coal. In the western part of the platform, primarily along the edges and in the centre of the broad Tungus syncline (a basin formed by a downward bend of rock strata), east of the Yenisey River, the Paleozoic strata are cut by numerous intrusive sills (layers) of a dark, fine-grained, basaltic rock known as traprock, locally associated with copper and nickel ores and volcanic pipes (cylindrical veins of volcanic origin) containing diamond-bearing kimberlite (an ultrabasic rock of a subcrustal origin). In the southern part of the platform, the lower layers of the sedimentary cover consist of thick layers of Precambrian marine sediments analogous to the Sinian System in China. They were slightly disturbed at the time of the Baikal folding, which occurred at the end of the Precambrian and during the Cambrian Period. Along the northern and eastern edges of the platform, there is widespread distribution of coal-bearing deposits from the Jurassic Period (from 190,000,000 to 136,000,000 years ago) and the Cretaceous Period (from 136,000,000 to 65,000,000 years ago). Mesozoic sedimentary series reach particularly great thickness in the depressions of the Vilyuy River and in the depression near the Verkhoyansk Mountains, which run parallel to the Lena River, both areas having natural gas deposits.

The Chinese Platform. The Chinese Platform consists of four separate massifs—the North Chinese, the South Chinese, the Tarim, and the Tibetan—which probably constituted a single platform in the Precambrian. This platform appears to have broken up at the beginning of the Paleozoic Era, at which time geosynclinal downwarps were formed between the separated blocks. As a result of the folding that took place in these downwarps during the Paleozoic and Mesozoic eras, mountain ranges arose, including the Kunlun and the Tsinling (both in China). The Precambrian bedrock is exposed on the surface primarily within the boundaries of the North Chinese massif, which contains large deposits of ferruginous (iron-bearing) quartzite.

The four Chinese massifs

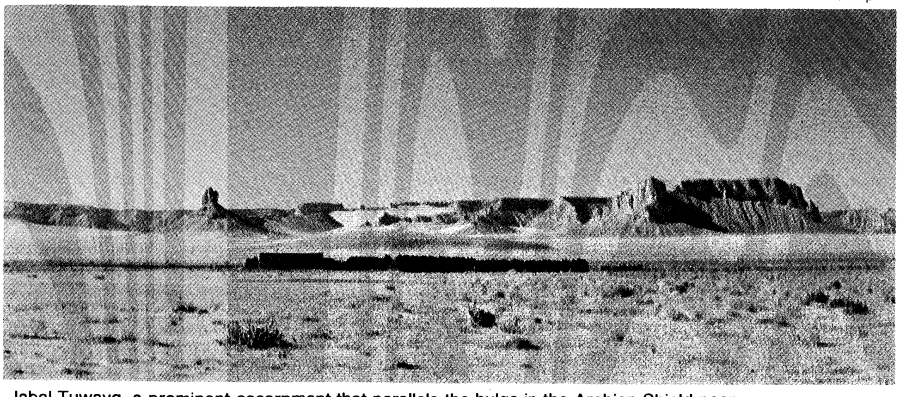
During the Silurian Period and the Devonian Period (from 395,000,000 to 345,000,000 years ago), the southern part of the platform—as well as almost all of it during the later part of the Cretaceous Period—was subjected to folding and faulting deformations that were considerably more intense than those experienced by platforms elsewhere in the world. These deformations were accompanied by intrusions of ore-bearing granite.

The sedimentary rocks overlying the folded bedrock (basement) of the Chinese Platform represent, for the most part, marine sediments of the Precambrian and Lower Paleozoic (Cambrian, Ordovician, and Silurian) eras; marine and continental deposits from the Devonian, Carboniferous (from 345,000,000 to 280,000,000 years ago), and Permian periods, some of which are coal-bearing; and thick layers of rocks of the Mesozoic Era, as well as layers from the Cenozoic Era in some of the depressions. Among the Mesozoic deposits there is an abundance of red fragmented rocks that were formed when the prevailing climate was hot and dry. The sedimentary cover is thickest in the synclines of the Ordos Desert region of Inner Mongolia, of the Chinese province of Szechwan, and of the Lower Huang Ho, as well as in the Mesozoic downwarp of southeastern Korea.

The Arabian Platform. The Arabian Platform, like the Indian Platform, is usually considered to have formed a part of the continent of Gondwana, which began to split apart during the Paleozoic Era. Gondwana is believed to have included Africa, Australia, South America, and possibly Antarctica, and the two platforms do have much in common with these continents. The ancient basement of Arabia is composed of Precambrian crystalline rocks, which crop out in the western part of the peninsula as the Arabian Shield. To the north and east the bedrock is buried under thick layers of horizontal or only slightly disturbed Paleozoic rocks; these are mainly sediments

The crystalline bedrock

Picturepoint



Jabal Tuwayq, a prominent escarpment that parallels the bulge in the Arabian Shield near Riyadh, in central Saudi Arabia.

from the Jurassic and Cretaceous periods and from the earlier part of the Cenozoic Era. These sedimentary strata lie within the bounds of the Mesopotamian downwarp, which occurs in the Persian Gulf region; here the bedrock has sunk to a depth of about two and a half miles and contains large amounts of petroleum.

The Indian Platform. The Indian Platform covers a large part of peninsular India and Sri Lanka. It consists of a Precambrian crystalline shield in which broad grabens (blocks that have been downthrown along faults in the rock on either side) along the Godāvari River and along the Dāmodar River farther to the northeast are overlain by horizontal deposits of the Gondwana system. These were laid down between early Carboniferous and late Jurassic time. The Deccan Plateau, which forms the backbone of peninsular India, is itself overlain by basaltic lavas (traprock from the Late Cretaceous Period and the Early Cenozoic Era). Along the edges of the platform the bedrock is covered by marine deposits from the Jurassic, Cretaceous, and Tertiary (from 65,000,000 to 2,500,000 years ago) periods; in the Gulf of Cambay region, off the west coast of India, these contain petroleum deposits. To the north, the bedrock lies under Cenozoic deposits that reach their greatest thickness in the northern part of the Ganges River Basin, in the marginal downwarp at the foot of the Himalayas, and in the Indus Basin. Deposits of gas and petroleum are associated with the Cenozoic marine and continental sediments, while the Precambrian rocks of the shield contain iron and manganese ores. A small Precambrian outcrop, the Shillong Plateau massif, occupies the eastern part of the platform.

Mountain folding. *The Late Precambrian and the Paleozoic eras.* Those Asian zones that were folded during the Late Precambrian and Early Paleozoic eras consist of various kinds of weakly and strongly metamorphosed volcanic and sedimentary rocks. They are cut by intrusions of granite, granodiorite (a quartz-bearing rock formed at great depth), gabbro (a coarse-grained, dark igneous rock), diabase (a rock of basaltic composition), and serpentinous ultrabasic rocks (*i.e.,* igneous rocks containing less than 45 percent of silica). Deposits of gold, copper, tungsten, polymetallic (*i.e.,* containing many metals), and other ores are associated with these rocks in many areas.

Regions of Late Precambrian and Early Paleozoic folding border the Siberian Platform to the west and south. They compose the Yenisey Ridge, the Eastern and Western Sayan ranges, the Kuznetsk Alatau, the higher (alpine) part of the Soviet Altai, the Mongolian Altai, and a considerable part of the Khangay and Tannu-Ola ranges. Caledonian folding—which is to say, folding that took place during the Caledonian (*i.e.,* Silurian–Devonian) orogeny (mountain-building process)—is also found in the Taymyr Peninsula and on the Severnaya Zemlya Islands, in central Kazakhstan, in the northern chains of the Tien Shan, in the Kunlun Range, and in southeastern China. In the portion of the Caledonian orogeny that occurred during the Devonian Period, a number of large depressions were formed. These included the Kuznetsk and Minusinsk basins and the depressions in the central part of Tuva. Large reserves of coal are concentrated in the deposits laid down in these regions in the Carboniferous and Permian periods.

Areas of depression

The fold structures of the Hercynian orogeny were formed as the result of tectonic activity (movements of the Earth's crust) that occurred during the Middle Paleozoic (Devonian) and Late Paleozoic (Carboniferous and Permian) eras; these constitute a broad arc traversing the central part of the continent. The Hercynian fold structures have a northeasterly and southerly trend between the Russian and Siberian platforms, and a northwest trend in Kazakhstan, along the Salair Ridge, and in the southern Altai; run east to west in the Tien Shan, Kunlun, and Tsinling ranges; and run northeast in northeastern China. According to geophysical data, the Hercynian fold system of the Urals, covered toward the south by a horizontal layer of Mesozoic and Cenozoic deposits, is joined with the Tien Shan. In addition, a branch proceeds from it which passes through the Mangyshlak Peninsula toward the Donets Basin.

A considerable area of the original Paleozoic fold zones was levelled by erosion and subsequently sank, forming large depressions that were filled to an overall depth of between one and four miles with virtually undisturbed marine and continental sediments of the Mesozoic and Cenozoic eras. Associated with these depressions are the vast West Siberian Plain; the Turgay Downwarp, located to the east of the southern Urals; the Turan Platform in the Amu Darya and Syrdarya basins; the depressions of Lake Balkhash and of the lower course of the I-li Ho (river); the Dzungarian Basin; the Fergana and Turfan depressions; the Tsaidam Basin; the Tungliao or Manchurian lowlands in the Sungari River Basin; and others. Only the intermontane depressions, those of Fergana and Turfan, were affected by slight folding; in the other depressions the strata lie almost horizontal. Deposits of petroleum and gas are found in almost all these depressions, from Western Siberia to the Tsaidam Basin and Manchurian Plain in China.

The Mesozoic Era. The belt of Mesozoic folding includes northeastern Siberia (the Verkhoyansk and Chersk ranges), the Sikhote-Alin Range in the Amur region, a large part of Indochina, and possibly the Trans-Himalayas. Folded structures, intersected by numerous granite intrusions, were formed in all these areas as a result of the crumpling of thick layers of geosynclinal deposits of the Permian, Triassic (from 225,000,000 to 190,000,000 years ago), Jurassic, and Early Cretaceous periods. Associated with these folding belts are gold, tin, and polymetallic ores. Moderately large blocks—such as the Kolyma, Indosinian, and other massifs, which resemble the Precambrian or Paleozoic platforms in their structure—are enclosed between the branches of the broad Mesozoic folding zones.

Mesozoic folding zones

Mesozoic folding and the upwelling of magma (molten rock) also affected adjacent areas, such as the Hercynian mountain zone of Mongolia and the Transbaikalia, the Caledonian zone of South China, and a considerable part of the Chinese Platform. Deposits of tungsten, tin, mercury, and other metals are found in these areas.

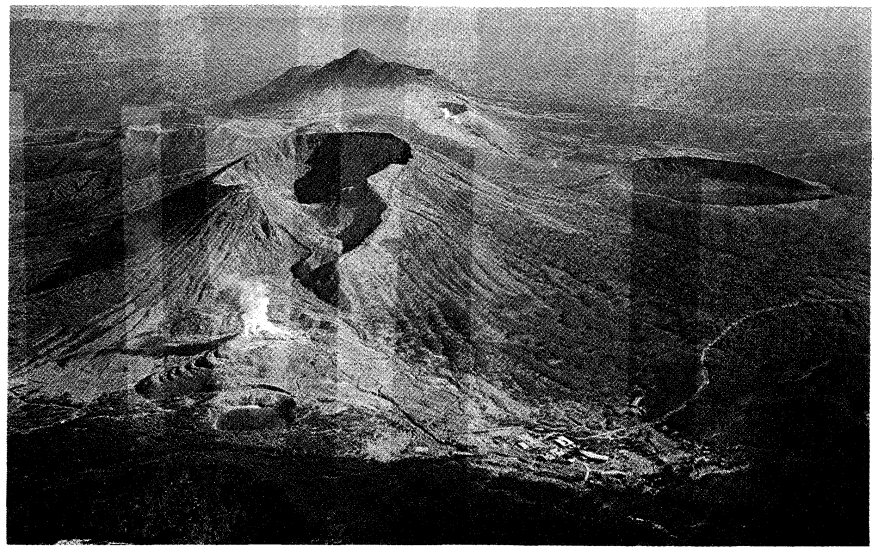
The Cenozoic Era. During the Early Cenozoic Era, a volcanic belt was formed that extends from the Chukotsk Peninsula and the shores of the Sea of Okhotsk through the eastern slopes of the Sikhote-Alin Range, in South Korea, to the southeast coast of China. It runs approximately along the border of the Mesozoic and Cenozoic fold zones. This volcanic belt is composed of massive accumulations of basaltic, andesitic, and acidic lavas dating from the Cretaceous and Early Tertiary periods, with many granite intrusions. It would seem that a series of volcanoes extended through this belt in a long arc, resembling the volcanic island chains of East Asia.

The two Cenozoic fold belts

Areas of Cenozoic folding are confined to two zones—the Alpine–Himalayan belt that traverses Asia from west to east, and the Pacific Ocean belt that runs through the island arcs to unite with the Alpine–Himalayan belt in Indonesia. The Alpine–Himalayan belt was essentially formed on the site of the extensive Tethys Geosyncline. This geosynclinal ocean, the remains of which are to be seen in the Mediterranean and Black seas and in the marine basins of Indonesia, divided two distinct continents—Gondwanaland and Angara, or Angarida—during the Paleozoic, Mesozoic, and Early Cenozoic eras. In the Alpine–Himalayan belt itself there may be distinguished two—and in places three—curving folded mountain ranges; at times these are close together, and at times they diverge. These ranges represent slabs or blocks that were overthrust onto one another or onto the edges of the neighbouring platforms, forming gigantic and complex anticlines (convex folds). The internal parts of their folded structures were sometimes already formed during the Mesozoic Era from Mesozoic and Paleozoic sediments that had accumulated in the Tethys Geosyncline, while some slopes and foothills were formed of Tertiary deposits as a result of less violent folding. The thick sand–clay strata of the Cenozoic Era, which accumulated at the base of the mountains and in depressions in front of them, often contain gas and petroleum.

Northern and southern anticlinoria

The northern series of alpine anticlinoria (*i.e.,* series of anticlines and synclines so arranged that together they form a general arch or anticline) is formed by the Greater



Kirishima Range on Kyushu, Japan's southernmost island; it includes several active volcanoes.
Kokunai Jigyo Kouku

Caucasus Mountains between the Black and Caspian seas; the Turkmen–Khorasan ranges east of the Caspian; and the Selseleh-ye Safīd Kūh (Paropamisus), Pamir, Gissar, and Alai ranges. West of this system of ranges runs still another series of folded structures, separated from the first series by deep troughs. These consist of the Pontic Mountains, the Lesser Caucasus, and the Elburz Mountains. In the related depressions—the Black Sea, the Kolkhida (Rion) and Kura-Aras Lowland, and the South Caspian basin—are concentrated layers, from three to six miles thick, of Cenozoic deposits, which are folded along the edges of the depressions. The southern series of anticlinoria is composed of the ranges of the Taurus, Zagros, Makran and Soleymān mountains, the Hindu Kush, and the Himalayas. Between the northern and southern series of folded mountain ranges are situated the massifs of Menderes and Kirşehir in Turkey, and the central Iranian massif. The small Georgian block occupies a similar position between the Greater and Lesser Caucasus. All these massifs represent regions consolidated by Precambrian or Paleozoic folding, with surface outcrops of bedrock in some places; in others, the bedrock is covered by weakly folded sedimentary deposits of the Paleozoic, Mesozoic, and Cenozoic periods.

Mountain folding in progress. In the folded ranges of Burma, Malaysia, and Indonesia there occurs a transition from the belt of alpine folding—formed on the site of geosynclines that were already landlocked and drained—to a contemporary geosyncline system in which folding is not yet complete. This system, the youngest, occupies the area between the Asian continent itself and the Pacific Ocean and includes the folded systems of the Koryak Range; the Kamchatka Peninsula; Sakhalin Island; Borneo; Celebes Island; the archipelagoes of the Komandorskiye (Commander), Aleutian, and Kuril islands; and the islands of Japan, the Ryukyus, Taiwan, the Philippines, the Moluccas, and the Sundas. Within its structure, geoanticlinal uplifts, consisting of island arcs and mountainous peninsulas, may be distinguished, as well as intermontane troughs, the contemporary geosynclinal downwarps of the seas bordering Asia, and deep-sea trenches on the periphery of the Pacific and Indian oceans. The axial parts of the folded ranges and island arcs in this zone are usually composed of Mesozoic and Upper Paleozoic deposits; they are cut by recent (Tertiary) intrusions and are crowned with a series of active volcanoes. The Tertiary deposits of the intermontane troughs attain great thickness and contain petroleum deposits, such as those of Sakhalin, Japan, and Indonesia; they were folded during the Miocene Epoch (from 26,000,000 to 7,000,000 years ago) and the Pliocene Epoch (from 7,000,000 to 2,500,000 years ago). In some places there are folds in the deposits sedimented in the Quaternary Period (which began 2,500,000 years ago), and as a result of large and recent fractures, Cenozoic lavas consisting of basalts and andesites cover vast areas of the island regions of East Asia.

The youngest geosyncline system

The Alpine–Himalayan belt and, in particular, the Pacific Ocean belt are both characterized by the active tectonic processes peculiar to geosynclinal systems. A comparatively rapid horizontal shifting of different parts of the Earth's crust at a speed of between one-quarter and three-quarters of an inch a year is taking place, according to geodetic measurements taken in Japan and Tadzhikistan. Also characteristic is intensive vertical movement. This occurs in the form of the uplifting of geoanticlines, accompanied by the sinking of neighbouring depressions and strong seismic activity. Disturbances of the isostatic equilibrium are concentrated in zones where the contrasting nature of the vertical movement is the most clearly manifest. The foci of earthquake shocks are not confined only to those fractures that rupture the Earth's crust but include also the deeply buried zones of folding that are associated with fractures. These zones are tilted from the deepwater trenches of the Indian and Pacific oceans toward the Asian continent at angles of from 20° to 70°. The sources of the earthquake shocks lie from 10 to 450 miles (16 to 720 kilometres) deep (about 150 miles deep in the Hindu Kush region). The characteristic movement originating from these foci is overthrust folding and upthrusting, and it indicates extreme compression of the Earth's crust. It is assumed that along the deep fractures there occurs an overthrusting of the island arcs onto the floor of the Pacific Ocean, an underthrusting of the Indian Platform under the Himalayas, and an analogous movement along the buckled edges of all the other outlying downwarps and deep-sea trenches. A considerable thickening of the Earth's crust is taking place in the high mountain regions of Central Asia and the Himalayas; here the Earth's crust is up to 40 or 50 miles thick, as opposed to 20 to 25 miles thick on the low plains and flatlands. This thickening may also be related to the lateral compression that is folding the Cenozoic strata.

Foci of earthquake shocks

THE PATTERN OF ASIA'S PALEOGEOGRAPHIC DEVELOPMENT

The Precambrian era. Although, on the crystalline shields of Asia, rocks may be found that are known to have been formed 3,000,000,000 years ago, an adequate description of the paleogeography of the continent—that is to say, of its geography in different eras and periods of geological time—can be given only for the last 1,000,000,000 years. At the beginning of this period, primitive forms of animal and plant life appeared, primarily as algae, when the vast marine basins of the Precambrian covered the

Asian seas of the Precambrian

Geological structure of Asia and (inset) New Guinea.

geosynclinal regions of the Urals, the Tien Shan, the Altai, the Western Sayans, and South China, as well as the southern parts of the Siberian Platform and the entire Iranian Massif. At that time continental sediments, including glacial deposits, accumulated over a considerable part of the Chinese Platform, while regions of erosion where detrital material was removed were represented by the shields.

The Cambrian and Ordovician periods. During the Early Cambrian Period, the seas began to transgress, covering almost all of the Siberian Platform, a large part of the Chi-

nese Platform, and the northern parts of the Arabian and Indian platforms. The transgression reached its maximum extent in the middle of the Cambrian Period, and the seas began to recede during the Late Cambrian, particularly from the Siberian Platform. But the sea advanced again during the Early Ordovician Period, covering almost the entire area of the Chinese Platform. The Cambrian and Ordovician deposits on the platforms were primarily limestone, but there were also red sand–clay rocks. Deposits of rock salt were associated with the sand–clay rocks in the southern part of the Siberian Platform, in the northwestern part of the Indian Platform, and in the eastern part of the Arabian Platform. Such deposits, and the nature of the marine animal life known from the fossil record, indicate that these three platforms were situated in a hot climatic zone; judging from the paleomagnetic data, the Equator then passed through the southern part of the Siberian Platform.

The Silurian Period. The seas receded during the course of the Silurian Period; red sediments and evaporites (sediments formed as a result of evaporation) continued to be laid down in the western half of the Siberian Platform and in other areas. At the end of the Silurian Period and in the course of the Devonian Period, all the platforms were uplifted and became dry land, except for some of the marginal areas and freshwater basins. Meanwhile, during the Cambrian, Ordovician, and Silurian periods, marine sandy-clayey and carbonate sediments, tuffs (rocks formed of compacted volcanic fragments), and lavas continued to be deposited in the geosynclines that separated the Russian, Siberian, and Chinese platforms from each other, as well as from Gondwanaland. The basins, which were separated from each other by islands, gradually dried up as the result of the Baikal and Caledonian folding; these included the basins of the Sayans, a considerable part of the Altai, the Khangarulsk, Tannu-Ola, central Kazakhstan, Kunlun, and South China. Large land areas were also subjected to volcanic activity.

The Devonian, Carboniferous, and Permian periods. The paleogeography of the Late Devonian, Carboniferous, and Permian periods is somewhat similar. In place of the present continent of Asia there existed, as before, the separate, small continental blocks of the three northern platforms, with the enormous continent of Gondwanaland to the south of them. Carbonaceous, detrital, and volcanic rocks continued to be deposited in the seas of the geosynclines separating these blocks. Calcareous sediments were also characteristic of the seas covering the massifs of Iran and Turkey, part of the Siberian Platform during the Devonian Period, and the southern half of the Chinese Platform in the Carboniferous Period. But the area of the marine geosynclinal basins was greatly reduced as a result of subsequent Hercynian folding. The basins between the Russian, Siberian, and Chinese platforms almost disappeared toward the middle of the Permian Period, at which time the area of present-day Mongolia and Western Siberia dried up. Coal-bearing strata that accumulated during the Late Carboniferous and Early Permian periods on the Indian, Chinese, and Siberian platforms and in the intermontane Kuznetsk, Minusinsk, and Tuva basins provide evidence that a humid climate then prevailed. In India these strata are underlain by glacial deposits of the Carboniferous Period.

Virtual disappearance of the landlocked basins

The Triassic and Jurassic periods. From the beginning of the Triassic Period, the northern platforms, together with the folded mountain ranges from the Paleozoic Era, formed a large continent embracing three-quarters of modern Eurasia. At the same time Gondwanaland split up into several gigantic blocks, separated from each other by a series of basins, some of which broadened and coalesced to form the Indian Ocean. The geosynclinal basins of the Mesozoic Era survived in the Tethys Sea, in Northeastern Siberia, in the Amur region, in Indochina, and in the belt of island arcs located in East and Southeast Asia. Shallow seas or coal-bearing freshwater basins and swamps, especially numerous during the Jurassic Period, at times covered the lowlands of Western Siberia, the Turan Plain, the Iranian Massif, and the Ordos Desert and Szechwan Basin.

The Cretaceous Period and the Cenozoic Era. Toward the middle of the Cretaceous Period, the areas of Mesozoic folding—located in what are now the Verkhoyansk Mountains, Sikhote-Alin, and Indochina—dried up, while in the course of the Late Cretaceous Period and the Early Cenozoic Era a rapid reduction in the size of the Tethys Sea took place. Finally, toward the beginning of the Early Cenozoic Era, folding in the Himalayas united India with the remainder of the continent, and Asia acquired approximately the outlines it has today. The island arcs and the basins of the marginal seas were shaped at the same time.

The joining of India and Asia

The Cretaceous Period and the Cenozoic Era were marked by mighty volcanic phenomena throughout East Asia. The Pliocene Epoch and Quaternary Period were times of vigorous tectonic movements that not only led to the uplifting of folded mountain ranges in the Alpine-Himalayan belt and in the island arcs but also rejuvenated the relief of the ancient folded mountains of the Urals, the Tien Shan, the Altai, the Sayans, and other ranges. Today an active belt of contemporary seismic activity stretches from the Hindu Kush through the Tien Shan, Mongolia, and the Baikal region to the Sea of Okhotsk.

Contemporary developments. The paleoclimatic and paleomagnetic data indicate that considerable shifting was taking place—in relation to the North and South poles as well as to each other—of those blocks from which the Eurasian continent was gradually formed. During the course of the early and middle periods of the Paleozoic Era, a considerable rapprochement (bringing together) of the Russian and Siberian platforms occurred. In the Late Carboniferous and Permian periods the Indian Platform was situated much closer to the South Pole and was partially subjected to glaciation. The Siberian Platform, on the other hand, was at that time located in the northern humid zone; in the Triassic Period the North Pole evidently lay at its northeastern edge. At the end of the Paleozoic Era the Tethys Sea was several times broader than the belt of folding that was formed within it in the course of the Mesozoic and Cenozoic eras. There are indications of the drift of the island arcs toward the shores of the Pacific Ocean and of the spreading of the bottoms of the marginal seas, such as the Sea of Japan and others, which originally developed and deepened in the middle of the Mesozoic Era.

(P.N.K.)

The land

RELIEF

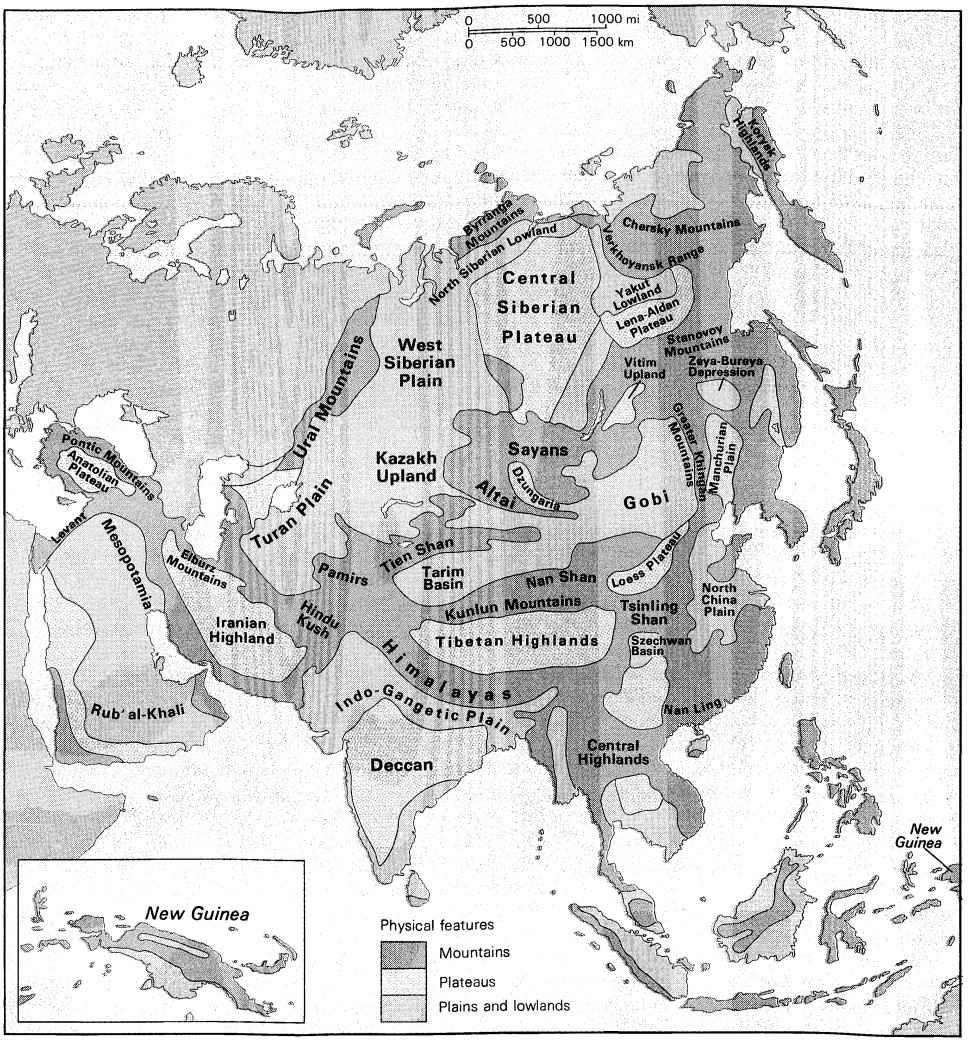
Asia is the highest of the continents and contains the sharpest relief. The highest peak in the world, Mt. Everest, which is 29,028 feet (8,848 metres) high; the lowest place on the Earth's land surface, the Dead Sea, which is approximately 1,300 feet (400 metres) below sea level; and the world's deepest continental trough, occupied by Lake Baikal, which is 5,315 feet (1,620 metres) deep and whose bottom lies at 4,250 feet (1,295 metres) below sea level, are all located in Asia.

Asia is also the most extensive of the continents. The farthest terminal points of the Asian mainland are Cape Chelyuskin in the Soviet Union (77° 43′ N) to the north; the tip of the Malay Peninsula, Cape Piai, or Bulus (1° 16′ N), to the south; Cape Baba in Turkey (26° 4′ E) to the west; and Cape Dezhnyova, or East Cape (169° 40′ W), also in the Soviet Union, overlooking the Bering Strait, to the east. The shores of Asia are washed by the Arctic Ocean on the north, the Pacific Ocean on the east, the Indian Ocean and the marginal seas of the Indian and Pacific oceans on the south, and by the seas of the Atlantic Ocean—the Mediterranean, the Aegean, the Sea of Marmara, the Black Sea, and the Azov Sea—as well as by the landlocked Caspian Sea, on the west.

The continent's extremities

Asia is separated from Australia to the southeast by the mingled waters of the Indian and Pacific oceans, and from North America on the northeast by the Bering Strait. The Isthmus of Suez unites Asia with Africa, and it is generally agreed that the Suez Canal forms the border between them.

The boundary between Asia and Europe is a historical–cultural concept that has changed more than once and is



Physiographic regions of Asia.

only as a matter of agreement tied to a specific borderline. The most convenient geographic boundary is a line drawn along the eastern slope of the Urals, then turning west along the Emba River to the Caspian Sea; west of the Caspian, the boundary follows the Manych River and the Kerch Strait to the Black Sea. From a statistical–economic point of view, the boundary is taken to run along those political–administrative borders of the republics and *oblasti* of the Soviet Union that most closely approximate this line, which is to say, along the eastern borders of the Komi Autonomous Soviet Socialist Republic; the *oblasti* of Arkhangelsk, Sverdlovsk, and Chelyabinsk; the western border of the Kazakh S.S.R.; and along the northern borders of the *kraya* (territories) of Stavropol and Krasnodar. Some authorities consider, however, that the boundary runs along the border between the Russian Soviet Federated Socialist Republic and the Transcaucasian republics (the Georgian S.S.R. and the Azerbaijan S.S.R.).

The area of mainland Asia, including the Caucasian isthmus, amounts to about 16,750,000 square miles (43,400,000 square kilometres), of which the peninsulas—Asia Minor to the west; the Arabian Peninsula, peninsular India, Indochina, and the Malay Peninsula to the south; Korea, Kamchatka, and the Chukotsk Peninsula to the east; Taymyr and Yamal to the north—make up about 3,000,000 square miles.

The islands—Cyprus, Sri Lanka, the Andamans, the Malay Archipelago, the Philippines, Hainan, Taiwan, the Ryukyus, Japan, the Kurils, Sakhalin, Wrangel Island, the New Siberian Islands, and Severnaya Zemlya—account for another 770,000 square miles.

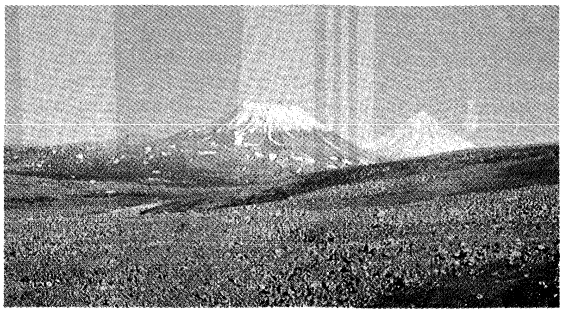
Asia's coastline is, variously, high and mountainous; low and alluvial; terraced as the result of the land being uplifted; or "drowned," where the land has subsided. The specific features of the coastline in some areas—especially in the east and southeast—are the result of active volcanism; of thermal abrasion (resulting from a combination of action by sea breakers and of thawing) by the subterranean fossilized ice (consisting of fossil ice, subsurface ice, and ice-formed rock), as in northeastern Siberia; and coral building, as in the south-southeastern area.

A characteristic of the surface of Asia is the predominance of mountains and plateaus, which form about three-quarters of the total area. The highest mountains and plateaus occur in Central Asia (Mongolia, Dzungaria, the Kashgar region, and Tibet) and Middle Asia (Turkmenia, Uzbekistan, Tadzhikistan, Kirgiziya, and Kazakhstan), which are also characterized by the vastness of their interior drainage basins.

Asia's mountain ranges

The mountain belts. The mountains are grouped in two large belts. One extends from the Chukchi Peninsula at the eastern extremity of Asia through the Kolyma Highlands, the Dzhugdzhur Mountains, and the Stanovoy Mountains to the mountains of Southern Siberia (the Stanovoy Uplands, the Sayans, and the Altai) and to the Tien Shan and Gissar-Alai. The Chersk and Verkhoyansk ranges are the western spurs of this belt.

The second belt, in the west, runs in a latitudinal direction and includes the West Asian highlands, the Pamirs, Karakoram, the Tibetan Highlands, and the Himalayas; it then turns to the south and southeast, running through the Arakan Mountains to the islands of the Malay Archipelago. A generally latitudinal branch springs from it in the Pamirs region and runs eastward through the Kunlun, Nan Shan, and Tsinling mountains. The western part of the second mountain belt consists, for a considerable distance, of two



Sopka (volcano) Krasheninnikova on the Kamchatka Peninsula in Northeastern Siberia.
Novosti Press Agency

series of mountain chains that converge in dense knots in the highlands of Armenia, in the Pamirs, and in the southeast of the Tibetan Highlands; the two chains then diverge to encompass the interior plateaus—the highlands of Asia Minor together with the Anatolian Plateau, and the Iranian and Tibetan highlands. At the margins of the highlands the mountain chains change direction abruptly in the congested mountain knots, but in the intervening areas they curve in flowing arcs.

Along the edges of the Central Asian plateaus extend the elongated mountain chains of the Greater Khingan, T'aihang Shan, and the Sino-Tibetan ranges. The Khingan–Bureya mountains (Bureya and Lesser Khingan mountains) demarcate the Zeya-Bureya Depression; the Manchurian–Korean mountains and the Sikhote-Alin ranges separate the plains of the Amur–Sungari, the Lake Khanka lowland, and the Manchurian Plain. The coastal ranges in the southeast consist of the Nan Ling (South China) and Annam mountains. In the east, the Koryak Highlands rise in the northern section of the Kamchatka–Koryak arc; in its southern portion they form the Central Range (Sredinny Khrebet) of Kamchatka. The marginal seas of the Pacific Ocean are bordered by the East Asian islands, which form the line of arcs that also extends partly onto the continent, running from Borneo to Taiwan, through the Ryukyu Islands to Korea, from Honshu to Sakhalin, and from the Kuril Islands to Kamchatka.

Detailed discussion of the Altai Mountains, Himalayas, Hindu Kush, Karakoram Range, Kunlun Mountains, Pamirs, and Tien Shan can be found at the end of this article.

The distribution of plains and lowlands

The plains and lowlands. Low plains occupy approximately a quarter of Asia, particularly the vast West Siberian and Turan plains of the interior. The remaining lowlands are distributed either in the maritime regions, such as Northern Siberia, the Yana-Indigirka lowland, Kolyma, and the Chinese coastal mainland, or in the piedmont depressions of Mesopotamia, the Indo-Gangetic Plain, and mainland Southeast Asia. In addition, there are the intermontane plains of Kashgaria, Dzungaria, Tsaidam, and Fergana and the plateaus of middle Siberia and the Gobi. The plateaus inside Tibet, the Tien Shan, and the Pamirs lie at altitudes as high as 12,000 feet. Many ranges in Middle and Central Asia reach elevations of 21,000 to 24,000 feet, but in other mountain chains and massifs they rarely exceed 12,000 feet. To the south of the zone of piedmont depressions, partially occupied by seas—the Arabian Sea, the Bay of Bengal, the Gulf of Oman, the Persian Gulf, and, in the Mediterranean, the Cyprus Basin—lie extensive tablelands and plateaus, including the Deccan Plateau in India and the Syrian–Arabian Plateau in the west; these are enclosed by marginal mountain ranges, such as the Western Ghāts in India; the Oman, Hadramawt, Yemen, and Hejaz mountains on the Arabian Peninsula; and the Lebanon and Anti-Lebanon mountains in the Levant. In Central Siberia are the isolated and uplifted Putorana Mountains and, situated to the north of them, the Byrrangas Mountains.

The islands. A large part of the islands of Asia are mountainous. The highlands of Sri Lanka rise to 8,279 feet (2,524 metres), Mt. Kinabalu in Malaysia reaches 13,451 feet (4,101 metres), Fuji-san on the Japanese island of Honshu has an altitude of 12,385 feet (3,776 metres), and many volcanoes of Sumatra, Java, and Mindanao reach 10,000 feet (3,000 metres). The Kuril–Kamchatka island arc, which extends onto the Kamchatka Peninsula, comprises the Vostochny (East) volcanic range. Especially high is the Klyuchevsk group of volcanoes, where the highest active volcano in Asia—Klyuchevskaya Sopka—rises 15,580 feet (4,750 metres).

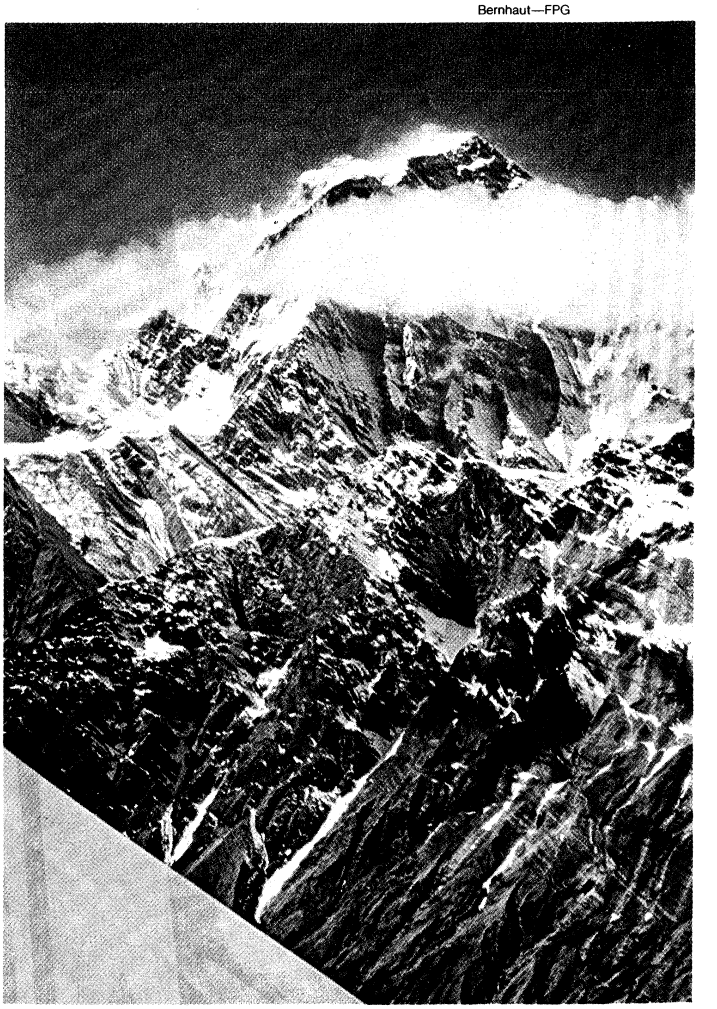
Forces shaping Asia's relief

Geologic and climatic influences. Mesozoic and Alpine foldings created boundaries between basic types of mountains over vast areas of Asia. The contemporary relief of Asia was molded primarily under the influences of: (1) ancient processes of planation (levelling); (2) larger vertical movements of the surface during the later Tertiary and Quaternary periods; and (3) severe erosive dissection of the edges of the uplifted highlands with the accompanying accumulation of alluvium in low-lying troughs, which were either settling downward or being uplifted more slowly than the adjoining heights.

The interior parts of the uplifted highlands, and the plateaus and tablelands of peninsular India, Arabia, Syria, and Eastern Siberia, which are relatively low-lying but composed of resistant rock, have largely preserved their ancient peneplaned (levelled) surfaces. Particularly spectacular uplifting occurred in Central Asia, where for the last 30,000,000 years the amplitude of this uplift of the mountain ranges of Tibet, the Pamirs, and the Himalayas has exceeded 13,000 feet. The eastern margin, meanwhile, underwent subsidences of up to 2,300 feet. Uplifting as a result of fractures at great depths, of which the Kopet-Dag and Ferghan mountains provide examples, and of folding over a large radius, of which examples may be seen in the Tien Shan and Gissar-Alai, played a large role.

Erosional dissection transformed many ancient plateaus into mountainous regions. Majestic gorges were carved

Bernhaut—FPG



Dhaulāgiri, in the Great Himalayas, Nepal, rising to a height of 26,810 feet.

into the highlands of the western Pamirs and southeastern Tibet; the Himalayas, the Kunluns, the Sayans, the Stanovoy Highlands, the Cherski Mountains, and the marginal ranges of the West Asian highlands were deeply cut by the rivers, creating deep superimposed gorges and canyons. In many areas, and especially in those regions with dry climates, erosion clearly exposed the structural forms, including rock layers of different erosional resistance.

Vast areas of Middle, Central, and East Asia, particularly in the Huang Ho Basin, are covered with loess (a loamy unstratified deposit formed by wind or by glacial meltwater deposition). There are broad expanses of badlands, eolian (wind-produced) relief, and karst topography (limestone terrain associated with vertical and underground drainage), and features associated with ancient glaciation.

olian
elief

The mantle of Quaternary glaciation embraced northwestern Asia only to 60° N. East of the Khatanga River, which flows from Siberia into the Arctic Ocean, only isolated glaciation of the mantle debris and of the mountains occurred because of the extremely dry climate that existed in the northeast even at that time. The high mountain regions experienced mainly mountain glaciation. There are traces of several periods during which the glaciers advanced—periods separated by warmer interglacial epochs. Glaciation continues in many of the mountainous areas and on the Severnaya Zemlya archipelago. Karakoram, the Pamirs, the Tien Shan, the Himalayas, and the eastern Hindu Kush are noted for the immensity of their contemporary glaciers.

There is an enormous area of permafrost in northern Asia that extends to lower latitudes than in any other part of the world. Little snowfall occurs, due to the aridity, and deep freezing of the soil takes place.

Several lowlands, primarily coastal plains, are covered with marine sediments as the result of recent advances of seas, such as the Caspian and the northern seas.

Volcanism added broad lava plateaus and chains of young volcanic cones to the relief of Asia. Ancient lavas and intrusions of magma, exposed by later erosion, cover the terraced plateaus of peninsular India and Central Siberia. Extensive zones of young volcanic relief and contemporary volcanism, however, are confined to the unstable arcs of the East Asian islands, together with Kamchatka, the Philippines, and the Greater and Lesser Sunda Islands.

olcanic
ones

Recent volcanism is also characteristic of the West Asian highlands, the Caucasus, Mongolia, the Manchurian–Korean mountains, and the Syrian-Arabian Plateau. In historic times eruptions have also occurred in the interior of the continent in the Lesser Khingan Mountains and the Anyuy highlands.

The regions of Asia. In geographical literature the practice of dividing Asia into large regions, each grouping together a number of countries, is common. These divisions usually consist of North Asia, including Siberia and the northeastern edges of the continent; East Asia, including the continental part of the southern Soviet Far East, the East Asian islands, Korea, and eastern and northeastern China; Central Asia, including the Tibetan Highlands, Dzungaria and Kashgaria in the Sinkiang Uighur Autonomous Region, Inner Mongolia, the Gobi, and the Sino-Tibetan ranges; Middle Asia, including the Turanian Plain, the Pamirs, the Gissar-Alai, and the Tien Shan; South Asia, including the Philippine and the Malay archipelagoes, Indochina and the Indian Peninsula, the Indo-Gangetic Plain, and the Himalayas; and West Asia, including the West Asian highlands (Asia Minor, Armenia, and Iran), the Levant, and the Arabian Peninsula. On occasion, the Philippines, the Malay Archipelago, and the Indochina peninsula, instead of being considered as part of South Asia, are grouped separately as Southeast Asia; the Arabian Peninsula and the Levant are also sometimes grouped together separately as Southwest Asia.

North Asia. The North Asia region includes platform plains, plateaus, and folded mountain ranges. Frost weathering and permafrost have influenced relief.

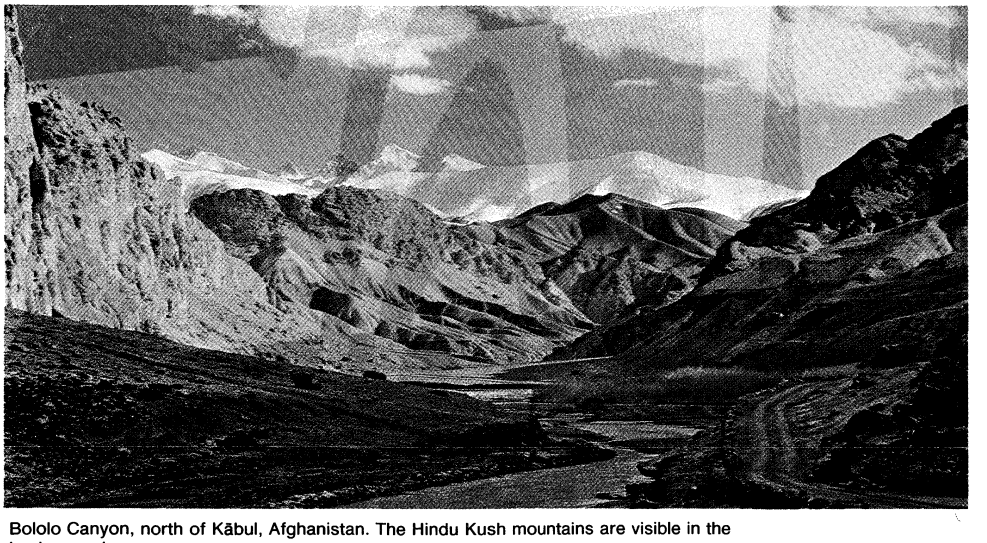
In Northeast Siberia are found faulted and folded mountains of moderate height, such as the Verkhoyansk, Chersk, and Okhotsk-Chaun mountain arcs, formed of Mesozoic structures rejuvenated by neo-tectonic uplifting; the Koryak Mountains, formed of Cenozoic structures, are also in this region. Volcanic activity took place in these areas during the Cenozoic Era. Some plateaus are found in the areas of the ancient massifs, such as the Kolyma massif. Traces of several former centres of mountain glaciers remain, as well as traces of lowland originally covered by the sea, such as the New Siberian Islands. The Aldan Plateau—an ancient peneplain resting on the underlying platform that sometimes outcrops on the surface as the Aldan Shield—is located in the region. Traces of ancient glaciation are also to be distinguished.

Northeast Siberia

The North Siberian plains consist of the Middle Siberian Tableland and the Lena–Vilyuy lowlands, which are platform plateaus and stratified plains that were uplifted in the Cenozoic Era. They are composed of terraced and dissected mesas with exposed horizontal volcanic intrusions; plains formed from uplifted Precambrian blocks; a young uplifted mesa, dissected at the edges and partly covered with traprock (Putorana Mountains); and the peripheral North Siberian lowland, covered with its original marine deposits.

The West Siberian Plain is stratified and is composed of Early Cenozoic sediments deposited over thicknesses of Mesozoic material, in addition to folded bedrock that is Hercynian in the west and Caledonian in the east. The northern part was earlier subjected to several periods of glaciation; in the south the predominant deposits are those laid down by glacier streams, as well as alluvial deposits.

Harrison Forman



Bololo Canyon, north of Kābul, Afghanistan. The Hindu Kush mountains are visible in the background.

In the northern part of the region are the mountains and islands of the Asian Arctic. The archipelago of Severnaya Zemlya is formed of fragments of fractured Paleozoic folded structures. Throughout the region vigorous contemporary glaciation has occurred.

East Asia. Mountains and plains are characteristic of the northern part of continental East Asia. The main features in the northern region include the Khingan–Burein mountains; the Sikhote-Alin ranges of Khabarovsk and Primorsky *kraya* (territories) in the Soviet Union; the Manchurian–Korean highlands running along North Korea's border with China; the East Korean range of the Korean Peninsula; the Zeya-Bureya Depression of Amur *oblast* in the Soviet Union; the Liao Ho in Liaoning Province, China; the Manchurian Plain and the North China Plain; and the Amur and Sungari rivers and the Lake Khanka lowlands. Most of these features were formed by folding, faulting, or broad zonal subsidence. The mountains are separated by alluvial lowlands in areas where recent subsidence has occurred.

The mountains of southeastern China were formed from Precambrian and Caledonian remnants of the Chinese Platform by folding and faulting that occurred during the Mesozoic and Cenozoic eras. The mountain ranges are numerous, are of low or moderate altitude, and occupy most of the surface area, leaving only small, irregular-shaped plains.

The East Asian islands

The islands off the coast of East Asia and the Kamchatka Peninsula are related formations. The Ryukyu Islands, Japan, Sakhalin, and the Kuril Islands are fragments, uplifted in varying degrees, of the Ryukyu–Korean, Honshu–Sakhalin, and Kuril–Kamchatka mountain-island arcs. Dating from the Mesozoic and Cenozoic eras, these arcs have complex knots at their junctions, represented by the topography of Kyushu and Hokkaido. The mountains are of low or moderate height and are formed of folded and faulted blocks; some volcanic mountains and small alluvial lowlands are also to be found.

Kamchatka is a mountainous peninsula, formed from fragments of the Kamchatka–Koryakskaya and Kurilo–Kamchatskaya arcs, which occur in parallel ranges. The young folds enclose rigid ancient structures. Cenozoic (including contemporary) volcanism is pronounced. Vast plains exist that are composed of alluvia with volcanic ashes.

Central Asia and South Siberia. Central Asia consists of mountains, plateaus, and tablelands formed from fragments of the Siberian and Chinese platforms, peripherally surrounded by a folded area formed in the Paleozoic and Mesozoic eras.

The mountains of Southern Siberia and Mongolia were formed by renewed uplift of old faulted and folded blocks; ranges are separated by intermontane troughs. The alpine mountains—the Altai, the Mongolian Altai, and the Sayano-Tuvan and Stanovoy highlands—are particularly noticeable. They have clearly defined features resulting from ancient glaciation; contemporary glaciation is also very active.

The Central Asian plains and tablelands include the Takla Makan Desert, the Gobi, and the Ordos Desert. Relief features vary from surfaces levelled by erosion in the Mesozoic and Cenozoic eras to stratified plateaus with low mountains, eroded plateaus on which loess had accumulated, and vast sandy deserts covered with windborne alluvium and lacustrine deposits.

Alpine Asia—sometimes known as High Asia—includes the Pamirs and the eastern Hindu Kush, the Kunlun Mountains, the Tien Shan, the Gissar-Alai Mountains, the Tibetan Highlands, the Karakoram Range, and the Himalayas.

The Pamirs and the eastern Hindu Kush are sharply uplifted mountains dissected into ridges and gorges in the west. There is thick glacial cover; alpine deserts occur on the plateaus.

The Kunlun Mountains, the Tien Shan, and the Gissar-Alai Mountains belong to an alpine region that was formed from folded structures of Paleozoic age. There are glaciers of impressive size centred in this alpine region.

The Tibetan Highlands

The Tibetan Highlands represent a fractured alpine zone in which Mesozoic and Cenozoic structures that surround an older mass in the centre have experienced more recent uplifting. Some of the highlands are covered with detrital desert; elsewhere in this region, alpine highlands are dissected by erosion or are covered with glaciers.

The Karakoram Range and the Himalayas include the highest mountains in the world; they were formed by uplifting that took place in a zone of Cenozoic folds and Mesozoic partially folded areas containing outcrops of the ancient bedrock. Contemporary glaciation in the region is vigorous.

South Asia. South Asia, in the limited sense of the term, consists of peninsular India and Sri Lanka and the Indo-Gangetic Plain.

Peninsular India and Sri Lanka are formed of platform plateaus and tablelands uplifted in the Mesozoic and Cenozoic eras and subjected to humid-climate erosion ever since. Tablelands with uplifted margins and terraced and dissected plateaus with lava mantles or intrusions may be distinguished.

The Indo-Gangetic Plain is formed from the combined alluvial plains of the Indus, Ganges, and Brahmaputra rivers, which lie in a deep marginal depression running north of and parallel to the main range of the Himalayas. It is an area of immature subsidence, in which thick accumulations of earlier marine sediments and later continental deposits that washed down from the mountains have been transformed into sandy deserts in the western arid region.

Southeast Asia. Southeast Asia comprises the Indochina Peninsula and the islands and peninsulas to the southeast of the Asian continent.

Indochina

The mainland consists of the western mountain area and

Brian Blake—Rapho/Photo Researchers

The Ch'iang-t'ang plain of northern Tibet, bordered by the Kunlun Mountains.

the central and eastern mountains and plains. The western mountain area of Burma is a zone of Cenozoic folding. Mountains of medium altitude are formed of folded blocks that decrease in size and altitude to the south; the valleys are alluvial and broaden out to the south. The central and eastern region of Thailand and of Vietnam is characterized by mountains of low and moderate height that have been moderately fractured. The region is one of Mesozoic structures surrounding an ancient mass known as the Cambodian saucer, with which are associated plateaus and lowlands filled with accumulated alluvial deposits.

Archipelagoes border the southeastern margin of Asia, consisting mainly of island arcs with which peninsulas are associated. The island arcs are bordered by very deep oceanic trenches. These arcs are characteristically very unstable and are volcanically active.

The Indian Ocean arcs—the island chains of Sumatra, Java, and the Lesser Sunda Islands—consist of fragments of alpine folds formed from materials of different ages. Cenozoic and contemporary volcanic activity is manifest, and volcanic mountains as well as alluvial lowlands may be distinguished.

Borneo and the Malay Peninsula are formed from fractured continental land situated at the junction of the Alpine–Himalayan and East Asiatic geosynclinal regions; contemporary volcanism is absent. The mountains are composed of folded and faulted blocks; the lowlands are alluvial.

The Pacific Ocean island arcs, including Celebes, the Moluccas, the Philippine Islands, and Taiwan, are fragments of folded alpine structures that were built up by volcanic products during the Cenozoic Era. Volcanic activity and the building of coral reefs continue. Mountain areas of moderate height, volcanic ranges, alluvial lowlands, and coral reef islets may all be distinguished in the region.

Middle Asia. Middle Asia includes the plains and hills lying between the Caspian Sea to the west and Lake Balkhash to the east.

The area between the Caspian and Lake Balkhash is composed of flat plains on continental platforms of folded Paleozoic and Mesozoic bedrock. Individual uplifted portions form low rounded hills (*melkosopochniks*) in the Kazakh region; low mountains on the Mangyshlak and Krasnovodsk peninsulas of the Caspian Sea; and mesas (isolated hills with level summits and steeply sloping sides) in areas of earlier marine sedimentation, such as the Ustyurt Plateau and the Kara-Kum Desert. Thick accumulations of alluvium have been transported by the wind, forming sandy deserts in the south. Original marine and lacustrine sediments adjoin the shores of the Caspian and Aral seas and Lake Balkhash.

West Asia. West Asia includes the highlands of Asia Minor, and the Armenian and Iranian highlands.

The highlands of Asia Minor—the Pontic mountain system that parallels the Black Sea, and the Taurus and Anatolian tablelands—are areas of severe fragmentation, heightened erosional dissection, and isolated occurrences of volcanism.

The Armenian Highlands

The Armenian Highlands, which include the Little Caucasus and the Kurd mountains, are severely fragmented. Recent uplifting, in the form of a knot of mountain arcs, took place during a period of vigorous volcanism that occurred in the Cenozoic Era.

The Iranian Highlands represent a combination of mountain arcs (in the north, the Elburz and Turkmen-Khorasan mountains, the Selseleh-ye Safīd Kūh, and the western Hindu Kush; in the south, the Zagros, Makran, Soleymān, and Kirthar mountains), together with the tablelands of the interior, and the Central Iranian, Eastern Iranian, and Central Afghanistan mountains. There are isolated Cenozoic volcanoes, a predominance of accumulated remnants resulting from ancient erosion, and saline and sandy deserts in the depressions and on the tablelands.

Southwest Asia. Southwest Asia, like much of southern Asia, is made up of an ancient platform—the northern fragments of Gondwanaland—in which sloping plains occur in the marginal downwarps. Its principal components are the Arabian Peninsula and Mesopotamia.

The Arabian Peninsula is a tilted platform, highest along the Red Sea, on which the stratified plains have undergone erosion under arid conditions. Block tablelands with uplifted margins, Cenozoic lava plateaus, stratified plains, and cuestas (long, low ridges with a steep face on one side and a long gentle slope on the other) may all be distinguished. Ancient marine sands and alluvia, resulting from previous subsidence and sedimentation, now take the form of sandy deserts.

Mesopotamia consists of the Tigris and Euphrates floodplains and of the deltas from Baghdad to the Persian Gulf. The original lowland is covered with late Cenozoic and Quaternary sedimentation; the elevated plain, on the other hand, has been dissected by erosion and denudation under the continental conditions prevailing in the Late Cenozoic Era.

CLIMATE

Air masses and wind patterns. The enormous expanse of Asia and the abundance of mountain barriers and inland depressions have resulted in great differences in existing conditions of solar radiation, atmospheric circulation, and climate as a whole. A continental climate, associated with large landmasses and characterized by an extreme annual range of temperature, prevails over a large part of Asia. Air reaching Asia from the Atlantic Ocean, after passing over Europe or Africa, has had time to be transformed into continental air. As a result of the prevalent easterly movement of the air masses, as well as the isolating effect of the marginal mountain ranges, the influence of sea air from the Pacific Ocean extends only to the eastern edge of Asia. From the north, Arctic air has unimpeded access into the continent. In the south, tropical and equatorial air masses predominate, but their penetration to the centre of Asia is restricted by the ridges of the latitudinal belt of highlands; in the winter months—November through March—such penetration is further impeded by the density of the cold air masses over the interior.

The continental climate

The contrast between the strong heating of the landmasses in the summer months from May to September and the chilling in winter produces sharp seasonal variations in the atmospheric circulation and also enhances the role of local centres of atmospheric activity. Winter chilling of the Asian landmass develops a persistent high-pressure winter anticyclone over Siberia, Mongolia, and Tibet, which is normally centred southwest of Lake Baikal. Within the zone of the anticyclone there is relatively little strong air movement in protected basins and lowlands, but strong winds may affect the higher mountains and passes. The anticyclone is fed by subsiding upper air, by bursts of Arctic air flowing in from the north, and by the persistent westerly air drift that accompanies the gusty cyclonic low-pressure cells operating within the Northern Hemisphere cyclonic storm system. Drifts of cold, dry air move eastward and southward out of the continent, affecting eastern and southern Asia during the winter. Only a few of the winter cyclonic lows moving eastward out of Europe carry clear across Asia, but they do bring greater periodic change in weather in Western Siberia than is typical in Central Siberia. The zone of lowest temperature—the so-called cold pole—is found in the northeast, near Verkhoyansk, where temperatures as low as −90° F (−68° C) are recorded. The outward drift of winter air creates a sharp temperature anomaly on Asia's eastern margin, where the climate is colder than the characteristic average for each given latitude. But episodic intrusions of oceanic air from the east and southeast moderate this anomaly, so that temperatures are not as severe here as in the centre of the anticyclone.

The polar front

The zone where the temperate and tropical air masses are in contact—called the polar front—shifts southward in winter. This movement is caused by a displacement, in the same direction, of the entire system of atmospheric circulation—a displacement resulting from the powerful climatic influence exerted by the chilled continent. The winter rainy season in the southern parts of the West Asian highlands, which is characteristic of the Mediterranean climate, is associated with this southerly movement of the polar front. In the more northerly areas of West and Middle Asia, the effect of cyclonic action is particularly

strong in the spring, causing the maximum in annual precipitation to occur at this season. In summer, the polar front shifts northward, causing cyclonic rains in the mountains of Southern Siberia. In West, Middle, and Central Asia, a hot, dry, dusty, continental tropical wind blows at this time. Over the basin of the Indus River the heating creates a low-pressure area, known as the South Asian (or Iranian) low. The southern monsoon (a rain-bearing wind) advances along its southern edge, bringing copious rainfall to peninsular India, the southern Himalayas, and mainland Southeast Asia. Farther to the west the hot, dry air of North Africa and the katabatic (downward) current of air from Europe, blowing from the northwest, sweep in the direction of this low-pressure area. The aridity of the desert-tropical climate of Arabia and Pakistan is related to this phenomenon.

In eastern Asia the Pacific Ocean polar front creates atmospheric disturbances during the summer. From the warm sectors of cyclones moving westward through this region, the warm and moist summer monsoon blows toward the continent. Becoming chilled as it passes over cold ocean currents, this air brings fogs and drizzling rains. To the south of 38° N, where the warm Kuroshio (Japan) current approaches the coast of Japan, the summer monsoon brings protracted rains and high humidity; together with high temperatures, this creates a hothouse atmosphere.

The summer period over China is a time of variable air movement out of the South Pacific. If that drift is strong and the summer continental low-pressure zone is marked, a strong summer monsoon may carry moisture well into Mongolia. If neither the drift nor the continental low is strong, the China summer monsoon may fail, falter over eastern China, or cause irregular weather patterns that may threaten China proper with crop failure.

Typhoons and monsoons

Tropical cyclones, or typhoons, occur along the East Asian weather fronts throughout the year but are most severe during the autumn months. These typhoons are accompanied by very strong winds and torrential rains so heavy that the maximum precipitation from the typhoons locally may exceed the total amounts received during the normal summer monsoons.

In winter the Pacific Ocean polar front is driven back to tropical latitudes by a steady drift of cold, dry Siberian air. On the East Asian islands the effect of the winter continental monsoon is tempered by the surrounding seas. In passing over them, it becomes warmed and saturated with moisture; then waters the northwestern slopes of the island arcs. Occasionally, however, strong bursts of cold air carry cold spells as far south as Hong Kong and Manila.

In winter, continental tropical air prevails in subequatorial Asia; in summer it is replaced by equatorial ocean air. The winter season's dry and warm winds, directed toward the equatorial low-pressure axis, are analogous to trade winds but simultaneously act as the South Asian continental monsoon. The dry spring that follows changes abruptly and dramatically into the rainy summer with the onset of the monsoon. The summer monsoon brings enormous amounts of rain (up to about 25 inches [635 millimetres] in a month). Over the areas of Asia close to the Equator—southern Sri Lanka, Malaysia, and the Greater Sunda Islands—equatorial air prevails continuously, accompanied by even temperatures and abundant rainfall at all seasons. The Lesser Sunda Islands have a subequatorial monsoon climate; their wet and dry seasons are regulated by the calendar rhythm of the Southern Hemisphere, which is characterized by a wet summer from November to February and a dry winter from June to October.

The influence of topography. Differences between the climatic conditions of the various regions of Asia are determined to a considerable degree by topography. Different altitudinal climatic zones are most clearly defined on the southern slopes of the Himalayas, where they vary from the subequatorial and tropical climates of the foothills, at the lowest levels, to the snowy climate of the peaks, at the highest altitudes. The degree of exposure also plays a large role—the different orientation of the opposite slopes of the ridges in relation to compass directions and to the prevailing winds. The sunny southern slopes differ from the shady northern ones, and windward slopes exposed to moist ocean winds differ from leeward slopes, which, lying in the wind (and rain) shadow, are necessarily drier. In addition to the physical isolation of the leeward slopes from the moisture-laden winds, the foehn effect is also found here. This occurs when a strong wind traverses a mountain range and is deflected downward as a warm, dry, gusty, erratic wind. Contrasts of climate resulting from exposure are manifested clearly in the Himalayas, the Elburz Mountains, Japan, Taiwan, the Philippines, the Tien Shan, the Transbaikalia, and many other places.

The foehn effect

The isolating barrier effect of the relief on the climate appears clearly in the West Asian highlands and in Central Asia. In these regions the surrounding mountains isolate the tablelands of the interior from the moisture-laden winds. The massiveness of the interior highlands is also a significant factor; it favours the formation of local anticyclones over them during the cold months of the year.

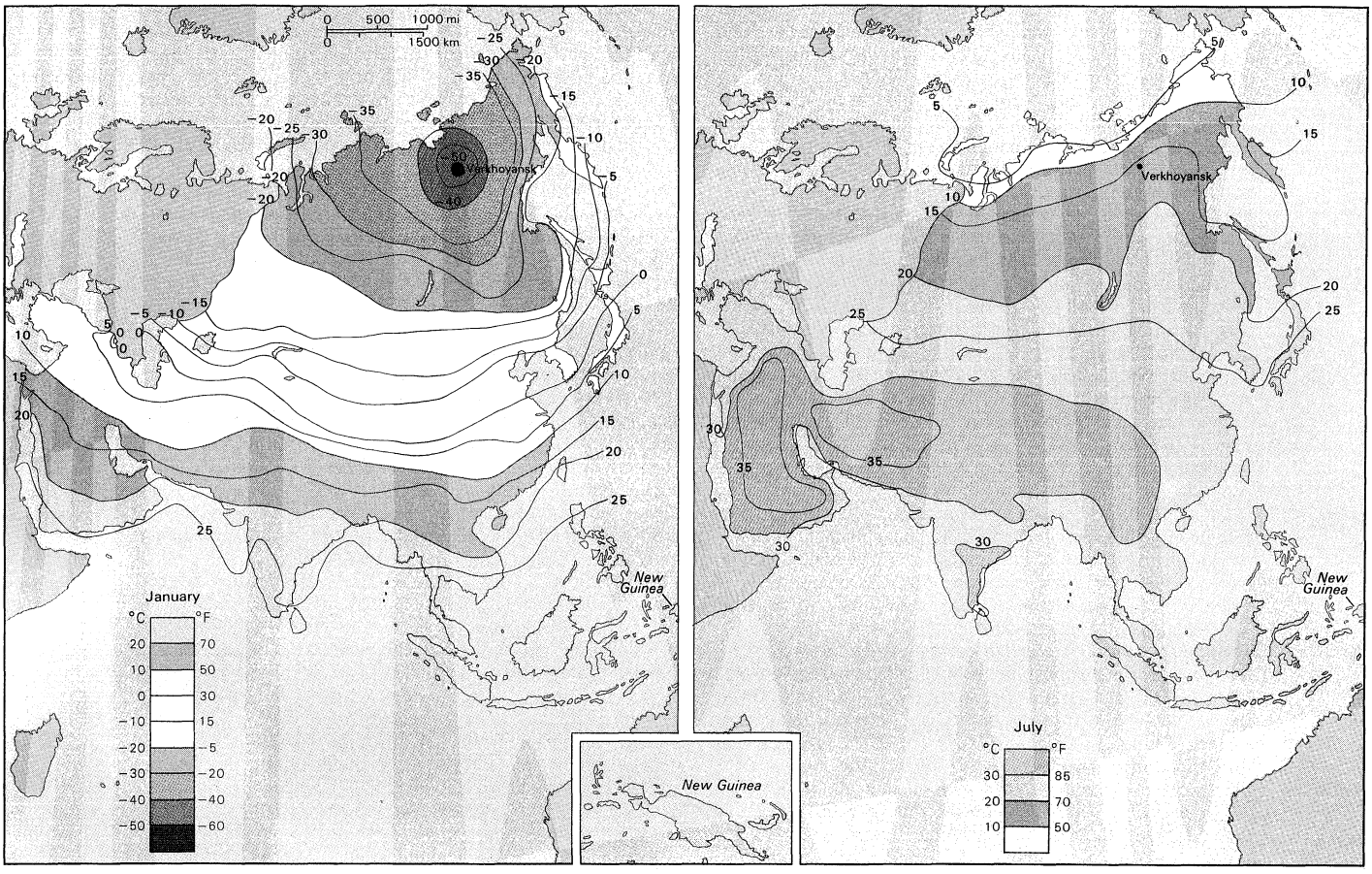
During the winter season some of the cyclonic storms that move eastward through the Mediterranean Basin are deflected south of the Tibetan Highlands, crossing northern India and southwestern China and then turning northeastward to return to the northern cyclic path. Such storms do not often bring winter rain, but they create short periods of cloudy, cool, or gusty weather and bring snow to the higher mountain ranges.

Temperature. The average January temperature over a considerable part of Siberia is below −4° F (−20° C), and in the Verkhoyansk region it reaches −58° F (−50° C). Along the coastal areas, the proximity of the Pacific Ocean moderates the temperatures to from 23° F to 5° F (−5° C to −15° C). The January isotherm (a line connecting points of equal temperature) of 32° F (0° C) passes through Samarkand, Peking, and the island of Honshu. An isotherm of 68° F (20° C) is traced along the Tropic of Cancer and one of 77° F (25° C) along the Equator. In July, when the average temperature is 86° F (30° C), the maximum temperatures are found in West Asia and in the Thar and Takla Makan deserts. The 68° F (20° C) isotherm moves as far as 55° to 60° N, but near the cool Pacific Ocean it bends to the south. Along the northern coasts of Asia the average temperature in July is below 50° F (10° C), which is typical for a tundra climate. The greatest amplitude in annual temperature range occurs near the "cold pole," which has surprisingly warm summers; the annual range may exceed 175° F (97° C).

Rainfall. Annual rainfall in the equatorial belt is approximately 80 inches (2,000 millimetres); it is 80 to 120 inches and more (300 to 500 inches in places) on the windward maritime slopes in South and East Asia. In Cherrapunji (Meghalaya) 900 inches of rain fell in seven months in 1891. Precipitation is less than 40 inches on the lee slopes of the subequatorial regions. In the subtropical and temperate monsoon climates there is adequate rainfall, amounting to about 24 to 40 inches. Precipitation is less than 10 inches in Eastern Siberia and averages six to eight inches (but may be less than four inches in some places) in the deserts of West, Middle, and Central Asia.

Climatic regions. The distribution pattern of rainfall throughout the year is varied. Relatively uniform moisture is characteristic of the Asian equatorial zone. Maximum summer precipitation and minimum winter precipitation are the rule in the subequatorial zones and in other regions with monsoon climates, as well as in those areas where there is summer movement of the fronts—the polar front in the mountains of Southern Siberia and the Arctic front in the subarctic regions. Wet winters and dry summers are typical of the Mediterranean climatic region in West Asia, where precipitation is associated with the winter activity of the polar front. This polar front activity, accompanied by maximum precipitation, occurs in the spring in the interior parts of the West Asian highlands. Summer and winter precipitation merges in some parts of Asia. In the Kolkhida climate the summer rains—brought by the northwesterly Atlantic air currents—merge with the cyclonic Mediterranean winter rains. In some areas of Japan and eastern China there is uniform precipitation when, in addition to the summer monsoon, the winter monsoon brings moisture.

As the aggregate result of these various meteorological



Average temperatures for January and July in degrees Celsius for Asia.

Types of climate

patterns, the following types of climate may be distinguished in Asia: the tundra climate (associated with the cold, treeless plains of the Arctic lowlands of Asia); the cold, sharply continental climate of Eastern Siberia; the cold, moderately humid Western Siberian climate; the humid, subtropical Kolkhida climate; the desert climate of the temperate zone; the Mediterranean subtropical climate of the western edge of West Asia; the subtropical desert climate; the mountain-steppe highland subtropical climate of West and Central Asia; the alpine desert climate; the climate of the Eastern Pamirs, Karakoram Mountains, and Tibetan Highlands; the climate of the tropical deserts; the temperate monsoon climate of the Soviet part of the Far East, and northern parts of Japan and East China; the subtropical monsoon climate of Southern Japan and of Southeastern China; the subequatorial monsoon climate of South Asia, eastern Java, and the Lesser Sunda Islands; and the equatorial climate of the Greater Sunda Islands.

Many climatic variants can be distinguished that are associated with such local topographical features as the degree of exposure of the slopes, the protective effect of the mountains, and altitudinal zonality, with temperatures dropping as the altitude increases. Low temperatures, however, are also found in low hollows where cold air stagnates or on coasts where air is chilled by cold ocean currents. The mountain climates, evidently, represent variants of those climates that are determined by latitude. All the various features of the types of climate mentioned exert a strong influence on other natural conditions, as well as on the landscape as a whole.

Urban climate. Distinctive variations of climatic characteristics result from the cultural and economic activities of human society. One example of this is provided by the microclimates associated with the cities and with large industrial complexes. The emission by the cities of quantities of dust and gases produces alterations of temperatures and changes in wind patterns. Such conditions are characteristic, for example, of Tokyo and the industrial region of northern Kyushu in Japan, of Calcutta and the industrial area of the northeastern part of peninsular India, and of the industrial regions of the Kuznetsk Basin in the Soviet Union.

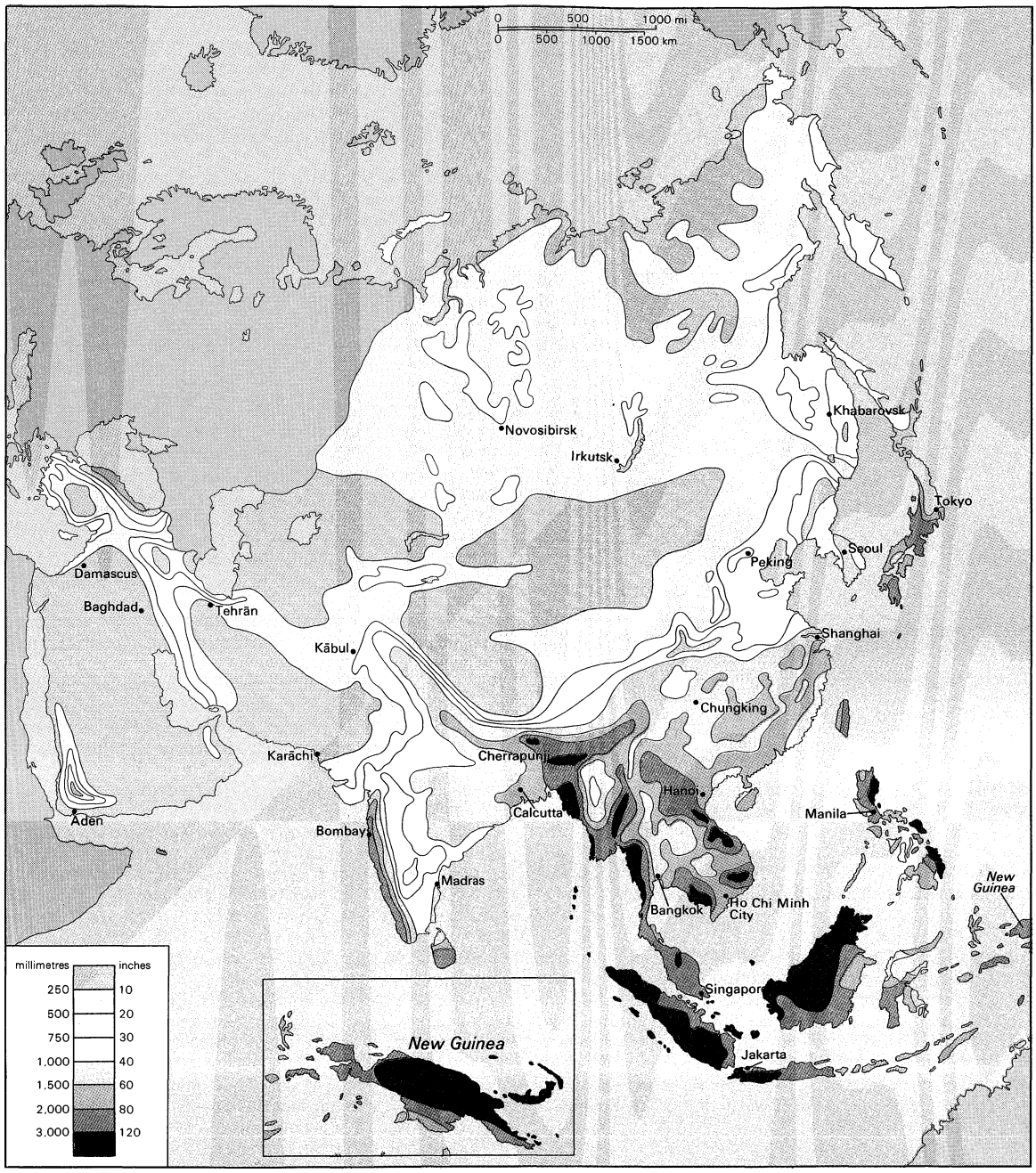
DRAINAGE

Rivers. Asia is a land of great rivers. The Ob, the Irtysh, the Yenisey with the Angara, the Lena (with the waters of the Aldan and the Vilyuy), the Yana, the Indigirka, and the Kolyma rivers all flow into the Arctic Ocean. Among rivers draining into the Pacific Ocean are the Anadyr, the Amur (combined with the Sungari and the Ussuri), the Huang Ho, the Yangtze, the Hsi, the Red River (Song Hong), the Mekong, and the Chao Phraya. The Salween, the Irrawaddy, the Brahmaputra, the Ganges, the Godāvari, the Krishna, and the Indus flow into the Indian Ocean, as also does the Shatt al-Arab, which is the confluence of the Tigris and Euphrates rivers. Only small mountain rivers flow from Asia into the Caspian, the Sea of Azov, the Black Sea, and the Mediterranean. The Amu Darya, the Syrdarya, the I-li, the Chu, the Tarim, the Helmand, and the Tedzhen rivers empty into vast interior basins. Some of these rivers end in lakes, some end in dry deltas in the sands or salt marshes, and some flow into oases, where all the water is used to irrigate fields or else it evaporates.

Rivers of the Pacific

All the Siberian rivers freeze over in the winter, and some freeze to the bottom. In spring widespread flooding occurs. These rivers are important communication routes, being used by boats during the summer and as roads for sleighs in winter; they also teem with fish.

In the dry regions, where drainage is landlocked, the only large rivers are temporary ones fed by snow and glacier water in the mountains; they reach their peak water levels in summer. Rivers that are not fed by mountain runoff have little water; their levels vary sharply, and periodically or occasionally they dry up. The rivers of the monsoon climate regions reach their maximum volume in summer and are utilized for irrigating the rice fields. The Asian rivers in the vicinity of the Mediterranean that are not fed



Average annual precipitation for Asia.

by mountain snows grow shallow in summer and sometimes even dry up. In the equatorial regions, however, the rivers are perennially full of water.

Lakes. The many lakes of Asia vary considerably in size and origin. The largest of them—the Caspian and Aral seas—are the remains of larger seas. Lakes Baikal, Issyk-Kul, and Hövsgöl (Khubsugul), the Dead Sea, and others lie in tectonic depressions. The basins of Lakes Van, Sevan, and Urmia are, furthermore, encircled by lava, and Lake Telets was gouged out by ancient glaciation. A number of lakes were formed as the result of landslides (Lake Sarez in the Pamirs); karst processes (the lakes of the western Taurus, in Turkey); or the formation of lava dams (Tsin Bokhu in Northeast China, and several lakes in the Kuril Islands). In the volcanic regions of the eastern Asian islands, in the Philippines, and in the Malay Archipelago, lakes have formed in craters and calderas. The subarctic has a particularly large number of lakes; in addition to lakes formed as a result of permafrost and subsidence, there are also ancient glacial moraine lakes. Many lagunal lakes occur along low coastlines.

The lakes in the internal drainage basins—such as Koko Nor, Tuz, and others—are usually saline. Lake Balkhash has fresh water in the west and brackish water in the east. Lakes through which rivers flow are freshwater and regulate the flow of the rivers that issue from them or flow into them; examples of these are Lake Baikal, associated with the Angara River; Lake Khanka (the Sungacha and Ussuri rivers); Tung-t'ing Hu and P'o-yang Hu (the Yangtze River); and Tonle Sap (the Mekong). Large reservoirs have also been created by the construction of hydroelectric stations.

Subterranean water. In arid regions, subterranean water is often the only source of water supply. Large accumulations are known to exist in artesian basins and beneath the dipping plains at the foot of mountains; these are associated with the extensive oases of Middle Asia, Kashgaria, and many other regions.

Detailed discussion of Asia's drainage systems and waterways, arranged generally from south to north, can be found at the end of this article. The discussion for Southwest Asia includes the Caspian Sea, Persian Gulf, Red Sea,

The Ganges River winds through the fertile plain east of Delhi, in northern India.
Harrison Forman

and Tigris and Euphrates rivers. Treatment of South Asia includes the Arabian Sea, Bay of Bengal, Brahmaputra River, Ganges River, Indus River, and Irrawaddy River. For East and Southeast Asia, the Amur River, China Sea, Huang Ho, Sea of Japan, Mekong Delta, Yangtze River, and Yellow Sea are covered. Discussion of Siberia's waterways includes the Lena, Ob, and Yenisey rivers.

SOILS

The soils of Asia are distinctly marked by the combined effects of climate, topography, hydrology, organic nature, and the economic activities of man. The horizontal zonality of the climate, the drainage conditions, the existing plant and animal life, and agriculture are each related to the considerable meridional extension of Asia, which is accompanied, of course, by a horizontal zonality of the soil cover, which is especially clearly defined in the plains of the continental sector.

The Arctic zone. In the Arctic, where glacial and Arctic deserts predominate, the processes of soil building are manifested only in rudimentary form. The soils are saturated and are low in humus. The subarctic north of Asia is occupied by a timberless zone of tundra vegetation. Beneath the tundras, specifically tundra-type soils are formed, which are characterized by poor drainage (associated with the proximity of permafrost) and only a short period in which the decomposition of organic substances is possible. This results in the accumulation of undecomposed organic residues in the form of particles of peat. The poor drainage creates an oxygen-free medium in which the bluish substance known as gley is formed. Thus, peaty-gley soils are most characteristic of the tundra. There are widespread occurrences of movement by mud glaciers; heaving of the ground because of frost; settling or caving in of the ground from thawing; and the formation of stone rings around central areas of debris in bouldery regions.

The forest tundra. Farther south stretches the transitional belt of the forest tundra, where tundra and sparse forest alternate with regularity. Tundra soils alternate with the soils of the taiga (the cold, swampy forested region to the south of the tundra, characterized by very low temperatures). The soils below the frozen taiga are called cryogenic (*i.e.,* having very low temperatures). In the mountainous regions the peaty-gley soils are replaced by mountain tundra and weakly developed, often embryonic soils of detritus and stony fragments.

The forest zone. The forest zone occupies the largest part of the temperate zone. Characteristic of soil formation in the forest zone is the leaching process. The forest leaves and needles that fall, together with dead remains of the sparse grass cover, are subjected to decomposition by organic acids in the litter of the forest floor; the duration of the summer season and the amount of precipitation are sufficient for complete decomposition of the soluble soil components; the soil solutions transport them and leach them into deeper soil horizons (layers). The undecomposed quartz grains remain in the upper horizon, which is therefore infertile; this layer resembles light-gray ashes, which is the reason soils of this type are called podzols ("under ashes"). The various subzones of the forest zone are subjected to different degrees of leaching. A dense, rusty-brown horizon of wash-down (deposition in an underlying layer of soil) underlies the podzolic portion of the soil profile (vertical section of the soil); its colour is related to the accumulation of iron and aluminum oxides. This layer, called orstein, or iron pan, is impervious to water and contributes to the self-swamping of the taiga forests. East of the Yenisey River, where the forest zone for its entire breadth is in the grip of permafrost, soil drainage (and consequently the leaching process) is made more difficult; the transfer of substances is complicated by freezing and thawing, and therefore the typical podzols are replaced by specific cryogenic taiga soils. Marshes and bog-type soils are widely distributed over a considerable part of the taiga subzones.

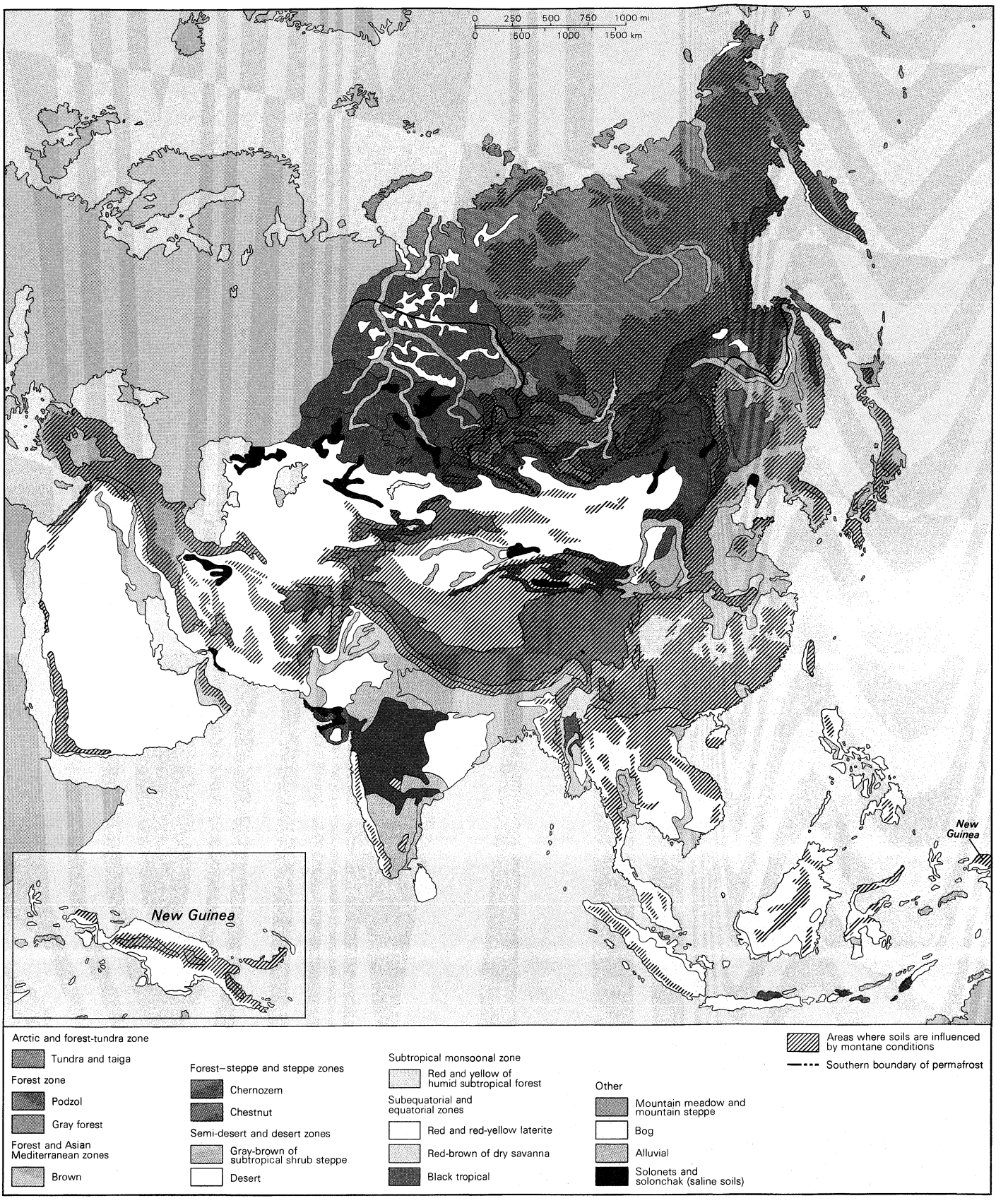
The leaching process

The deciduous forest subzones of Asia form two distinct areas. In Western Siberia there are small-leafed, primarily birch or aspen, forests on gray forest soils. They are more gray in colour than the podzols because of the greater amount of organic substances—such as tree leaves and a more abundant grass cover—feeding these soils. This explains their higher content of humus, as well as their greater fertility. The second section of the deciduous forest subzone has survived in the Far East, stretching from the Lesser Khingan Mountains in the north to the Japanese island of Honshu; in this subzone abundant warmth and moisture intensify chemical weathering, and iron oxides accumulate even in the surface soil horizons. In this manner brown forest soils, known as forest burozems, are formed.

The forest-steppe and steppe. Soil cover in the forest-steppe region is formed when the ratio of precipitation to evaporation is in equilibrium and as the leaching process of the wet season alternates with the upward flow of the soil solutions during the dry period. Under these conditions, with abundant organic material resulting from the dense vegetation, intensive accumulation of humus takes place in the soil, and dark-coloured soils are formed that are the most fertile in all of Asia; known as chernozems, they are the most fertile as well as the thickest of the forest-steppe and mixed grass subzones. Characteristic of the wooded-meadow plains of the Amur Basin (the "Amur prairies") are meadow soils that are dark, semiboggy, and often composed of blue gley. In the drier steppes, where vegetation is sparse, the amount of humus is reduced and the content of unleached mineral salts is increased; transport of the dissolved salts to the surface by the upward flow of soil solutions is also intensified. Associated with this process is a bleaching and salinization of the soil. The drier steppes thus form a transitional zone from the shallow southern chernozems to the chestnut soils. Broad expanses of the forest-steppe and steppe are under cultivation and serve as rich granaries for the cultivation of grain crops. Severe wind erosion occurs during the hot, dry seasons. In many areas impoverishment of the soil has also developed as the result of surface washout and gully erosion, despite the preventive efforts that have been made.

Soils rich in humus

Semidesert and desert. Through inner Kazakhstan and Mongolia stretches a zone of semidesert, and through Middle Asia, Dzungaria, Takla Makan, and Inner Mongolia a belt of temperate zone deserts. A belt of subtropical deserts extends through the Levant, the Iranian highlands, and the



Soils of Asia.

southern edge of Middle Asia. Beneath the semideserts, with their mosaic of desert and arid-steppe vegetation, light-chestnut and light-brown semidesert soils form; these are low in humus but contain an abundance of strongly alkaline soil. Beneath the deserts, where the supply of organic substances, as well as the humus content, is extremely low, gray-brown soils form in the temperate zone; gray desert soils (sierozems) form in the arid subtropics. Here there is a great deal of saline soil, and agriculture is possible only with the use of irrigation, which is feasible in the infrequent oases, where specific cultivated types of sierozems have formed.

Only in western Asia is the tropical desert zone clearly defined. Broad expanses of this area are characterized by embryonic soils and desert crusts, as well as by blowing sands.

Iron-bearing soils

The Asian Mediterranean. In the maritime areas of the Asiatic Mediterranean—Asia Minor and the Levant—xerophytic vegetation (vegetation structurally adapted to exist with very little water) of the Mediterranean scrub-woodland types, known as maquis (evergreen), shiblyak (an association of bushes and scrub characteristic of the Balkan Peninsula in Europe), and frigana (low-growing, prickly, and cushion-like bushes), is prevalent. The predominant soils under such vegetation are brown; they have accumulated iron as a result of the intense chemical weathering during the wet Mediterranean winter and of the upward flow of soil solutions during the dry summer. Frigana vegetation, characterized by thorn bushes, is widely represented in the West Asian semidesert highlands. Here soils have developed that are transitional between the brown soils and the sierozems.

Detailed discussion of the Arabian Desert, the Gobi, the Kara-Kum, and the Takla Makan can be found at the end of this article.

The subtropical monsoonal regions. Typical of the monsoon subtropics are the evergreen forests of the southern portion of the Korean Peninsula, of southwestern Japan, and of southeastern China. Intensive chemical weathering during the simultaneously warm and wet summer monsoon season results—as it also does in the more southerly torrid zones—in the decomposition and carrying away from the soil of many minerals, the accumulation of residual iron and aluminum oxides, and the consequent predominance of red and yellow soils as well as of podzolized soils. Agriculture, with the irrigation of rice fields, is especially widespread on the alluvial soils of the plains, which have been cultivated continuously by farmers for thousands of years. Terracing of the slopes is a widely applied practice.

The subequatorial and equatorial regions. The subequatorial zones of Asia are covered by savannas (grassy parklands) and dry-tropical deciduous forests, primarily situated in the rain shadow on the leeward slopes, and by wet-tropical evergreen forests on the rainy windward slopes facing the sea. Intensive leaching followed by evaporation is characteristic of these soils. Under the wet tropical forests, red-yellow lateritic (leached and hardened iron-bearing) soils predominate; beneath the savannas and dry tropical forests there are red lateritic soils that change, with increasing aridity, to red-brown and desert brown soils. Beneath the dry savannas of peninsular India there are unique black soils called regurs that are thought to be developed from basalt country rock.

Charles Lenars—Atlas

Tropical vegetation surrounding fertile rice paddies on the island of Java, Indonesia. The active volcano Gunung Merapi rises above the clouds in the background.

In the equatorial zone (southern Malaysia and the Greater Sunda Islands), typical tropical rain forests have developed. In southwestern Sri Lanka and in Java they have been almost entirely replaced by an agricultural landscape in which mountain slopes and hills are covered with plantations of tea, coconut palms, and rubber trees. The soils are lateritic and are red-yellow or brick-red, with marginal degrees of laterization.

In the valleys of the subequatorial and equatorial zones, alluvial soils predominate; they have been developed by thousands of years of cultivation and irrigation of the rice fields. Artificial terracing of the slopes is practiced on a very large scale in the mountainous regions, both for purposes of irrigation and in order to prevent soil erosion.

Underdeveloped mountain soils

The mountains. In the mountains, zones of different soil types are found at different altitudes. As a rule they are skeletal, underdeveloped soils, clearly reflecting the differences in rock structure and origin and in the degree of exposure of the slopes. The boundaries of the vertical zones become higher from north to south; the number of zones increases. Mountain soils also correspond to the different vegetation zones occurring at different altitudes. Under the mountain forests of the northern portion of the temperate zone are mountain-podzolic soils. Mountain variants of the gray forest soils and taiga-cryogenic soils have also developed; above the tree line they are replaced by mountain-tundra soils. Mountain chernozem and mountain chestnut soils develop beneath the mountain steppes in the southern part of the temperate zone; brown mountain-forest soils are found in the wetter regions; beneath the alpine and subalpine meadows the soils are of the mountain-meadow type. Mountain red podzolized, yellow earth, and other lateritic soils are predominant in the wet areas of the lower latitudes; in the dry areas mountain brown and gray-brown soils, as well as mountain gray soils, occur. The alpine steppe and desert soils of Central Asia have a number of distinctive characteristics.

The correlation of the vertical soil zones and of the landscape zones varies with the whole spectrum of vertical zonality. A zone of forest, followed higher up by meadows, with snow cover at the highest altitudes, is characteristic of the western maritime regions. On lower slopes in the western Caucasus, for example, broad-leafed mountain forests occur on brown mountain-forest soils; above these are coniferous forests on mountain podzolic soils, followed by stunted trees, followed in turn by subalpine and alpine meadows on mountain-meadow soils, while near the highest ridges perennial snow and glaciers are found. Associations of desert, steppe, meadowland, and snow zones are widespread in the interior of Asia and sometimes include mountain-forest zones. Thus, characteristic of the Tien Shan, for example, is the predominance of mountain-desert and semidesert landscapes, which occur in association with gray-brown and brown mountain soils in the foothills of the ranges, while higher up are mountain steppes associated with mountain chestnut soils and mountain chernozems. Under parts of the mountain forest-steppe and the mountain forests, the soils are podzolized. Above the stunted forest zone in the maritime sector, mountain-meadow soils occur beneath the meadows, but here, too, a distinctive snowy type of landscape occurs in the vicinity of the ridges.

Typical of the mountains of Eastern Siberia are the taiga-tundra spectra that occur in vertical zones; thus, mountain taiga on taiga-cryogenic soils is followed by a zone of dwarfed trees, followed by mountain tundra, and then finally by bald peaks.

In eastern Asia, the subalpine and alpine meadow zones with mountain-meadow soils usually disappear; instead, mountain-forest landscape extends as far up as the vicinity of the crests and is succeeded only by a zone of stunted trees. The spectra of the alpine regions of South Asia (the

Himalayas) are distinguished by the most complex variety of vegetation and soil types. (Y.K.Y.)

PLANT LIFE

In Asia an immense range of vegetation is found, resulting from the continent's wide diversity of latitude, altitude, and climate. Natural conditions, however, are not entirely responsible for the associations of trees, plants, and grasses of Asia; natural landscapes have been transformed by 80 centuries of farming.

The geographic pattern of vegetation. *North Asia.* The natural landscape has been least affected by man in sparsely populated North Asia. Vast plains, continentality, and the nearness of the Arctic Ocean explain the presence here of a zone of tundra—cold, treeless plains with permanently frozen subsoil—similar to that found in the western Soviet Union and in Canada. In more flourishing parts the tundra has a discontinuous covering of lichens, mosses, sedges, rushes, some grasses, cushions of bilberries, and dwarf trees of willow and birch; in the far north, lichens grow on favourable hillsides. Thanks to the greater number of hours of daylight during the summer solstice in June, when the Arctic Circle receives the same amount of light energy as the tropics, the tundra at this season is covered with bright flowers. Nevertheless, climate conditions are extreme; in Severnaya Zemlya, along the Arctic coast, thawing begins in May and frosts begin in August, although in some years frosts may occur at night throughout the short summer. The soil never thaws below a depth of two or three feet; consequently, hollows are badly drained and turn into peat bogs. Windy conditions speed up evaporation, and the frozen soil cannot absorb water to compensate for this, so that surface drought often allows wind erosion and the transport of sediments deposited by annual riverine floods.

The tundra

The tundra belt extends still farther south on higher ground. In the Arctic Urals tundra begins at about 3,000 feet, but at latitude 53° N it begins at 4,250 feet. Tundra extends over large areas of the Chersk, Verkhoyansk, and the Kamchatka mountains.

The taiga

The taiga zone—a belt of coniferous forest—begins south of the tundra, after a transitional zone of "wooded tundra" and forest galleries, found along streams between the tundra-covered watersheds. Taiga, although essentially coniferous, is mixed with hardy deciduous trees such as aspen and birch; there are sections of grass and shrub steppe in the drier zones. Larches account for 37 percent of the Siberian forest, which covers 2,700,000 square miles; pines cover 24 percent and spruce 4 percent. The geographic distribution of particular types of vegetation is determined chiefly by climate. Spruce, for example, unable to survive temperatures below −36° F (−38° C), is not found east of the Yenisey River. The taiga has a thin undergrowth of cranberries and bilberries, and there are numerous extensive peat bogs.

Sovfoto

Taiga vegetation growing on the Siberian Lowland of south central Siberia.

In Soviet Asia broadleaf deciduous forest does not extend eastward beyond the Yenisey River, where it gives way to the coniferous forests of Central Siberia, reappearing in Eastern Siberia near the Okhotsk Sea; here poplars, birches, and alders are numerous, as well as various conifers and larches. Forests around the Ussuri River include maples, ashes, walnut, elms, and lindens, in addition to species already mentioned. In the direction of China, as described below, the landscape becomes transitional.

South of the Siberian forest, the zone of prairie (continuous herbaceous cover) is not uninterrupted; forest frequently gives way to steppe (discontinuous cover).

Tibet, which is chiefly dry and cold, has a scattered vegetation of halophilic bushes (bushes flourishing in a salty environment) and Artemisia's tufts.

The Far East. In the Far East, the monsoon climate brings hot and rainy summers, giving rise to a great variety of temperate and tropical vegetation. China has the most varied vegetation of any country in the world, with about 15,000 species, excluding mushrooms and mosses. Far Eastern forests are fascinating to botanists because of the variety of their plant life; many trees have large, bright evergreen leaves, and there is a dense undergrowth with abundant creepers.

Japan has 68 percent of its area under forest, whereas China is almost entirely deforested, although sizable tracts remain untouched in the remote, rugged regions and many small areas have been reforested. The reason for this is that Japanese scenery was traditionally respected, and strict forestry regulations were severely enforced. The best examples of Far Eastern forest are found in Japan; for example, in the Kii Peninsula. Conifers are the principal species used for reforestation in Japan's policy of restoring its forests in order to meet the industrial need for wood.

The "true" Chinese forest

North of the Yangtze River, much of China was covered by primeval deciduous forest, most of which has been removed through farming. South of the Yangtze the "true" Chinese forest was prevalent before 1800. A wild growth of trees and shrubs survives, however, throughout the cultivated areas, and park-like tree growth and stands of bamboo are widespread. The "true" forest included 60 different genera of tall trees, including—among the temperate genera—oak and maple, linden, chestnut, hornbeam, and a species of hickory. Tropical genera included magnolia, the tulip tree, the camphor tree, the Spanish cedar, liquid ambar (a tropical tree of China), catalpa, and lianas (vines). A variety of conifers of both hemispheres was also to be found, and in the mountains of eastern Szechwan there grew a rare and ancient Chinese conifer, the metasequoia. Palm trees are found throughout South China and South Korea as well as in the southern half of Japan; many varieties of bamboo are also found in these regions.

The Peking government is proceeding energetically with a program of reforestation. The new forests, however, consisting largely of pines, do not resemble the primeval Chinese forest.

South Asia. The wettest parts of peninsular India, such as the Western Ghāts, and of Southeast Asia have magnificent forests noteworthy for the variety of their plant life. The culturally controlled forests of Java and Sumatra alone include over 3,000 species of trees. The variety of tropical vegetation is accentuated by the diversity of influences, such as that of altitude upon climate. Temperate pines are found in Sumatra and in the Philippines, where eucalyptus also thrives; oaks occur in the mountains of New Guinea; and austral *Podocarpus* (evergreen trees with a pulpy fruit) in the eastern Himalayas. In the seasonal monsoonal climatic zone of central Indonesia, Thailand, Burma, and southern India, the teak forest thrives as an open park-like cover with little underbrush.

A notable feature of South Asian vegetation is the Dipterocarpacea family (yielding aromatic oils and resins), which

Vegetation zones of Asia.

is here represented by more than 500 species. Mangrove thrives in muddy deltas along the South Asian coastline. In the southernmost areas, the bogs on the land-side edges of the mangrove swamps abound with the semi-aquatic nipa palm, the leaf fronds of which are widely used for thatching.

Hevea brasiliensis, the most successful rubber plant, is found in tropical Southeast Asia, where it was introduced from South America in the 1870s; it is particularly important in plantations in Malaysia and Indonesia.

Primeval evergreen rain forest remains in a few parts of South Asia. Secondary forest, which develops where temporary clearings have been made in the original forest, covers a much larger area. In drier tropical areas, secondary forest is deciduous; where the dry season is particularly long, park forest, with trees spaced at wide intervals, is found, as in the "sal" (East Indian hardwood) forests of India. Extensive fires in such areas have resulted in a herbaceous landscape, as in the cogonales (areas of coarse tall grasses, used for thatching) of the Philippines.

Mountain vegetation

In the higher mountains of Southeast Asia the cooler humid tropical climate gives rise to deciduous and coniferous temperate forest at altitudes of between about 4,250 feet and 10,000 feet. Above this level, low forests of plants, mostly shrubs of the heath family, are often found. Diverse types of trees grow in the mountain forests of the region. The Arakan Mountains of Burma, for example, are covered with a thick mantle of little bamboos. In the eastern Himalayas sal is intermingled with *Castanopsis* (a small genus of nut-bearing trees) and pines. Above these are found forests of shrubs and trees of the laurel family and, higher still, oaks and conifers; between about 10,000 feet and 13,000 feet, forests of firs occur. The central Himalayas present strikingly beautiful landscapes in the following upward succession: dry sal forest; pine forest; cedars, spruces, pines, and oaks; firs, birches, and tall rhododendrons; above 13,000 feet, bushes of rhododendrons, together with junipers; above 16,000 feet, perpetual snows.

West Asia. In western Asia naturally wild vegetation no longer occurs in clearly defined zones but is dispersed in small areas. The region is predominantly arid: desertlike depressions such as the Kyzylkum in the Kazakh and Uzbek regions of the Soviet Union and the Rub' al-Khali (Empty Quarter) of the Arabian Peninsula contrast with the moist, well-forested mountains that lie between them. Three climatic zones, however, characterize western Asia: a continental climate in the northern regions; a dry zone, except where northerly winds bring moisture to the mountains, to the south; and a Mediterranean climate along the western edges.

A few examples of the variety of vegetation associated with these climatic zones may be cited. In the valleys of the dead rivers of the Kara-Kum Desert grows a strange tree, the saxaul, which is oddly shaped, gnarled, and leafless; between the galleries of saxauls the desert is interspersed at very wide intervals with bushes and tufts of grass. A fringe of steppe covers the area between the Fertile Crescent (which sweeps in an arc from the Tigris–Euphrates Valley to the Mediterranean) and the north and west of the Syrian Desert. With more than 2,000 species of plants—more than in the whole of the Sahara—the borders of the latter desert are noteworthy for their floral variety. The moist northern slopes of the Pontic Mountains in northern Turkey are covered by magnificent forests of beeches and conifers, with an undergrowth of tall cherrylowels, hollies, and creepers. This type of forest is also found in Georgia and on the northern slopes of the Elburz Mountains in Iran. Along the Mediterranean border of Asia the vegetation is similar to that in other parts of the Mediterranean region: holm oak (an evergreen oak), Aleppo pine (characteristic of the city of Aleppo), cistus, mastic tree (which yields mastic, used as a chewing gum), and other species are found in landscapes of thick underbrush and open scrubland.

"Pontic" forest

Man and vegetation. *Vegetation in traditional civilization.* Asia's indigenous vegetation has provided many edible products, such as stone fruits, citrus fruits, bananas, mangoes, soybeans, and tea; building materials, such as wood, bamboo, and thatch; cotton and straw for clothing; bamboo, widely used in the making of utensils; and the bark of the paper mulberry, used for making bark cloth and paper. In addition, silkworms are fed upon mulberry leaves; lacquer is made from *Rhus vernicifera* (lacquer tree); and a multitude of other items are obtained from plants, including a styptic for stopping hemorrhage, an anti-asthmatic agent made from mahuang (a plant yielding ephedrine), and a fine fibre, extensively used in weaving, derived from ramie, a plant of the nettle family.

Similarly, the forested areas of Southeast Asia provide the sparse population with a wide variety of products: firewood; timber for construction; foodstuffs from a variety of trees, plants, and fungi, including sago, taro (a plant with an edible, starchy, tuberous rootstock), and mushroom; scented resins from sandalwood and eaglewood (an East Indian tree with soft resinous wood); and dyes from a variety of tubers. Probably the irrigated cultivation of taro and rice began in Southeast Asia.

Wheat is indigenous to the humid hill margins of western Asia; it still grows spontaneously in the hilly fringes of Asia Minor. Cherry, peach, and pistachio trees, as well as vines, were domesticated in the mountains of western Asia. The large acorn cups from the forests of *Quercus aegilops*—a species of oak, in western Asia Minor—are valued for the abundant tannin they contain.

Commercial forestry. China and the Indian subcontinent, with their enormous populations, are today poor in timber resources. The best resources of timber for building and for paper manufacture are in Siberia. Japan produces about 1,600,000,000 cubic feet (45,000,000 cubic metres) of timber annually from its controlled forests. The tropical and equatorial forests of South Asia are difficult to exploit; because of the diversity of species found in these forests, commercially valuable trees are mixed with a majority that is of no economic value. Some deciduous forests, however, are of commercial interest since they are more homogeneous and consist of good quality trees. Included in this category are the forests of teak, sal, and ironwood. Such forests are important in the Burmese economy, timber being among the more important exports. The countries of Indochina, as well as Thailand, Malaysia, and the Philippines, are covered with extensive forests, but only a small fraction of their timber resources are commercially exploited. As in Burma, teak is the most valuable wood, but much of the remaining natural timber is of low quality and not easily marketable. It is, nevertheless, increasingly being recognized that the forests of these countries form potentially valuable resources. Malaysia, in particular, is encouraging the development of forestry, and although overshadowed by rubber and tin, forestry is one of the most important Malaysian industries. In western Asia, Turkey has launched a program to increase production from its forests and thus reduce lumber imports. (P.Gu./Ed.)

Timber

ANIMAL LIFE

The Himalayas, stretching from east to west, form a barrier largely preventing the movement of animals southward or northward. Thus, Asia north of the Himalayas, with parts of western Asia and most of the Far East, belongs to the Palearctic (literally "ancient Arctic") zoogeographical region. Asia south of the Himalayas is called the Oriental, or Indian, Region. The boundary dividing these zones east and west of the Himalayas is not well marked, however, as there the mountain chains often have a north–south trend facilitating migration of animals between them.

The Palearctic Region. A distinction can be made between the animal life of the tundra in the north and that of the adjacent taiga farther south; the taiga in turn merges into the steppes, which have their own distinctive forms of animal life. The tundra subsoil is frozen throughout the year; hence, burrowing animals cannot live there, and, as the tundra is partly free from snow only during the short summer, conditions for life are poor. Most animals, including reindeer, Arctic hare, Arctic fox, and wolf, live here in summer only and migrate in autumn, but the lemmings (small rodents of circumpolar distribution) stay, feeding on the herbage buried beneath the snow. Hibernation is impossible, for the short summer does not allow the necessary accumulation of food reserve in the body.

During the summer, birds are numerous but they also desert the tundra in winter, except for such birds as the willow grouse and the ptarmigan, which live in tunnels in the snow, feeding on berries and leaves. Many species of waders, such as the gray plover, the sanderling, and several kinds of sandpipers, migrate to the tundra and breed there in the summers, feeding principally on the mosquitoes in the wet areas. Mosquitoes are also the staple food of passerine birds (true perching birds), such as the snow bunting and the Lapland bunting. Gyrfalcons (a subgenus of large Arctic falcons), rough-legged buzzards, and skuas (large, dark-coloured rapacious birds of northern seas) prey on these smaller birds and on lemmings. Several kinds of geese and ducks, Arctic tern, and three species of divers occupy the moist parts.

The taiga fauna

The taiga fauna is much richer than that of the tundra. The taiga is the haunt of brown bear, wolf, glutton (a kind of wolverine), otter, ermine, sable, lynx, elk, forest reindeer, hare, and several kinds of squirrel. Birds include various kinds of grouse and woodpecker, pine grosbeak, crossbill, siskin, redpoll, red-spotted bluethroat, rubythroat, redwing, fieldfare (a medium-sized thrush), nutcracker, Siberian jay, and others. Wading birds include the terek sandpiper, which frequents marshes and pools.

The rivers of North Asia are inhabited by many common freshwater fishes and by several kinds of sturgeons, including the sterlet. Lake Baikal has a peculiar animal life, including many native species of sponges, worms, and crustaceans and a native species of seal.

The animal life of the steppes differs as much from that of the taiga as that of the tundra. It includes many burrowing rodents, such as jerboas, marmots, and piping hares, and, among larger animals, large numbers of antelope. The steppes were the original home of the northern cattle (*Bos taurus*), the horse, and probably the Bactrian (two-humped) camel; it is doubtful that any of these remain as truly wild animals. Typical birds are bustards, quails, sand grouse, and the red-legged hobby. Hoopoes and rollers are common locally, and bee eaters and the common sand martin nest along riverbanks. Waterfowl inhabit the reed beds of the great rivers, as do locusts, which migrate in almost unbelievably large numbers, devastating crops.

Wild sheep and goats live in the mountains and on the plateau to the north of the Himalayas. Tibet is the home of the wild yak, which is in great danger of extermination, although the domesticated yak survives.

The eastern part of the region, comprising Northeast and eastern China, has several peculiar kinds of deer. The Siberian tiger, originally native to southeastern Siberia, Manchuria, and Korea, has spread southward through eastern China into all of Southeast Asia and northern India. The giant panda inhabits the lower mountain margin of China bordering Tibet; the lesser panda is a Himalayan animal. Associated with the wastelands of the higher Himalayas is a legendary form of higher animal life—the yeti, or "Abominable Snowman." Some species of animals are peculiar to Japan, including a monkey related to the tailless Barbary ape of Gibraltar.

The large rivers of China have a rich fish life, among which *Psephurus gladius* (paddlefish) from the Yangtze and Huang Ho is of interest, as it is one of the two survivors of an otherwise extinct family, the other remnant of which is the paddlefish of North America. Another freshwater animal is the giant salamander, found in Japanese waters. Southeast Asia and southern China are the home of most of the carp family, from which the various forms of goldfish are derived.

The animal life of Asia Minor is much like that of other Mediterranean countries, but that of Israel, Syria, and Arabia also includes an African element, such as a species of coney and—in Lake Tiberias and the Dead Sea—fishes of the African genus *Tilapia* (the Nile perch). The donkey may have been domesticated in Southwest Asia, and the dromedary (one-humped) camel was originally native to the drier portions of Transcaspia.

The Oriental Region. The greater part of the Oriental Region is tropical. The northwestern part is dry and partly desert, so that animal life is chiefly confined to the forms related to those of the dry parts of the Ethiopian and Palearctic regions. Elsewhere, monkeys are common. Apes are found only in tropical rain forests—being represented by gibbons in Assam, Burma, the southeastern peninsula, and the Greater Sunda Islands—whereas the orangutan is restricted to Sumatra and Borneo, where it is in danger of extermination.

Large game animals

The Asiatic distribution of the African lion is now confined to the Gir Forest of the Kāthiāwār Peninsula in India, where it is protected, but a few specimens may still occur in southeast Iran. The tiger is now found from the Himalayas to Sumatra, Java, and Bali, but not in Borneo or Sri Lanka. Panthers range all over the region, except in Sumatra. Civets and mongooses are numerous. Among badgers, the ratel lives in the hilly districts of peninsular India and is even to be seen as far west as Israel. Jackals are plentiful in India; the striped hyena is confined to drier parts. Both are absent from the east.

Flying and ordinary squirrels are common in woodlands; the gaur (a large wild ox) is found in India and Burma, the banteng (the Malayan wild ox) in Burma and south to Borneo and Java, but not in Sumatra.

The most common antelope is the black buck, found in open brush-covered wild areas and cultivated plains all over India, except on the Malabār Coast; the nilgai, or blue bull, and the chousingha (a four-horned antelope of northern India) occupy hilly regions south of the Himalayas. Species of deer include musk deer in the pine zone of Kashmir, Nepal, and Sikkim; sambar deer practically over the whole region; and barking deer ranging northward into southernmost China.

Chevrotains (very small, hornless, deerlike ruminants) are typical, and wild pigs are also widely distributed. The Indian one-horned rhinoceros is protected and confined to Nepal and Assam; the Javanese rhinoceros is now restricted to Malaya, southern Sumatra, and western Java; the two-horned rhinoceros ranges from Burma to Sumatra. The Indian tapir lives in dense forests in southern Tenasserim, Malaya, and Sumatra. The Indian elephant is found throughout the region. Scaly anteaters, or pangolins—also found in Africa—are characteristic. The tropical cattle (*Bos indicus*), known as Zebu or Brahman cattle and recognizable by its shoulder humps, was domesticated in India, as was the water buffalo, which is now distributed from Egypt to central China and the Philippines.

Birds

Game birds are important. The Indian peacock can be seen throughout India, whereas another species of peacock (*Pavo muticus*) is restricted to Java. Numerous species of pheasants live in the forests of Burma, Thailand, Indochina, Malaya, Sumatra, and Borneo. Jungle fowl are unique to the Oriental Region and are the source of all domesticated chickens. Pigeons occur in great variety, but the number of species of parrots is small compared with other tropical regions. Water and wood kingfishers are represented by many species. Hornbills show their greatest development in the Oriental Region. The Indian hoopoe is common in India but is only a migratory bird in the southeastern part of the region. Among cuckoos the brain-fever bird—an Asian hawk cuckoo that takes its name from the suggested effect of its repetitious cry—is well known. Eagles, osprey, falcons, hawks, kites, and buzzards all occur; in the western part vultures are numerous and are found even in towns. The forests are inhabited by many species of woodpeckers. The barbets (loud-voiced tropical birds) are characteristic, the best known being the coppersmith bird. Bee eaters and rollers are common in India, but whereas the former can be found as far as the Malay Archipelago and beyond, rollers are absent in the southeast except in Celebes and beyond. The passerine birds are very numerous. The house crow, the Indian grackle, and the common mynah are familiar birds in India. Drongos (Old World passerines, usually black with hooked bills), flycatchers, bulbuls, tailorbirds, orioles, and many others are widely distributed, and broadbills are typical. Among the herons the white cattle egret is common throughout the region, whereas spoonbills, cranes, and gulls are mostly confined to the western part.

Of the crocodiles the gavial, which has long slender jaws and a soft, inflatable nose tip, is restricted to the large rivers of northern India; a species of an allied genus

is found in Sumatra and Borneo; and the mugger (the common freshwater crocodile) and the estuarine crocodile have a wider distribution. Freshwater turtles and land tortoises are well represented. Lizards are numerous, and flying lizards are also typical of the region. Chameleons are chiefly African, but one species is found in peninsular India and Sri Lanka. Snakes are numerous, among them the poisonous krait, cobra, and Russell's viper. Frogs and toads are abundant.

The freshwater fish life of the Oriental Region is rich. The carp and catfish families have many native genera and species. The labyrinth fishes (so named for a labyrinthine outpocketing of the gill chamber that permits them to take oxygen from air as well as from water), to which the climbing perch and the gourami belong, are characteristic of the fish life of the region, as are spiny eels.

Insects, arachnoids (scorpions, spiders, and mites), mollusks, and other invertebrates inhabit this region in great numbers. Large bird-winged butterflies, allied to the well-represented swallowtails, are typical. Almost all known families of scorpions are present. Among land shells the absence of Helicidae (a family of land snails having lungs), common in the Palearctic Region, is noteworthy. Their place is taken by other forms, such as *Hemiplecta,* and by land mollusks having horny or shelly plates on their posterior dorsal surfaces. (L.F.deB.)

The people

A discussion of Asia and its peoples cannot entirely exclude other parts of the Old World when the origins of man, his ethnic divergence, and his migrational wanderings, or the evolution and historic development of his linguistic and culture systems are considered. The relatively modern division of the largest of the continents into Europe and Asia derives from arbitrary decisions made by peoples living in the western part of the peninsula of Europe; this division has only minor significance in relation to the historic patterns of human occupation of the continent. The ethnic and linguistic diversity of Asia is greater than that of any other continent, because it represents ethnic types and linguistic systems that have evolved in separated regional homelands, as well as repeated patterns of modification and intermixture, resulting from both peaceful and militant migrations. Ethnically and linguistically, some territories have become highly diversified mosaics in which there are mixed and overlapping elements.

EVOLUTION OF ETHNIC PATTERNS

The three major racial stocks

Original racial stocks. The peoples of Asia include all the three major racial stocks of *Homo sapiens*—Congoid (Negroid), Caucasoid, and Mongoloid. The view is here taken that Congoid racial stocks derived from tropical origins in the zone extending from Africa eastward through the Indonesian Archipelago. It would, however, appear that to the east of Africa their original numbers were insufficient to ensure that they became the predominant element among the populations of South and Southeast Asia. The contrary belief that Congoid groups east of Africa originated in Africa leads to the conclusion that the migrating Congoids scattered out very thinly to the eastward and were able to contribute only to the skin pigmentation of South India and part of Southeast Asia. Whatever their origin, Congoid racial stocks at no time seem to have penetrated deeply into the Asian mainland. At the present time they constitute an element in ethnic stocks only in the southern peninsular fringes of Asia. The view here taken is that the Caucasoid racial stocks derive from a western Eurasian racial hearth, or region of origin, that also includes North Africa, and that Mongoloid racial stocks derive from an eastern Eurasian racial hearth. A very simplified view of the history of man in Eurasia, then, is that the western zone of the continent came to be populated primarily by the Caucasoid, or white, ethnic groups; that the eastern zone was populated primarily by the Mongoloid, or yellow, ethnic groups; and that the southern fringes of the continent were lightly occupied in early time by the Congoid, or black, ethnic groups. This pattern was apparently derived from the evolutionary beginnings of *Homo sapiens* possibly about 40,000 years ago, out of predecessor racial stocks, since the basic differentiation of the human species predates both modern man and the last glacial era. In considering modern man in Asia we cannot speak of distinct races in any definite way, as repeated migrational movements since the last glacial era have resulted in such intermixed racial stocks that the modern Asian cannot be racially defined. While it is possible to give a general description of an Indian, a Chinese, or a European, this can be done only in terms that connote ultimate regional origin in the broadest sense. In discussing Asian ethnology, therefore, we cannot refer to races but only to ethnic groups that speak particular languages and that have particular cultural characteristics. Thus, a Hunan Chinese from central China may be compared to a Czech from central Europe, or a Japanese from the southern island of Kyushu may be compared to an Iranian from southern Iran. All four belong to the single species but show marked differences in physique, language, and culture. It is, nevertheless, convenient to stereotype Chinese, Japanese, Czechs, and Iranians as members of "geographic races," meaning members of normally separated breeding populations.

Prehistoric migrations

Ancient migrations. The two primary prehistoric centres from which migrations over the continent took place were Southwest Asia and a region comprising the Mongolian plateaus and North China.

From prehistoric to historic times, possibly beginning as early as 30,000 years ago, movements from Southwest Asia continued toward Europe and into Central Asia; significant movements also took place into India. There were probably small divergent migrational movements in other directions that became swallowed up in later patterns of mixing. The Greeks were one of the late groups moving westward, about 2100 to 1900 BC, as were the Aryans who moved east to invade India from 1600 to 1500 BC. Mongoloid migrational movements have always been primarily toward Southeast Asia.

Important Mongoloid components, however, also moved westward through Central Asia toward the European peninsula. Such movements must have begun as early as 10,000 years ago, but they continued into the Christian Era as Mongols pushed Turkic peoples westward, setting off additional displacements of such peoples as the Finns and the Magyars. These westward Mongoloid movements also produced, over a period of time, much mixing of early Caucasoid and Mongoloid stocks in Central and West Asia. Northern Eurasia continued to be inhabited chiefly by thinly distributed residual elements of very early eastern Asian stocks, although some fairly late northward movements of Turkic peoples did take place.

There have been many small-stream movements away from the main trends, and these have often complicated the ethnic picture of any one region. At least one prehistoric Caucasoid movement penetrated East Asia and today is represented by the historic aboriginal population of Japan known as the Ainu. A countermovement out of India by a nomadic ethnic stock about AD 1000 contributed the Gypsy strain now so widespread in Europe.

Prehistoric countermovements along the China coast carried early Mongoloid migrants of Southeast Asia northward again into southern Korea and Japan, to leaven the later Mongoloid and Ainu stocks, from all three of which modern Japanese are derived. Similar northward drifts of early Mongoloid Indonesians account for a significant share of the ethnic ancestry of the population of the Philippines. Within the broad zone of Central Asia, prehistoric and historic movements have often retraced older migratory routes, creating overlapping and fragmented distributions of stocks that have yielded the many ethnic groups found there today. Secondary and tertiary intermixing of many of these regionally derived ethnic groupings has resulted in the emergence of regionally and physically distinct ethnic groups. Thus, the Uzbeks may originally have derived from a Mongoloid stock; some of them migrated westward to near the Volga River at an early date, then moved southward to become intermixed with Caucasoid-derived stocks. Uzbeks are now widely distributed in Central Asia and show considerable physical and linguistic variation.

Modern movements of peoples. Within historic time the aggressive expansion of particular ethnic groups has either driven weaker groups away from their territory or has resulted in the newcomers' assuming control of the territory and reducing the older inhabitants to the status of ethnic minorities. Some of these weaker ethnic stocks eventually became so diluted by intermixture as virtually to lose their identity. In some instances a new and variant ethnic stock with a different dialect resulted from the mixing. Some areas are now given over to distinct enclaves occupied by several diverse ethnic stocks, each following its own way of life. Thus, in Southeast Asia, from the riverine and coastal lowlands to the higher mountain uplands, different ethnic stocks have been migrating southward to become resident in separate altitudinal layers, one above the other. Some of this migrational movement in Southeast Asia is as late as the 19th century, but it has been going on for thousands of years. Within what are now India and Pakistan the general trend, for many centuries, has been eastward and southward, producing very discontinuous patterns. Discontinuity also characterizes the ethnic patterns in Central and Southwest Asia.

The mixture of ethnic groups

Militant campaigns of Arabs spread Islām and Arab political structures out of Arabia westward into Africa and Spain and eastward through the Levant into Asia Minor, Transcaspia, and India. Beginning in the 7th century AD and lasting until the 16th century, these efforts spread Arab ethnic elements widely in Southwest Asia and northern Africa.

Within recent times, movements of Caucasoid European Russians eastward along the Central Asian routes of exploration, and the penetration of the oceanic fringes of South and East Asia by Caucasoid western Europeans, have carried Caucasoids to all parts of the Eurasian continent. This has resulted in an interbreeding that has produced many local and variant mixtures. Since the 17th century, intermarriage between Europeans and indigenous Asians has produced many mixtures, including the Anglo-Indians of India and the Burghers of Sri Lanka. Intermarriage between Chinese men and local women has produced many hybrid strains in Indonesia, Malaysia, Thailand, and the Philippines. The introduction of American white and black soldiers to East and Southeast Asia, during and after World War II, has further complicated the ethnic mosaic in China, Korea, Japan, Vietnam, and the Philippines. Such modern racial mixings are often viewed apart from the historic patterns of migration and racial mixing, but essentially they form a part of them.

POPULATION DISTRIBUTION AND REGIONAL ECOLOGY

The background. Around 1750 the distribution of ethnic groups was relatively easy to describe. The whole of northern Eurasia was rather lightly populated by diverse ethnic groups of Paleo-Asiatic, Tungusic, and Turkic peoples who engaged in hunting, collecting, fishing, or herding; some groups, such as the Samoyed, Yakut, and Chukchi, had somewhat distinctive single economies or had economies that were seasonally mixed. Central Asia, Tibet, and Mongolia formed a mixed zone dominated by nomadic pastoralism, but the lower plateaus and lowlands were sprinkled with agricultural oases in which towns and villages were occupied by sedentary crop growers. Population was relatively light; mountain regions were occupied only in summer, but there were locally dense populations centred on such large oases as Tashkent, Samarkand, Kashgar, and Urumchi, with smaller groupings around lesser sources of water. The Buriat Mongols and the Kirgiz were pastoral, whereas the Tadzhiks, Uighurs, and Uzbeks were sedentary oasis dwellers. Southwest Asia was then inhabited by Iranian, Arab, and Turkish peoples, with a scattering of minority ethnic stocks, practicing either traditional pastoralism or the agricultural economy of the oasis. Population was concentrated around cultivable areas, water resources, or grass pastures.

Life-styles in the 18th century

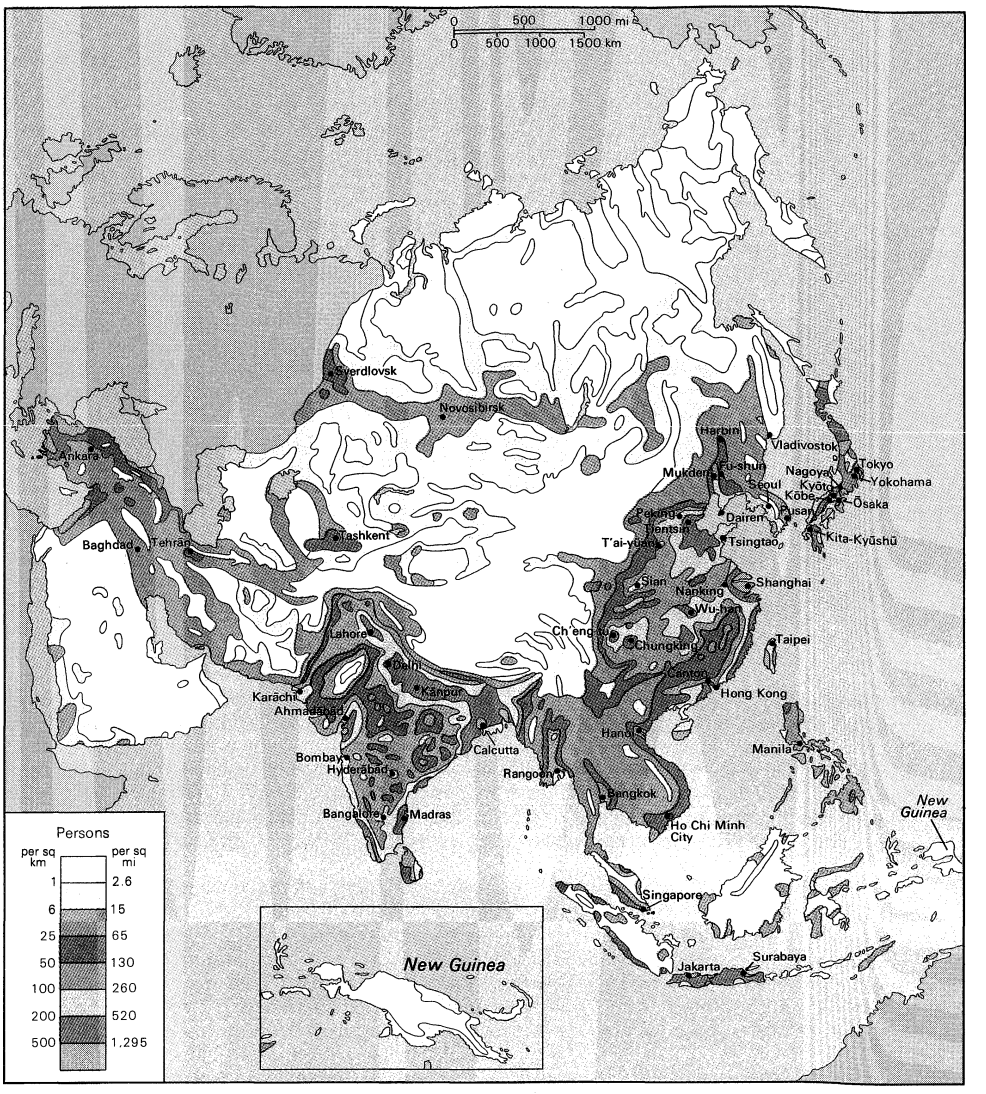
South and East Asia showed a more complex dual set of patterns. The largest components consisted of the highly civilized lowland populations, long settled on their land and engaged in sedentary agriculture and handicraft manufacturing. Market towns and cities were scattered over the countryside, and many small port towns dotted the seacoasts. Population density was heaviest in the best agricultural lowlands, which had also been occupied the longest, such as the North China Plain, southern Japan, coastal Vietnam, and the Ganges Valley.

Lesser components included the many diverse ethnic groups scattered in wet deltaic lowlands such as those of the Ganges, Irrawaddy, Chao Phraya, and Mekong rivers; the central plain of Luzon Island in the Philippines; and the north coast of Sumatra. Groups were also scattered throughout most of the hill and lower mountain country. Their living systems varied from simple hunting and collecting economies to more complex systems in which, apart from shifting cultivation (the shifting of cultivation to new land after a crop has been raised), hunting and collecting are also practiced. Generally these lesser areas had small populations scattered in hamlets or village settlements sustained by subsistence economies; only a little handicraft manufacturing took place, and trade was confined to minor products. The Nāga of northeast India, the upland Karen of Burma, and the Miao of Laos exemplified this life-style. Toward the end of the 18th century, European colonial efforts were beginning to shape the production systems of eastern Eurasia to conform to patterns of integrated world trade. The supplying of Europe with raw materials, which was to characterize the early 20th century, was also begun at this time.

The pattern of ethnic distribution. By the late 20th century, great changes had taken place in both the ethnic patterns and the associated life-styles in Asia. The divisions that prevailed in the 18th century were dissolving as the Soviet Union and China extended their economic and political control over Siberia and Central Asia, as the former colonial lands of South Asia established political independence, and as some of the component territories of the old Ottoman Empire were reshaped into the modern nations of Southwest Asia. Many of the hundreds of small ethnic groups were being absorbed into the populations of nation-states, many old languages were declining, and many formerly distinctive living systems were remaining in existence only as remnants or artificially preserved societies. The expansion of dominant ethnic groups was steadily restricting the territory available for older, simpler societies; and aspects of modern economic development were replacing the earlier systems. It is possible, still, to identify the region in which the Yukaghir formerly lived as a separate culture group in Eastern Siberia, but—for the few hundred Yukaghir who remain—political absorption, modernizing acculturation, and internal social decay have made the classic description of the group largely a historic one. Many former horse-riding, tent-dwelling, sheep-herding Kara-Kirgiz today ride tractors on Soviet state grain farms, live in permanent villages, and speak Russian in public. Some men of the Chotanāgpur hill region of eastern India, who formerly engaged in hunting and practiced shifting cultivation, today work in the steel mills of Jamshedpur. The remnant Ainu of northern Japan today are gathered into "cultural villages" where their continued woodcarving and bear dances attract a flow of tourists from southern Japan.

Social changes of the 20th century

Contemporary developments. Population densities have everywhere increased, and planned programs of modernized agriculture, mineral exploitation, and industrialization are bringing cultural change. Some of the small ethnic groups are dying out, but larger groups have often accepted change and are again increasing in numbers. In South and East Asia, increasing lowland populations are pressing hard upon the available land as population densities exceed 2,000 persons per square mile. In Central Asia, both Chinese and Russian settlement programs are moving peoples from heavily populated regions into frontier zones, in order to develop both agricultural and industrial resources. In Southern Siberia the Soviet settlement program has spread a thick wedge of European Russians and assorted ethnic minorities eastward to the Pacific and northward along every river valley to the Arctic Ocean. As a result, the Paleosiberian ethnic remnants are being submerged and absorbed. Old trading posts, oasis towns, and



Population density of Asia.

the few old cities of Southern Siberia and Soviet Central Asia are being developed into modern industrial centres; these are linked to modern transport systems by which raw materials and manufactured products flow back to the European regions. Most new cities are populated largely by European Russians, with the former Asian ethnic stocks remaining chiefly in the rural areas. The modernization of Southwestern Asia, through the renascence of Turkey and the impact of petroleum exploitation on the Arabian Peninsula, has altered many of the old patterns of ethnic groupings in these areas. A further alteration of the historic pattern came in 1948 with the creation of the State of Israel, to which more than 1,500,000 Jews from the Middle East, Europe, and North America have migrated.

Urbanism

Urbanism is becoming marked in many parts of Asia, thus heightening regional contrasts in population densities. Israel and Japan are among the most highly urbanized countries in the world. Asia's cities are among the largest in the world, and many of the newer cities resemble those of the West in terms of population, buildings, facilities, and congestion.

Ecological factors. Agriculture remains the mainstay of Asia, though the percentage of the population engaged in agriculture is steadily declining.

Although marginal lands in many parts of South and East Asia have been brought under cultivation, and many former pastoral ranges in Southwest and Central Asia are now irrigated, the broad ecological factors touched upon above continue to give rise to zonal distinctions in population and economic activity. Parts of South and East Asia can support dense populations. Favoured localities in the southwest—for example, in Turkey and northern Iran—support large populations. In Southwest and Central Asia in general, however, agricultural productivity and population density vary markedly with the regional pattern of rainfall or the availability of water from humid highlands nearby. In the Soviet sector the older pastoral nomadism has been transformed into organized transhumance (seasonal migration of stock between lowlands and mountains); consequently, the families that were formerly nomadic have become permanent residents in villages, and only herders accompany the flocks and herds. Northern Asia remains a semi-developed frontier region with short-season crop growing in favoured southern localities, even though breeding of newer varieties has extended agriculture northward. The Arctic fringe is being developed on the basis of mineral resource exploitation, but only in particular localities. Siberia continues to be lightly populated, with the population primarily grouped around local centres.

The importance of water

The pattern of language distribution. Language maps of an area tend to reflect conditions of the past because speech systems are constantly undergoing change. In the past, language maps for the Eurasian continent have often shown eight languages—Turkic, Slavic, Tungusic, Chinese, Tibeto-Burman, Indo-Aryan, Iranian, and Mongol—almost blanketing the main portion of Asia and leaving other languages predominating only on peninsular appendages, island fringes, and in small pockets. Except for the large eastward expansion of the Slavic group, the map

reflects distributional patterns that prevailed in the 18th century. At that time, by far the largest language groups, by number of speakers, were the Chinese and Indo-Aryan, but the Tungusic languages were probably used over wider areas. In recent times many of the small ethnic-group languages have, for practical purposes, been dying out, to be preserved only by professional linguists. The Tungusic group shows this decrease in usage, in spite of the Soviet practice of publishing books and newspapers in regional languages and the encouragement given to the preservation of the more important ethnic languages. Russian is now the dominant public language throughout the Soviet Union, and in the Asian territories it is spoken by large numbers of non-Slavic inhabitants. Similarly, Mandarin Chinese is expanding in China at the expense of local languages and dialects. In India, however, local languages are not losing ground, and language has become a territorial political issue. Meanwhile, the Indian government continues to use English as an official language. In Indonesia, which has many local languages and dialects, Bahasa Indonesia (the national language) has not yet spread throughout the Indonesian state. English remains the most commonly spoken single language in the Philippines, despite the adoption of a national Pilipino language.

The loss of small ethnic languages

Whereas many languages are dying out, as ethnic groups disappear or become merged into larger groups, some ethnic populations are increasing in numbers, thus increasing the relative importance of their languages. The large increase in the population of China now means that Mandarin Chinese is the world's leading language by number of speakers. The marked increases in population in Japan and the island of Java mean that Japanese and Javanese rank much higher on the list of languages, by speakers, than formerly. Similarly, Western Hindi, spoken in northern India, is one of the larger languages by number of speakers. Though many of the Paleosiberian (Paleo-Asiatic) languages are dying out as their ethnic users decline in numbers, the Uzbeks and the Tadzhiks, for example, have adjusted to Russian control and are again increasing in numbers, forming significant ethnic components of the Soviet Union, with the result that their languages are being maintained. It appears that many ancient languages spoken in Asia have disappeared within the last millennium and that others have been greatly modified by linguistic change.

Regional situations have sometimes produced multiple and overlapping language patterns. Around some of the old Central Asian oases and in Southern Siberia, migrants from Russia and exiled ethnic groups have created ethnically mixed regional populations. A comparable pattern may be discerned in Chinese Central Asia. Such large cities as Manila, Singapore, and Bombay show complex linguistic patterns. As European Russians have moved into the new cities in Transcaspia and Western Siberia, Russian has become the language of the cities; the older languages have been confined chiefly to the countryside.

FORMS OF ETHNIC ADMINISTRATION

Imperial administration. The older forms of administration, by which political states controlled adjacent and frontier ethnic groups, generally used local native leaders, who normally were given honorific subordinate titles and made responsible for the orderly control of their territories. The Chinese Empire sometimes entered into treaty-like agreements with subordinate states on its periphery and either subsidized the non-Chinese states or exacted tribute as a token of subordinate or feudatory status. The British in India and Burma, the Dutch in Indonesia, and the French in Indochina developed systems of frontier agencies that employed resident officials to supervise local leaders, who exercised autonomy over what amounted to ethnolinguistic groups. The Thai maintained their control of Siam (now Thailand) as a buffer state between the French and British regions but in their northern area maintained the older form of control through native leaders. Beyond the reach of these larger imperial states simpler ethnic groups maintained their local sovereignty under the rule of chiefs, shamans, or clan leaders, sometimes forming limited confederations.

Multiethnic states. The development of modern forms of political administration among Asian states has produced some distinctive regional patterns. The Soviet Union was the first state to organize administrative districts on an ethnolinguistic basis. There are about 100 separate ethnic groups publicly recognized in the Soviet Union, as well as some minority groups never identified, and about 60 of these are represented by political administrative territories at major or minor levels. China under the Communist regime has adapted this system and has modified the Imperial political structure in regions containing ethnic or linguistic minorities—primarily in South and Southwest China, Northwest China, and Central Asia. In the Soviet Union such ethnic territorialism is relatively fixed and stable, but in China there continue to be changes in spatial arrangements of autonomous regions as various pressures are exerted, for not all minorities have yet been given internal territorial autonomy.

Ethnolinguistic political units

In India, with several hundred languages and many varieties of ethnic groupings, ethnolinguistic recognition is made only at the state level. Several of the political states of the Indian Union are now bounded by linguistic limits. Many minorities are not recognized at present, and the question of spatial ethnic and linguistic autonomy has given rise to considerable unrest within the Indian Union. The former northeastern Nāga tribal agency, however, has become a full state on the basis of its cultural unity. In Pakistan the tribal and frontier agencies formed during British Indian rule are still preserved; in these agencies, spatial autonomy derives from the ethnolinguistic situation. Burma worked to resolve the problems of integrating ethnic minorities into a modern political structure after several upland ethnic minority groups expressed militant opposition to the forms of limited territorial autonomy offered by the government. Throughout the rest of Southeast Asia, except for Malaysia, ethnic minorities have generally not received formal recognition, and each country has tended to adopt different means of integrating its minorities into the national life.

Malaysia is a multi-ethnic state in which Malays total just less than half the total population; Chinese total just over one-third; and Indians, Pakistanis, and tribal groups almost equally split the remainder. The constitution makes no recognition of the plural ethnic composition; Malay is the official language; Islām is a state religion (although religious freedom is guaranteed); and the head of state must be a Malay. Quasi-legal political parties, however, represent ethnic groupings, and there are—in practice—many ways in which all ethnic elements are represented.

In Southwest Asia, minor populations exist in most political states without formal recognition of their status, the minority position deriving from ethnic, linguistic, or religious factors. Only in Saudi Arabia and Yemen (Aden) is there homogeneity in the three elements. Lebanon is ethnically and linguistically Arabic, but its population is almost equally divided between Christians and Muslims. Israel has a sizable Arab minority, and Iran is only half Persian in ethnic and linguistic terms. Most other states in Southwest Asia have comparable ethnic conditions.

(J.O.Es.)

DEMOGRAPHIC PATTERNS

Asia, which covers about a third of the total land area of the world, has a population of more than half the world total. Asia includes the two countries that have the largest populations in the world, China and India; these two countries alone have populations that together are estimated to comprise more than a third of the world's people.

The age structure of Asia's population, particularly in the developing countries, is predominantly young. One consequence of this is that the number of dependents—particularly children—is disproportionately large in relation to the number of employed adults. Another is that, in view of the high birth rate, the age structure favours large additions in the future to the massive population already in existence.

Youth of population

Several Asian countries, aware of demographic trends and their adverse effect on economic growth and social progress, have embarked on official birth control pro-

grams, which have met with considerable success. Japan's program has perhaps been the most effective. In operation since World War II, it includes well-publicized family-planning services, legalized abortion, and the provision of all forms of contraceptive devices. Programs in China, South Korea, Taiwan, India, Pakistan, and Sri Lanka offer family-planning services, birth-control clinics, vasectomies, and contraceptives (including intrauterine devices). The Soviet Union has an ambivalent population control policy, and birth rates in Soviet Asia continue to be relatively high. The Southeast and Southwest Asian countries lag behind in formal programs, but public consciousness and basic planning have grown.

In nearly all countries of the world, more male than female babies are born. In advanced industrialized countries, where maternal mortality is low and where infant girls receive as much care as do boys, the male death rates are higher than female death rates at every stage of life; the numerical excess of males at birth is, in consequence, gradually reduced until females outnumber males in the older age groups.

Some Asian countries, particularly India and Sri Lanka, as well as Pakistan and a few predominantly Muslim countries, have a high sex ratio—*i.e.,* the number of males per thousand females—in the sense that males outnumber females in all age groups, even though in a few categories of the population females predominate at birth. This sex ratio is unusual, and there is controversy about its cause. In some countries social attitudes are held to be responsible for the difference in mortality rates of the sexes after birth. Early marriages—if not quite child marriages—increase the initial balance in favour of males because of the relatively high mortality rate of mothers in childbirth.

While Asia and Europe are the two most densely populated continents, Asia is slightly less overcrowded than Europe. The distribution of Asia's population has traditionally followed a pattern of dense settlement in river valleys, where the soil is fertile because of perennial irrigation and where double-cropping (the harvesting of two crops a year) can sustain large numbers. This traditional pattern still exists, despite the emergence of another pattern, represented by the attraction of large and congested industrialized cities such as Tokyo, Calcutta, Bombay, and Shanghai—cities to which the underemployed rural population often migrates. The depressed rural economy, the periodic failure of the monsoon, and the near-famine conditions that result have contributed to the continual drift of population to the towns and cities; often, however, the shortage of available jobs results in many unskilled village farmers, who come to the city in search of work, ending up in slums rather than as operators of factory machines.

While haphazard, unplanned urbanization is proceeding in many Asian countries, the population in general remains predominantly rural. Asia's population is most urban in the Middle East, and in East Asia. (S.Ch./Ed.)

Traditional cultures

Within each area of Asia, different traditional peoples and their patterns can be identified, described, and classified together in many ways. The following represents simply one of many useful classifications.

SIBERIA

The main discussion of Siberian cultures is to be found in the section on *Siberia* in the article UNION OF SOVIET SOCIALIST REPUBLICS. Only a sketch is offered here.

Located east of the Ural Mountains, Siberia lies in a transitional area in which tundra or treeless plain becomes taiga or swampy coniferous forest. Farther south, the taiga blends into a forest steppe with alternating stands of trees and open, seasonal grasslands. Summers, except in the extreme south, are too short to permit agriculture; spring floods often inundate the taiga. Farther north are the Altai (Kirgiz) and Sayan (Kazakh) mountains, which abut Central Asia in the west and the Mongolian People's Republic in the east. Farther east are the mixed broadleaf forests of the Amur-Ussuri valleys, with hot summers and severe winters. The Pacific maritime coast and Sakhalin Island complete the geographic picture. The northeastern region of China (formerly called Manchuria) sits to the south.

Mongoloids of Siberia

The indigenous populations are primarily Mongoloid racially, although generally they are outnumbered in most areas by Russian, Ukrainian, and Belorussian immigrants. The aborigines still live mainly in separate, virtually endogamous villages, though many groups (including reindeer herders) have been collectivized by the Soviets. They speak several languages and dialects of the Uralic-Altaic family, among them Samoyedic, Tungusic, Mongolian, and Yakut (or other Turkic dialects).

The main economic pursuits include small-scale reindeer herding (seven to 10 reindeer per herd, small when compared to groups farther north), hunting (deer, elk, squirrel), trapping (sable, fox), and fishing. Traditionally, the inhabitants have lived either in longhouses and semi-subterranean huts in permanent villages or in conical, bark-covered tepees at temporary campsites and, for transportation, have used canoes, rowboats, skis and snowshoes, sleds drawn by reindeer or dogs, and, in the east, horses. Many aborigines now work in mines and other Siberian industrial projects.

Most aboriginal groups are small. The 80,000 Evenki (Tungusic-speaking reindeer herders), for example, live and wander in an area about the size of the continental United States.

CENTRAL ASIA

The main treatment of Central Asian cultures is given in the section on *Soviet Central Asia* in the article UNION OF SOVIET SOCIALIST REPUBLICS. Only a brief characterization is offered here.

Central Asia extends from western Manchuria to the Caspian Sea, along the grasslands, mountains, and plateaus of Inner and Outer Mongolia, Tibet, the Central Asian Muslim republics of the U.S.S.R. (Kirgiziya, Tajikistan, Uzbekistan, Kazakhstan, and Turkmenistan), and Afghanistan north of the watershed of the mountain range of the Hindu Kush. An interesting man-made boundary, the Great Wall of China, exists in the east. Altitudes and local ecology vary considerably from region to region and even within regions. The western Turkistan deserts and steppes lie close to sea level, whereas the upland Mongolian steppes often exceed 5,000 feet (1,500 metres) above sea level. Basically, Central Asia is an arid zone that has inland drainage and a continental climate.

Mongoloids of Central Asia

Racially, all Central Asians are Mongoloid variations, more Caucasoid to the west (Slavic) and southwest (Mediterranean subvariant). Most speak Ural-Altaic languages—usually Turkic dialects but also Mongolian. Other language families include Sino-Tibetan (Tibetan in Tibet and Chinese near the eastern and southern borderlands) and Indo-European (Tadzhik and Dari Persian). Cyrillic has replaced the Arabic script in Soviet Central Asia and Mongolian in Mongolia. Inner Mongolians now use simplified Chinese characters, but Tibetans, though under great pressure from the Chinese, cling to their own traditional script.

Islām pressed into Central Asia early in the 8th century AD, culminating in the Arab sack of Samarkand in 712. Most people as far east as Outer Mongolia and south into Chinese Sinkiang became Sunnī Muslims, constituting the main orthodox school of Islām, but pockets of Ismāʿīlīs, who consider the Aga Khan as their leader, live scattered in such mountainous areas as the Pamirs and the Hindu Kush. Buddhism centring on teachers known as lamas and heavily laden with animistic and shamanistic beliefs and practices dominates Tibet and neighbouring parts of Mongolia and China. Anti-religious drives by the Communists of the U.S.S.R. and China have been only partly successful.

Often called the belt of "pastoral nomadism," Central Asia also includes sizable sedentary agriculturists, usually existing in the loose-knit symbiotic system described earlier, with modifications wherever Communism has introduced collectivization of herds and land. Political ties are largely kin oriented, and a loosely stratified class system exists along with a hereditary nobility. Slavery was common until the Russians and British virtually wiped it out in the 19th century. Mongol nomadic patterns dominate:

groups move year-round within well-defined territories; there are few fixed economic centres.

Local ecology greatly influences livestock distribution. Fundamentally, sheep and goats are mountain animals (though sheep prove to be less adaptable than goats and tend to flounder in snow); cattle and camels appear more numerously in transitional forest steppes and semideserts. Yak and yak-bovine hybrids become important in the Pamirs and the higher mountains of Sinkiang, Mongolia, and Tibet. Donkeys and mules are found mainly in the west, where Central Asia nudges the Middle East. Horses, prestige animals throughout Central Asia, are ridden, and sometimes their milk is drunk, often as koumiss (fermented mare's milk). (Ed.)

SOUTH ASIA

The subcontinent of South Asia lies south of the Himalayas and Afghanistan. It comprises the countries of India, Pakistan, Bangladesh, Sri Lanka (Ceylon), Nepal, and Bhutan. While these countries together may be thought of as constituting a single region, they contain a multitude of ethnic groups. The peoples of the Himalayas have Mongoloid features. The upper-caste groups of the Indo-Gangetic Plains are basically Mediterranean in appearance. The tribal and semi-tribal populations of middle and southern India have dark skins, flat noses, and wavy or curly hair. The Pashtun (Pathan), Kashmiri, and Sindhi of the northwest are tall with long heads, white complexions, and wavy hair. The Kadar of Cochin, the Andamanese and Onge of the Andaman Islands, and the Vedda of Sri Lanka have the Negrito characteristics of dark skins, small stature, and frizzly hair. The tribes of the jungles of Bihār, Orissa, and Madhya Pradesh share certain characteristics with the aboriginals of Australia and are sometimes referred to as Proto-Australoid; some of them (the Oraon and the various tribes of Gonds) speak Dravidian languages, while others (the Mundas, the Santāl, and the Ho) speak Austro-Asiatic languages.

The South Asian population may be divided into tribal and nontribal peoples. The tribes are somewhat detached from the mainstream of civilization, having been isolated in the forests and hills. They constitute about 6 or 7 percent of the total population of India and varying proportions of the populations of the other countries.

A vast number of languages are spoken in South Asia, belonging to four different families of languages. In India alone there are more than 100 separate languages, of which 15 have been designated as national languages.

Selected ethnic communities of South Asia.

The South Asians are predominantly religious in their outlook. Their religious orientation pervades every sphere, from social and economic life to science, literature, and art. Whether one thinks of Hinduism in India, Islām in Pakistan, Buddhism in the Himalayan states, or tribal religions in the hilly areas, one cannot escape the fact that the traditional way of life is largely determined by ritual and belief. Along with this religious outlook there appears to go a tendency to fatalism, a sense that every course of action as well as its outcome is up to the gods. This is true not only of Hindus and Buddhists but even of Muslims; as S.M.H. Zaidi observed in *The Village Culture in Transition: A Study of East Pakistan Rural Society* (1970), the Muslims "have a predominantly fatalistic attitude to change, believing that no human hand without the will of God can change anything."

Another characteristic of the South Asian peoples is that they are predominantly peasants. More than 90 percent of them live in villages and depend on simple agriculture for their livelihood. The average person lives his whole life in the same village. There is much interaction between the village societies and the important religious cities such as Vārānasi (Benares), Mathurā, Ayodhyā, Gayā, Prayāga (Allāhābād), Rāmeswaram, Madurai, Cape Comorin (the southernmost tip of India), Dwārkā, and Kāmākhya.

Caste groups and tribes. South Asian society is characterized by a hierarchical social system with a vast efflorescence of caste groupings. A caste group is hereditary, usually localized, and associated with a particular occupation; its members do not marry outside the group. There are complex relationships of superiority and inferiority among caste groups, including restrictions on eating and drinking with one another. At the bottom are the so-called scheduled castes, or untouchables; in India these amount to about 15 percent of the population. These groups engage in various "impure" occupations such as scavenging, sweeping, cattle herding, and leather working.

At the lower levels of society, caste groups are not easy to distinguish from tribes. Many scholars, in fact, have suggested that caste groups evolved from primitive tribal groups. One difference is that while caste groups are endogamous (the members of a group do not marry outside it), the marital rules among tribal groups are somewhat more complex. Each tribal group confines its marriage relations to the same tribe or subtribe. But each tribe is divided into clans or moieties, often named after animals, birds, or plants, and these groups are exogamous; *i.e.,* the members of a clan must marry outside it. Such division of a tribe into several social groups or clans is found both in matrilineal tribes like the Garo and Khāsi of Meghalaya and in patrilineal tribes like the Nāga, the Munda, the Santāl, the Bhīl, and the Gond. Some tribes of southern India have a complex social organization. The Urāli, for example, who are matrilineal, have two sets of clans: brother clans, among whom marriage is prohibited, and brother-in-law clans, among whom marriage is sanctioned.

Tribal organization

Even among the tribes of the Himalayas, where there is a great tolerance of casual sex relations, both premarital and extramarital, the rule of clan exogamy is strictly followed. Among the Sherpa of Nepal, the term for clan is *ru* ("bone"), implying that the members belong to the same father's bone or blood. A recent study of this tribe failed to reveal even a fleeting amorous attachment between members of the same clan and no marriage between members of brother clans. The Dimasā Kachāri of Assam have different sets of clans for different sexes: there are 40 male clans and 42 female clans, all practicing the system of double descent, by which the male child is given the clanship of his father while the female child inherits the mother's clan. Very few tribes lack the totemic clan division. The Saora (Savara) and Māler of middle India have territorial divisions rather than clans, and their marriage relations are exogamous within these divisions. The Saoras are divided into 17 such exogamous units.

Certain tribes of middle India have clans organized in groups or moieties. Among the Gond of Bastar, for example, the moieties are composed of 96 and 69 clans. The Rāj Gond Ādilābād have four groups of clans, and the Muria have five.

While most tribes lead a peaceful life, there have been some warring groups. The Baluchi nomadic groups used to fight among themselves over pastureland, leadership rights, and women. Descriptions of the Nāga tell of their weapons, their methods of warfare, and their custom of head-hunting.

Caste and social structure

The basis of the social structure in South Asia is the caste system. Although it is Hindu in origin, the caste system has had a pervasive influence on all South Asian society, whether based on tribalism, Islām, Buddhism, Sikhism, or Christianity.

In Nepal both Hindus and Buddhists are subject to the code of the caste system. Among the Hindus of Nepal there are high castes such as the Brahmins, Ṭhākurs, and Chetris. The followers of Buddhism in Nepal are also divided into three main caste groups. There are the Barnas, the Udās, and a mixed caste of heterodox Buddhists. The Barnas enjoy a status equal to that of the Hindu Brahmins and observe the same rules of commensality and intermarriage. The Udās are inferior to the Barnas in the caste hierarchy. The third group of mixed castes includes numerous occupational castes (barber, carrier of light, etc.) and professional castes considered to be equivalent to the untouchables among the Hindus. The tribes also have several categories: the Gurung and Magar are at the top; the Newār are in second place followed by the Kiranti, the Khambu, the Limbu, and the Yākhas; below them are the Sunwar and Tamang (Mūrmi), who are given approximately equal status, while the Tibetan tribes including the Sherpa are ranked slightly below them; at the bottom of the scale are the Tharu, the Thami, the Hayu, the Thakali, and numerous other minor tribes.

In Sri Lanka the Goyigama (cultivators) enjoy the highest position, while the castes of potters and washermen are ranked low. These caste groups are endogamous.

The caste hierarchy, even in the traditional system, was not entirely rigid. It was possible for caste groups to move upward, although slowly. Even tribes have been able to acquire an appropriate place in the caste hierarchy. The Rāj Gond of central India, for instance, successfully claimed for themselves the rank of Kṣatriyas (a ruling and warrior class) on the basis of their acquisition of political power. Recent researches have revealed how caste-like groups such as the Bhumij, the Mahali, the Bauri in West Bengal, the Cheros, the Kharwār, and the Manjhi in Bihār, as well as various sections of the Bhīl and the Gond, have been assigned appropriate status in the Hindu caste systems in their regions.

Islām's claim of equality for all those who profess the faith has been greatly weakened in South Asia because of caste practices. The Muslim caste system, unlike the Hindu, is not based on purity and pollution concepts; nor is it sanctioned by religious ideas. Among the Muslim castes, the highest are the sayyids and the lowest are the sweepers. In Bangladesh the landholding groups of foreign (Persian or Arabian) descent are generally ranked high, while those who do manual labour that involves ritual contamination are ranked lowest.

The hierarchical system in Pakistan may meaningfully be compared to the Hindu caste system. In Swāt in northern Pakistan there are seven social groups; they are, in order of descending rank: (1) persons of holy descent, (2) landowners and administrators, (3) priests, (4) craftsmen, (5) agricultural tenants and labourers, (6) herders, (7) the despised group.

Sikhism also, despite its emphasis on equality, has a caste structure. The Sikhs are broadly divided into Sardars and Mazhabis, the former consisting of higher caste groups and the latter of sweepers. Caste restrictions are also rigidly observed among the Christians of the west coast, particularly in marriages.

The joint family

Kinship patterns. The joint family is another basic institution in the societies of South Asia. In India the joint family has existed since prehistoric times. The coming of Islām failed to modify the structure of this most ancient institution. Despite the industrial revolution and other modernizing influences, the joint family continues to exist on a large scale.

There are two types of joint families. In the northern

type, the family unit includes three or four generations of males and their dependents: one's grandfather and his brothers, one's father and his brothers, one's own brothers and cousins, one's sons, and the nephews and wives of all these male relatives plus one's own unmarried sisters and daughters. The southern type of joint family includes one's mother's mother and her sisters and brothers, one's mother and her sisters and brothers, one's own brothers and sisters, one's mother's sister's sons and daughters, and the children of one's sisters. The husbands of the women of the family live in the houses of their mothers, occasionally visiting their wives and children. Throughout Nepal the members of several generations in the father's line maintain common arrangements for cooking and have joint ownership of property under a single head. As in India, in some families the relatives on the mother's side may also live with the household in an extended family. Throughout Nepal, the kitchen is regarded as a symbol of family unity.

In Bangladesh the largest unit of family organization is the patrilocal extended family, in which two or three generations may live together. The extended family includes two or more nuclear or joint families united by male kinship bonds and living in a cluster of adjacent dwellings. It may have as many as 50 or 100 members or as few as seven or eight, the average being about 20. The extended family includes a number of component property-holding units such as the nuclear family, the joint family, and the irregular family (centred around a widow or a widow and her unmarried children). As long as the father remains alive after his son's marriage, the joint family is the rule rather than the exception. It is clearly related to the system of inheritance, by which land is allotted to the son only after his father's death. The joint family is slightly less frequent in Bangladesh than in West Bengal but more common than in Pakistan.

In Sri Lanka the nuclear family prevails. A Sinhalese family generally consists of a husband and wife and their unmarried children. It leads an economically self-sufficient life under the benevolent and mildly patriarchal rule of the father-husband.

Most tribal societies in South Asia have a nuclear family.

Marriage. Patterns of family organization can also be viewed in terms of the marriage relationship, the kinship system, and the position of the women in the family. The organization of the family throughout northern India is that of the patriarchal family. The traditional rule of avoiding marriage with somebody who is removed by less than seven degrees from the father and five degrees from the mother is practiced by all castes, from the highest to the lowest. In southern India a man preferably marries his maternal cross-cousin—*i.e.,* his mother's brother's daughter.

Polyandry among the Sherpa

Among the tribes of the high Himalayas, such as the Sherpa, the Bhutia (Bhotia), and the Lepchā, it is common to have more than one husband, usually brothers. The Sherpa believe that a polyandrous union is desirable because it prevents the fragmentation of property and fosters solidarity among brothers. It has been observed that since such a marriage improves a woman's economic prospects and assures her the advantage of a younger husband in later years, she prefers to marry two brothers. The Sherpa style of polyandry differs from the Tibetan practice that permits any number of younger brothers to share an elder brother's wife; this is also prevalent among the Khāsa, the Toda, and the Nāyar, who allow more than two younger brothers to share the same woman.

The custom of paying a bride price is prevalent among almost all the tribes of South Asia. The bride price may be livestock, food, clothing, or even service by the bridegroom in the bride's father's house. The ease of acquiring wives and the freedom of sex relations in the tribal communities make divorce an easy process.

In Islāmic law, marriage is essentially a contract rather than a sacrament, as it is among the Hindus. The rules of marriage prescribed by the Qur'ān have been modified in parts of India by the customs of the particular region. Marriage among the Muslims as among the Hindus is an important family affair, arranged by the elders of the family. Marriage is possible between almost any boy and girl under Muslim law, even within an extended family and between cousins. While traditional Muslim law does not sanction dowry, the Hindu influence compels it. The bridegroom's family expects the bride's father to give enough in dowry to establish a home for the couple. A Muslim can divorce his wife at his own will, but no such privilege is accorded to the wife. The custom of *pardah* (wearing the veil) is observed strictly among the Muslims of South Asia.

Socialization and education. The birth of a child in the societies of South Asia is considered a divine blessing. A childless woman is viewed unfavourably. In Hindu society a very important ceremony called *śrāddha,* conducted after a person's death, can be performed only by a son. It is believed that without this ceremony the departed soul is not likely to attain salvation. To the peasantry of South Asia, a male child means a pair of extra hands to help on the farm. Girls are not highly regarded; having a female child is preferred only to remaining childless.

Rites of passage

In Hindu society an individual passes through a series of rites from his conception to his death. Several rites are performed before the birth of a child to ensure his safe arrival. After the child is born, other ceremonies are conducted to provide the incarnating soul with a good body and prepare it for life.

Next are the rites known as *vidyārambha* (commencement of education), *upanayana* (the sacred-thread ceremony), *vedārambha* (commencement of Vedic recitation) and *samāvartana* (convocation). Traditionally a Hindu boy was taught by the guru in the *gurukula,* where he led a life of rigorous discipline and purity. He was then ready to undergo the ceremony of marriage and become a householder in his joint family. A man's death is mourned by the members of the family, his kinsmen, and other members of his larger communities. A series of death rituals are observed by his eldest son in cooperation with the kin groups to ensure the deceased freedom from rebirth. The Indian joint family continues to practice some of the traditional rites or *saṃskāras,* especially the birth rites, the sacred-thread ceremony, the marriage ceremony, and the death rites.

Most of the Muslim communities are converts from the Hindu religion, and there is considerable similarity in the methods of child rearing. One important difference is the Muslim ritual of circumcision that usually takes place between the ages of three and five.

The tribal communities have relatively informal relationships among kinsmen, in contrast to the highly formalized Hindu relationships. The youth dormitory so common among the tribes is the traditional institution of the unmarried boys and girls; it provides training in folk songs and dances, the moral code, customs and manners, and sexual relations. The young learn the arts of life mostly by observation and imitation.

Economic life. *Villages and settlements.* Most of the people of South Asia still live in villages and follow the traditional pursuits of an agricultural and pastoral economy. In the high Himalayas the tribespeople live with their flocks in established summer and winter encampments, each of which remains unoccupied for some months in a year. The pastoral nomads of the Baluchi tribe in Pakistan generally have only temporary encampments. The Dafla, Apa Tani, Gallong, and other tribes of Assam site their villages on the highest parts of the hills, from whence the whole valley below is visible. Many tribes site their villages in accordance with the will of God as interpreted by their shamans. Hundreds of villages of the Santāl, Munda, and Ho tribes in middle India are in the forest clearings. In southern India many tribal villages are located in dense forest, and their economy revolves around the forest. In western Rājasthān there are dispersed settlements in desert or semidesert areas.

In the deltaic plains of West Bengal and Bangladesh, the countless rivers and streams have given birth to compact, nucleated settlements. More typical of the area are clusters of houses scattered along the banks of streams and rivulets. Each cluster usually forms a separate residential unit consisting of a number of huts built around a courtyard.

The linear settlement is much in evidence in the plain areas of Orissa, Madhya Pradesh, Andhra Pradesh, Kerala, and Tamil Nadu. The abundance of river valleys and the prevalence of the *jajmānī* system (in which lower castes perform specialized artisan services for the landowners and cultivators) serve to maintain the traditional linear settlement pattern. The linear settlement is also found in fishing villages on the sea, or estuaries, or along big rivers.

Villages vary in size from more than 1,000 people to fewer than 50. The villages of the rich agricultural plain areas are usually larger, while those of the hills and forests are smaller. Some tribal villages are temporary encampments of five to 10 huts that provide accommodation for up to 30 persons. The tribal villages of the Hill Māler, the Hill Khariā, or the Bondo highlanders in middle India are slightly bigger, composed of 10 to 30 huts accommodating 30 to 100 persons. Large villages like those of the Waromung in Nāgāland, having a population of 1,600, are exceptional.

The villages of the settled cultivators are large and somewhat elaborate in layout. Depending on the extent and fertility of the land, the population of an average village varies from 100 to 500 persons. Hindu villages, especially those in the Indo-Gangetic Plain, may have as many as 5,000 inhabitants, divided into a number of hamlets. The central part of a village is occupied by the dominant castes, while the untouchables live in a far-removed area. Every village has a temple located near the settlement of the high castes. The houses of the landlords are elaborate, while those of the tenants are simple. Most houses are built of mud and tiles, with courtyards and parlours.

Food getting. Almost all of South Asia lies in the monsoon belt, and its agriculture depends on the torrential rains associated with the monsoon. The vagaries of the weather pose a great problem to the peasantry. When there is a lack of rain they turn to the rain god, Indra. In Nepal the guardian deity Machendranāth, who presides over agricultural affairs, is elaborately worshipped; the ritual of asking for rain is observed collectively by the villagers, who go from shrine to shrine in a procession, shouting, "O Lord Mahādeo, give us rain!"

Agricultural organization

Agricultural technology is still very simple. While among the forest tribes the primitive occupations of hunting, food gathering, and collecting still predominate, the basis of the South Asian economy is the plow. The Indo-Gangetic Plain, extending from the Punjab in the west to Bangladesh in the east, along with the coastal deltaic regions of eastern India and Sri Lanka, is the largest rice-growing zone in the world. In Nepal, Sikkim state of India, and Bhutan, intensive plow cultivation is practiced in the valleys and plains, but a primitive slash-and-burn type of cultivation is found in the hills. Pastoral pursuits are supplemental to agriculture in most South Asian villages. There are, however, communities, such as those of the Pathan and Baluchi in northwestern Pakistan and of the Gūjar and Gaddi in the western Himalayan districts, whose entire economy revolves around pastoral nomadism. During the summer they move up the mountains in search of pasture, returning in wintertime to the foothills.

The economy of the tribal peoples is based on hunting and gathering. They eat edible roots, fruits, vegetables, honey, insects, fish, and various kinds of birds and game. Some tribes such as the Birhor in Bihār, the Chenchu in Andhra Pradesh, the Juang in Orissa, the Kadar in Kerala, and the Paliyan and Paniyans in Tamil Nadu depend on the forest as their primary source of food. The tribes of the Assam Hills, of Orissa, of Madhya Pradesh, and of Nepal, who practice shifting cultivation (moving from one location to another), supplement their economy with occasional hunting and gathering. Even agricultural tribes such as the Munda, the Oraon, and the Ho of the Chota Nāgpur Plateau depend on the forest to supplement their harvests.

The hunting tribes use various types of bows and arrows and trapping devices. They also have simple missiles such as the throwing stick of the Oraon and the nonreturning boomerang of the Bhīl. Slings are also used to throw stones and hardened clay balls. Spears with iron heads and wooden shafts are very common. The Buna use multipronged harpoons with barbs for fishing. The Jarawa of the Andaman Islands poison their spearheads.

There are many forms of land tenure in South Asia, ranging from common ownership by villagers to individual ownership by the head of a nuclear family. In the past, many landlords gained their property under British occupation when they were given tax-collecting rights; the tax-farming system led to a great stratification of the peasantry. At the same time, the pressure of the population upon the land produced a fragmentation of holdings in some areas. Since independence there have been various attempts at land reform.

The *jajmānī* system

A common institution of Indian village life is the Hindu *jajmānī* system, practiced all over India under different names. Each caste group within a village gives certain standardized services to the families of other castes. A member of a carpenter or a barber caste, for instance, works for a particular family or group of families with which he has hereditary ties. The association passes from father to son. The system serves to reduce the need for money and to minimize competition. It also tends to assure a stable labour supply for the dominant agricultural caste in a particular region by severely limiting the mobility of the lower castes. If a worker leaves his village, he must supply a substitute, usually a member of the same joint family; but the system gives him rights and advantages that make him reluctant to leave and endanger his social and economic security.

Another traditional element of economic organization is the local market. The weekly or biweekly markets are located in open areas between villages where the means of communication provide a central place for the economic exchange of the region. In the nonmonetary economy of the tribal areas of India and Bangladesh, villagers exchange their produce for clothes, utensils, spices, and other necessities. Occupational specialists such as barbers, tailors, blacksmiths, and potters provide their goods and services. The agricultural produce from the surrounding regions is collected here and sold. The intertribal market is also a centre of social, religious, and political life.

Moneylending is an integral part of the agrarian economic structure of South Asia. In Bangladesh, the so-called middle-class farmers lend out their surplus money. In Indian villages the moneylenders come mainly from business castes and traders who have settled in the tribal villages.

The worship of spirits

Religion and art. Life in South Asia revolves around religious belief and practice. Belief in supernatural entities is strongest among the tribal communities; they believe that disease and death, failure in hunting or cultivation, and all the other hazards of life are caused by ghosts and spirits who can be propitiated with the sacrifice of animals. A study of the religion of the Ho and other tribes of the Mundari group has shown that this belief in the power of spirits to control the nature and destiny of man is basic to their religion. The Buddhist Sherpa observe certain rites to protect themselves and their land from evil spirits. Among the Muslims of South Asia, the matrix of primitive faith is found in the worship of *pīr*s, or saints, a practice for which there is no authority in the Qur'ān.

The tribes and also the Hindus worship certain forest and tree spirits, water spirits, and other spirits in natural objects. Among the agricultural tribes, every village has in its vicinity a grove (*sarna*) reputed to be a remnant of the primitive forest left intact for the local gods when the clearing was done. This sacred grove is the chief village shrine among many Mundari and Dravidian tribes of middle India. These tribes have special festivals associated with trees. In the festival of Karma, a branch of a sacred tree is cut and planted in the village assembly ground, and the unmarried young of both sexes dance around it. The Karma festival is also observed by Hindus in north Indian villages.

The Hindus consider the pipal tree (*Ficus religiosa*) sacred, believing that its trunk provides a habitation for Brahmā, its twigs for Vishnu, and its leaves for other gods. Water is poured on its roots by pious people after their morning bath. The bel tree (*Aegle marmelos*) is associated with the Śiva (Shiva) cult. The small plant tulsi (*Ocimum sanctum*) is associated with Vishnu, and its leaves are of-

fered to him during worship. The presence of a tulsi plant in the courtyard is the mark of a pious Hindu house.

The worship of rivers and springs is of great significance among the tribal people, who believe that the indwelling spirit causes the water to move. The tribes of Chota Nāgpur offer sacrifices to the water spirits. The people of the plain consider some of the rivers, particularly the Ganges, to be sacred. The spirit of the Ganges is recognized by many Hindus as one of the wives of Śiva, and there are periodic rituals involving bathing in the Ganges and other rivers. The non-Aryan hill tribes worship mountains.

Worship of the sun has been common among both the tribes and the Hindus. The supreme spirit of the tribes of middle India is Singh Bonga (Sun God). Some temples have been erected to the sun, notably those at Konārak near Puri in Orissa and at Deo near Gayā.

Many tribal peoples and Hindus also worship the mother earth. Pious Hindus offer a prayer to her before walking in the early morning, and some farmers offer a sacrifice in her name before digging the land. Hindu mythology and folklore hold that the mother earth rests on the hood of a cobra; he is also considered sacred, and a festival is devoted to him. Some deities are associated with certain diseases, as the goddesses of smallpox (Sītalā) and of cholera (Olābibi), who are worshipped by singing devotional songs in their praise.

According to the scriptures, the Hindu pantheon has 330,000,000 deities, although even the most detailed texts have never tried to list all of them or even all the categories of them. There are many shrines to gods such as Śiva, Vishnu, Gaṇeśa, and their wives, as well as to the mother goddess Śakti (Kālī or Durgā). A shrine may be a pile of stones set up under the village tree or an elaborate iconographic creation in stone, marble, bronze, copper, silver, or gold. The latter are found at places of pilgrimage. The ancient Kṣetra (sacred grounds) celebrated in the scriptures have played a significant role in bringing together peoples from north and south and from different sociolinguistic groups.

Types of priesthood

Diverse types of priesthood are found among the tribal and Hindu communities of India. There used to be tribes such as the Korwa and other primitive communities that had no priests because they had no gods, their worship being confined to the spirits of their ancestors. Some tribes such as the Munda, the Oraon, and the Gond have their traditional priests, who perform ceremonial observances. There are also shamans who are said to have the power to cure diseases or other evils caused by malignant spirits. They give statements believed to be oracular.

In the Hindu religious system the highest caste group of Brahmins are traditionally and professionally the priests. Every orthodox Hindu of clean caste must have a Brahmin family priest to whom he pays customary dues in return for services at weddings, funerals, and other ceremonial occasions. The sacred specialists popularly known as *paṇḍā* are found at places of pilgrimage. Hindu religious organizations also sanction a number of ascetic orders, mainly at places of pilgrimage.

Islām as a religion is optimistic and transcendentalist; it does not favour either asceticism or extremes of ecstasy as does Hinduism. Three types of religious leaders are commonly found in the Muslim rural areas of South Asia: *mullah*s, *mawlānā*s, and *pīr*s. A *mullah* is a prayer leader and religious functionary who performs various ceremonies for the villagers, usually in association with a particular mosque or sector of the village. His office is hereditary, and he exercises control over secular affairs through discourses on Fridays on nonreligious matters. The *mawlānā* is better educated than the *mullah* in theology and serves as an interpreter of the Qur'ān. The *mullah* ordinarily supports himself entirely through his religious occupation, but the *mawlānā* generally earns his living from a secular occupation. A *pīr* is a holy man who commands a following of both Hindus and Muslims in the South Asian countries.

Religious amalgams

The remnants of Buddhism in South Asia are found in the Himalayan states of Bhutan and Nepal, as well as in the Indian territories of Himachal Pradesh, Kashmir, Sikkim, and northern West Bengal. The Sinhalese of Sri Lanka also practice Buddhism. The Himalayan Buddhists worship Hindu deities and observe certain Hindu ceremonies; sometimes they employ a Brahmin priest. Certain Himalayan tribes such as the Lepchā, the Bhutia, the Limbu, and the Lahaula, who practice Buddhism, combine it with animistic beliefs. The Chakmā of Bangladesh profess to be Buddhists, but their religious practices are mixed with the folk Hinduism of West Bengal. The same fusion of beliefs appears among the Magh of Chittagong, who have added the worship of Śiva and Durgā to the Buddhist observances.

Several other religions, including Jainism, Sikhism, Christianity, and Zoroastrianism, are also found in South Asia. Jainism is the chief religion of the district of Jodhpur and of certain trading and business castes. It may be regarded as a sect of Hinduism, and intercaste marriages are frequent. Sikhism, like Jainism, has a close affinity with Hinduism. Sikhs are disciples of their 10 gurus, some of whose writings are compiled in the sacred book the *Adi Granth*. The Sikhs are concentrated in Punjab, though migrants are found all over India. One of the most remarkable figures in the religious history of India was Kabīr (d. 1518), a mystic poet who established a new religious sect linking Hinduism and Islām. Groups of his followers are scattered through northern India.

Religion has been both the inspiration and the matrix of South Asian art, which is discussed at length in the article SOUTH ASIAN ARTS.

The cultures of the tribes of all India, the Himalayan states, and Sri Lanka are characterized by dances and songs. Most of these reflect the community spirit of the tribes and their patterns of life. The village *akharā* (dancing ground) and the youth dormitory, found in most of the tribes of India, are important in the development, maintenance, and transmission of the tribal dance pattern. The numerous traditional festivals are inseparably linked with singing, dancing, and drumming.

Modern developments. The traditional cultures of South Asia are rural, caste oriented, conservative, and religious in outlook. They have altered somewhat in the last 150 years under the impact of British rule and in response to modern education, technological development, and political change.

The tendency of the lower castes to rise upward accelerated under British rule. This process of upward mobility has been variously named Brahmanization, Sanskritization, and Kṣatriyaization. It may be observed both in tribal and in nontribal communities. Some of the tribes have been greatly influenced by the Kṣatriya *varṇa*, or class. This *varṇa* includes a number of caste groups, headed by the artistocratic Rājput lineages and differing widely in status. Some tribal and lower caste groups have been able to distinguish themselves by using the class name Kṣatriya or claiming Rājput descent. They marry their daughters to men of Rājput lineage. There are also examples of certain tribal groups claiming the status of Brahmins in the Hindu caste system. The Bauri of West Bengal claim to be Brahmins by virtue of their observance of the mourning period of 11 days, wearing of the sacred threads, and worship of Hindu deities.

Attempts to escape caste

Lower caste Hindus have made even more vigorous efforts to raise themselves in the caste hierarchy. Organized violation of caste rules by the so-called untouchable castes developed under the leadership of B.R. Ambedkar, who finally despaired of eliminating untouchability from Hinduism and led his followers into Buddhism. A similar turn to Buddhism took place quietly in Sri Lanka. There are instances in India of lower caste groups seeking to adopt the customs and rights of higher caste groups and being prevented from doing so. Such efforts have led to the partial breakdown of the *jajmānī* system in the villages and have also been a factor in the migration of the numerically weaker castes to the towns.

Some of the tribes in the hilly and forested areas of India were converted to Christianity by European missionaries, and made a conscious effort to abandon their tribal traditions. After India gained its independence, however, some of them foresook their Christianity in order to participate in certain political and economic privileges enjoyed

by the traditional tribal groups. In northeastern India, where Christianity is strong among the tribes, they pressed demands for separate statehood within India, resulting in the establishment of Nāgāland, Meghalaya, and other tribal states.

The process of modernization and Westernization has impinged upon every sector of the South Asian cultures. The rate and direction of change differ from region to region. South Asia has been selective in what it has accepted from the West, mingling its acquisitions with many traditional institutions. Thus the new industrial towns are treated as modern places of pilgrimage; the secularization of society is pursued in terms of religious tolerance; and the parties in democratic elections contend in terms of old values. South Asia is clearly in a period of transition from traditional societies to modern ones, and the course it is following is unique. (L.P.V.)

EAST ASIA

East Asia as a cultural entity is here defined as comprising, first, the Chinese and groups that have been derived from the Chinese ethnically or culturally and, second, those cultures that were greatly influenced by China at the time that they became states. Among the first category, of course, are the Chinese themselves, the Taiwanese, and, with some equivocation, the Miao and Yao. The second category comprises such peoples as the Japanese, Koreans, Manchus, and Min-chia. Other peoples have also been influenced by the Chinese, but, because they have not formed states of their own, this influence is difficult to determine and is often not emphasized. This is the case among, for example, the Nakhi of southwest China and the Ainu of Hokkaido.

Recognition of ties with China is the most characteristic feature of East Asian cultures. It unites peoples as disparate in many aspects of life as the Hui-hui of China's northwest, the Japanese, and the Manchus. The area thus has a special significance for many of the peoples themselves and is not simply a construct of social scientists.

People and languages. Some peoples in Asia more or less form a buffer zone around East as well as South Asia. These ethnic groups are of diverse origins and cultures. The Ainu, situated on the island of Hokkaido, appear to represent in ethnic terms a conservative autochthonous refugee group that cannot easily be connected with any other group. They are linguistically distinct from other groups but culturally akin to the peoples of the Soviet Maritime Province and Eastern Siberia. The Manchus, Xojen, Daghur, and Monguors are small ethnic groups situated along China's northern and northwestern frontier. The first two speak a language of the Tungusic branch of the Altaic linguistic family, while the Daghur and Monguors speak Mongolian languages, also of the Altaic family.

The Islāmic Hui-hui

The Hui-hui (also called Tungan in Chinese Turkistan and Panthay in southwest China), or Chinese Muslims, take their ethnic identity from their religion. Those aspects of life in which they differ from Chinese are related to their practice of Islām, although they also exhibit a greater esprit de corps and competitiveness. There is a natural tendency for members of an ethnic unit who can communicate readily with members of the dominant group to merge with that group, and many Hui-hui have seen themselves as predominantly Chinese. But it is also true that Chinese Muslims have regarded themselves as an alien and superior people to the Chinese.

Extending eastward from the Tibetan ranges and plateau, the Tsinling Mountains form a dividing line between border groups of Ural-Altaic linguistic affiliation to the north and Sino-Tibetan, Tai-Kadai, Austronesian (Malayo-Polynesian), Austro-Asiatic, and perhaps independent linguistic affiliations to the south. Most of the non-Chinese ethnic groups found along the borders of Tibet and China speak a language affiliated to the Tibeto-Burman branch of the Sino-Tibetan linguistic family.

The Lolo (or Yi) and Min-chia have occupied the highland area of southwest China for over 1,000 years. A number of Lolo traits—among them a language that indicates an affinity to Tibeto-Burman, a tall stature with a high-bridged nose, the use of felt, a mixed pastoral and agricultural society, and the use of poisoned arrows—appear to indicate a culture with diverse and distant origins. The Lolo seem to be most alike in culture, language, and physical type to the peoples of southeastern Tibet. They appear to have originated in the same borderland that they now occupy, although perhaps more to the west.

The Min-chia have been a part of a centralized political system since the 7th century, when their principal city, Tali, was the capital of the Nan Chao state. The Yüan dynasty emperor Kublai Khan conquered Nan Chao in 1253, but the area around Tali is to this day primarily inhabited by the approximately 1,130,000 Min-chia. The Min-chia have been rice cultivators for over 1,000 years and now are highly acculturated to Chinese ways of living. Many speak Chinese as well as Min-chia.

The linguistic affiliation of the Miao and Yao to major linguistic groups is still in dispute, but there is no doubt but that the two are linguistically related to each other and have borrowed heavily from Chinese. Miao names for themselves are Hmong, Hmung, or Hmu, followed by a further word indicating subethnic affiliation. The Yao refer to themselves as Kim Mien, Kim Mun, and Yu Mien. The Miao number about 5,000,000 in China and 400,000 in Southeast Asia. The Yao number more than 1,400,000 in China and more than 200,000 to the south. Both derive from the Yangtze Valley region and, in the case of the Yao, probably from the mountains along the east coast of China. They appear to have been recognized by the Chinese as ethnic units for over 2,000 years. Both are distributed for the most part in small enclaves sometimes no larger than a village, although there are also larger settlements. The People's Republic of China has recognized a number of Yao autonomous counties (*hsien*), and an autonomous special administrative district (*chou*) has been set aside for the Miao in Hunan Province, as well as two autonomous regions in Kweichow Province. The sociopolitical fragmentation of the Miao and the Yao has apparently been caused by their countless attempts to escape domination and assimilation by the Chinese. The last independent Miao group disappeared at the end of the last Miao rebellion in 1870. Both Miao and Yao have drifted southward, and now both groups are found as far south as Thailand.

Ethnic and language groups in Taiwan

The original inhabitants of Taiwan (Formosa) were Austronesian (Malayo-Polynesian) speakers, but, within the last 300 years, after the establishment of trading posts on the island by the Dutch, the Austronesian speakers have been gradually pushed back into the mountains, in part by Hakka Chinese from Kwangtung Province but mostly by Minnan (Amoy) Chinese from Fukien Province, directly across the Formosa Strait. There are now about 16,000,000 Taiwanese. The majority speak one of several dialects of the Fukienese languages, a Sinitic tongue. A minority speak Hakka, another Sinitic language. Many mainlanders in Taiwan speak Mandarin Chinese. Within the last 25 years many Chinese have come to identify with the family, associational, and professional ties that they have established on Taiwan. As a result they and especially their children have come to associate more with Taiwan than with China.

Japanese is a composite speech in which Austronesian was the earliest ingredient. A later Altaic influence came to overlay Austronesian forms, and many Chinese words have subsequently been borrowed.

Korean linguistic affiliations seem to parallel those of Japan but with different emphasis. Altaic, specifically Tungusic, linguistic influences are very strong, and cultural influences from the north and east have been more prominent. The Chinese, for example, conquered Korea and established military colonies in the 1st century BC, one such outpost lasting until the 4th century AD.

The dominance of China

By definition the peoples of East Asia turn toward Chinese culture but with varying degrees of ambivalence. Perhaps the Koreans, Manchus, and Min-chia have been the more assiduous students. One can contrast the attitudes of these nationalities with the open defiance and antagonistic acculturation maintained by the independent Lolo (Yi) of the Ta-liang Shan and the belief of the Japanese from the 10th century on that they should be treated as equals

rather than clients by the Chinese. On the other hand, such groups as the Miao, Yao, and some Lolo speakers lost their ability to control their affairs to any large extent beyond the village level, thus tending to accept the Chinese great tradition. In early China may be seen an approximation of the feudal system of Europe. Men who held political power and an authoritative position were overlords and not simply bureaucrats. The political structure was that of a lineage state; that is, it was considered that everyone, or at least everyone of importance, within the state was related. Ideally, therefore, a vassal was not only a subordinate—a client—but a relative as well. The lord should be a more direct descendant, a member of a senior generation, or in some way more closely related to the founder of the line than the vassal. If the kinship relationship between two men who wished to establish a vassal and lord relationship was not close enough, it appears that genealogical charters could be altered to fit social and political needs. The regularized system of interstate relationship, called tribute in the literature, between China and other countries was modelled upon a system of exchange between patron and client that existed in feudal China. According to the world view by which the tribute system operated, there was only one true universal culture—that of the Chinese. The tribute system was the means of emphasizing this fact. It consisted of a number of ceremonial acts, the most famous of which was the full kowtow, in which the ambassador from the foreign court prostrated himself full length in front of the emperor or the latter's representative. He thus expressed not only submission and subservience in military and civil affairs but also obsequious deference in cultural, ceremonial, and ethical matters. The tribute system was a means of carrying on diplomacy between the dominant power—that is, the "Middle Kingdom"—and other less equal powers and civilizations. Tribute was mostly an exchange of gifts, the Chinese often giving more in value than they received. The important element was the bringing of the various peoples together under the common umbrella of customary law and ritual, or attitudes toward virtue. The adoption of the Chinese ideographic system, calendar, Classical literature and art, court ceremonies, methods of government and law, and even dress and music were further aspects of the Chinese cultural hegemony.

The tribute system

The crucial period for the development of China and also East Asia was after the fall of the Han dynasty, about AD 220. Han society was comparable to Rome in its general complexity of organization and in its decline into equally fragmented pieces. Yet a central power was to reappear in the 7th century. The reasons for this revival were several, but the most important was that although the barbarian areas, especially those that threatened China, could fight well and at times conquer China, they never emulated it to the same extent that the Western barbarians emulated Rome.

There is some agreement that the pre-modernization periods in East Asia can be called "traditional" ("classical" has a different connotation and should refer to a period preceding the traditional). The traditional period has lasted a long time, not without change but with essentially the same types of social structures. In China the traditional period began during the T'ang dynasty; that is, during the 7th century AD. In Korea it began with the Yi dynasty in 1392 and in Japan with the Tokugawa shogunate in 1603.

Kinship patterns. The common bases upon which the luxuriant growths of East Asian social structure grew are few in number and relatively simple. One is the tendency to maintain extended patrilocal families; that is, for sons to bring their wives to live within the same household or, if not, at least close enough for all descendants to form a single social, economic, and ceremonial unit during the lifetime of the father and very often the mother. Due to family or sometimes circumstances of wider importance, it may not be practical or even considered desirable for sons to live and work together, but almost always one of the children will be on hand to care for the aged parents. Sometimes, if there are no sons, a man will marry into the family, the husband of a daughter living with his wife's family.

Whether it is a woman or man who marries in, the effect seems to be the same. The spouse, or his or her children, tends to establish an identity for inheritance and descent through residence in the house into which they marry or into which they are born.

Another basic element of East Asian social organization is the ego-centred kinship system. In such a system an individual does not place himself within a kinship group by reference to a group in which he has membership but rather calculates the distance of each relative from himself. He is thus the centre of the kin group, and those too far on the periphery cease to be relatives.

Main characteristics of the kinship system

This ego-centred system has two forms. One is the kindred form, the type of kinship system found in the West, in which relatives on both the mother's and father's side are equally considered as kin. The second form may be regarded as an asymmetrical extension of the kindred form. This is the patrilineal segmentary lineage, which emphasizes the relationship between the parent and child on the father's side of the family. The line of descent from grandson to great-grandfather is thus given prominence.

Characteristic of the segmentary lineage is a tendency for residence to equal kinship affiliation. Thus, most of those who live within the same village or encampment are believed to be related. Members may move out to form colonies, however, and in this way lose their original identity. Other individuals and even families move into the residential unit, and in time they or their descendants are accepted as relatives. Within any residential area, however, there is a core group of old families, so to speak, toward whom the others gravitate and claim relationship of some sort.

The last basic element of East Asian social organization is ancestor worship, especially of one's patrilineal ancestors. Shamanistic religious practices often related to ancestor worship are prominent.

Apparently most of Europe and northern Asia at one time maintained a social organization of the type depicted above. It is also found in northern and eastern Africa. East Asia is exceptional because not only have many of the features of this model been maintained until recently, even down to the present, but because they were also accompanied by the rise of a centralized state.

Sociopolitical organizations. As stated above, Chinese civilization began as a lineage state (one in which political subordinates were ideally relatives of inferior rank in the lineage) and one in which the patrilineal segment of a lineage was of particular importance. Japanese political structure was also a political system tending toward a state based upon the segmentary lineage. It is almost impossible to maintain a strong centralized system of government along with a segmentary lineage social organization because forces for centralization are countered by those for segmentation. In China this "familial feudalism" came to an end with the Ch'in dynasty (221 BC), when a bureaucratic, authoritative system of government was introduced for all China. Although China subsequently became fragmented, indeed for a time became a feudal society, it never reverted to the lineage state as a form of government.

Japan entered into its bureaucratic period under the impact of Chinese influence, which was at its height during the Taika reform era beginning in 645. By the 11th century, however, during the Fujiwara period, local family-centred warrior bands held Japan in de facto control. With the rise to power of Minamoto Yoritomo (died 1199), the welding into a national state of local centres of power based on residence and descent groups reached its climax.

Korean sociopolitical organization managed to combine a number of seemingly conflicting factors as late as the 12th century: a bureaucratic structure, in part oriented around a Chinese-like civil service examination system, combined with a very class-conscious society in which local power rested largely in the hands of aristocratic estate holders. The residents on these estates—slaves, serfs, and nobles—were all united by common genealogical charters. Unlike Shang dynasty China, however, these lineage estates did not integrate to form a lineage state. Linguistic evidence for medieval Korean social organization can be found in the use of place-names along with patronymics.

There are, for example, 84 place-names that qualify the Kim surname.

The traditional social structure in all three nations is similar in one respect: although segmentary lineages continued to exist, they were unable to dominate and determine the organization of the state. Nor can these be called feudal societies.

Significance of the tax system

Taxation and tax reforms and their relation to land tenure were important factors in traditional Chinese society. Reforms were initiated by the Northern Wei dynasty, which controlled North China from AD 386 to 534/535. At first, the reforms consisted of limitations to the amount of land any individual could control and of qualifications on inheritance rights. During the T'ang dynasty (618–907) this system was changed to a direct and, as compared to the past, heavy tax on land. The emphasis shifted from taxing each person equally, as in a poll tax, to a tax on resources. As a result, large estates apparently were broken up into small, independently owned farms and relatively small landlord-owned estates.

A second factor of great importance in shaping Chinese traditional society was the commercialization and urban orientation that developed in China. The percentage of persons living in urban centres in China even today is small, but by at least the Ming dynasty (1368–1644) and probably already by the Sung (906–1279) the influence of city life upon the nation was very strong.

Gentry society. This period in Chinese history is usually referred to as that of a "gentry society," a term of considerable ambiguity. On the one hand, one can speak of the scholar gentry or, more accurately, the degree-holding gentry—those called *shen-shih* in Chinese. Such individuals were granted a special legal as well as social status because of their success in the examination system. Some few members of this group, by continuing up the degree ladder, received bureaucratic appointments. These may be called the bureaucratic scholar gentry. Then there were others who, because of their position as landlords, may be termed landlord gentry.

To understand traditional Chinese society one should translate gentry simply as important and influential persons and then attempt to understand in what manner their pre-eminence was obtained and by what means their influence was exerted. Government service, the examination system, education, landed wealth, wealth in general, tenantry, and social mobility can best be examined as interrelating within a system of government, as well as of society. The summit of the system was the Imperial court supported by a minute elitist bureaucracy. The court and civil service system were organized as though the government was an autocratic, although benevolent, despotism. In a formal sense this was true. There was little an individual could do to counter the will of the central government if he met with its full displeasure. It is not fully understood how the theoretical or formal structure of the government, which can be deduced from examining old statutes, actually operated. It is, however, evident that the government of the various dynasties depended upon a client–patron system of structured relationships and on informal government to maintain stability in the countryside.

The personalization of contractual relationships

By late Ch'ing (1644–1911/12) and Republican times this relationship was called *kan-ch'ing.* To have good *kan-ch'ing* is to have a sense of well-being, a feeling of being at ease, especially in relationships between parties of unequal status who are not kinsmen. Those who have good *kan-ch'ing* are in a position to help each other rather than to compete with each other. In other words, it is a personalization of otherwise contractual relationships.

Landlord–tenant relations. Despite the superficially despotic central government, the landlord was primarily dependent upon the goodwill of the tenant if he was to get his rent. It is true that, should the tenant refuse to pay the rent, the landlord could ask for help from the magistrate, but such aid, in the short run at least, might prove to be very costly, as the landlord would be saddled with all court costs besides having to satisfy those greedy lower officials upon whose goodwill he had made himself dependent. Thus it was better for both parties, tenant and landlord, to keep the central government at a distance—especially military officials. The tenant, of course, normally paid the rent. For him not to do so and present no excuse would be tantamount to rejecting the landlord's right to the land. But the actual amount to be paid was open to negotiation. If drought, flood, too much rain at the wrong time of year, insect plagues, or other factors adversely affected the harvest, it was the landlord's obligation to remit as much of the rent as necessary for the tenant to live. The tenant was in a better position to know the actual growing conditions and the real condition of the harvest. If good *kan-ch'ing* existed, the landlord would not press overly for the rent, and the tenant would be fairly honest about paying it in good years.

The lack of security in rural areas meant that good relations between landlord and villager were necessary to protect the landlord against bandits. In an area in which the gentry were well respected, they were able to maintain an effective militia (mostly composed of villagers) for the defense of the village and their own property. But if the villagers hated the landlord, residents of the rural area where he lived might offer him no protection against bandits coming from a neighbouring area or might simply attack him themselves.

More symbolic contributions made by tenants to the landlord consisted of gifts of food and produce presented at festivals. During rites of passage (at birth, marriage, death, etc.) in the landlord's house, tenants as well as freeholders would contribute labour. The landlord would make monetary contributions to his neighbours when they needed help. Tenant children often worked as domestics in the "big house," and sometimes a landlord would take a tenant girl to be his "little wife" (second, subordinate spouse). The tenant was courteous to the landlord, calling him "master" or "elder."

In addition to forgiving rents at times of distress to the tenant who fulfilled his obligations, the landlord acted as an intermediary between the rural area and the functionaries of the central government. He had the poise, education, and background that made it possible for him to argue the case for an accused villager or attempt to have the tax rate adjusted or obtain a *pai-lou* ("memorial") in honour of a faithful widow or earnest local historian. It was also expected of landlords to obtain financing for bridges, schools, temples, and other semipublic buildings. With the aid of the village headman, the landlord organized the maintenance of public works by means of corvée labour (labour exacted in lieu of taxes). At the village level this was probably more like a series of days of cooperative community effort than unrewarded drudgery. This unofficial body of village headman and landlord was also responsible for morality in the community. They often helped to work out a compromise in a dispute or even acted as judges. During the Ch'ing dynasty they were responsible for bimonthly public readings of the emperor K'ang-hsi's "Sacred Edict" (*Sheng yü*):

The "Sacred Edict" of K'ang-hsi

> Be filial and respect the social relationships, be frugal and diligent, esteem scholarship and eschew unorthodoxy, be law abiding and pay your taxes.

Sometimes local children were taught along with the children of the "big house" by a tutor, or a school was established elsewhere for the neighbourhood, usually at the Classical (Confucian) temple. These illustrations indicate only some of the functions of the client–patron system of government.

The burdens of this system lay initially upon the tenant but tended to drain the resources of the landlord. It is difficult to imagine a landlord maintaining good *kanch'ing* and making a great deal of profit from his landholdings. But, by investing in urban-based enterprises, such as pawnshops, grain shops (in which a surplus could be sold at times of greatest demand), and small industries, the gentry could recoup its losses.

It was difficult to become wealthy through the ownership of land alone and thereby to acquire gentry status because of the rules of equal inheritance by all the male heirs. Wealth was often initially acquired in the cities or, if acquired in the countryside, frequently by such dubious

means as banditry. Later this wealth was transferred to land both for economic security and for social prestige.

Families and lineages. The extended families of the "big house" were complex organizations including often only remotely related individuals seeking patronage. The peasant families, on the other hand, tended to be small, approximating a nuclear family (parents and children), although the large, extended pattern was the ideal. For many the cyclical joint (nuclear-extended) family corresponded closest with reality. This pattern, which is a basic feature of East Asian social organization, assumes the dissolution of the extended family after the death of the parents. To the individual the pattern is cyclical, as he is born into a household composed of grandparents and their married sons, which breaks up into nuclear households upon the death of the grandparents. As sons mature it is recreated, only to dissolve with the father's death.

In South China, especially in the provinces of Kwangtung, Kwangsi, and Fukien, the client–patron relationship was solidified by the continued power of segmentary lineages, the *tsu.* In Taiwan large lineage organizations were rare. Most individuals of Chinese descent resident on the island maintained only shallow lineage organizations and remembrances of ancestors.

The kindred form of kinship organization was an element within Chinese and thus Taiwanese kinship organization, but it was not as important for the Chinese as it was for the Koreans and certainly not as important as it was for the Japanese.

Korean Yi dynasty society developed a gentry similar to that of China, but, while in China the importance of the gentry lay in their function as community leaders in villages where they owned only a small percentage of the land, in Korea the gentry (known in Korea as *yangban*) owned most of the land. Koreans possessed a stronger class consciousness, and land ownership was limited mostly to the bureaucratic class.

In recent times at least, the Koreans have tended to practice primogeniture rather than equal inheritance by all sons. The joint family described for the Chinese has thus not been common for the Koreans, although the eldest son, after receiving his inheritance, is responsible to some extent for the welfare of his younger male siblings. The result is a weakly united extended family, a family that does not live together but still feels responsible for its members. The system of kinship terminology indicated by Korean terms of address is much like that of western Europe. It indicates the importance of the nuclear family and bilateral (paternal and maternal) kindred, while the Chinese terms of address indicate an emphasis on patrilocal residence. Other systems of kinship terminology in Korea show the importance of patrilineal ties, as well as the former presence of segmentary lineages.

Tokugawa Japan (1603–1867) was a highly bureaucratic state working within a feudal framework. During this time there were probably as many civil servants in Japan as there were in all of China at the same period, despite the difference in size and population. Admittance to the bureaucracy was limited to the former warrior class, Samurai or *bushi,* who were placed on stipends in times of peace and whose caste position was protected by law.

Because of the exercise in Japan at this time of tighter political controls, especially vigilance by the Tokugawa and daimyo (hereditary feudal lord) for signs of disloyalty, which might be punished by confiscation of land, there was less need for a client–patron system to emerge, although the Japanese form of the segmentary lineage, the *dozuku,* only gradually declined in importance. A client–patron relationship existed naturally between the daimyo-executive and *bushi*-bureaucrat, but this relation varied from area to area. It tended to be thought of more in terms of formal obligations and efficiency rather than as one based on personalities. While in China there had been a personalization of contractual relationships, in Japan there was a ritualization and formalization, if not to say contractualization, of personal relations.

The Tokugawa period was one of increasing urbanization and commercialization. There was a development of small industries and an improvement of agricultural techniques, based in part upon specialization and the partial commercialization of agriculture. It became more profitable to work the land in small nuclear family groups now that farm management demanded more care and a growing demand for labour existed in the urban centres. As a consequence, the nuclear family came to be the dominant form in rural Japan. The position of the nuclear family in traditional Japanese society was also strengthened by the gradual emergence of a pattern of primogeniture. Primogeniture is relatively less painful in Japan because of the respect given adopted sons-in-law, who often are the younger brothers who did not inherit the estate into which they were born but whose children inherit their wives' estates.

Socio-economic change in Tokugawa Japan

The other groups covered in this section (except the independent Lolo) were dominated by or integrated into the sociopolitical organizations of other East Asian groups or cultures during the traditional period.

The Hui-hui, for example, are very similar in social and familial form to the Chinese, differing only in those aspects in which religion has been influential. One finds in China Chinese-Islāmic communities existing as compact segregated enclaves, focussing on a single mosque. There is more community spirit than among the Chinese, and the Islāmic minister, *chiao-chang,* is also the leader of the community.

The situation of the independent Lolo (Yi) of the Ta-liang Shan is complex because of the interrelation between caste, class, and lineage affiliation. Only about one-tenth of the population of the Ta-liang Shan were Black Lolos (Mosu, or Nakhi)—that is, the upper caste of "pure" Lolos. The sociopolitical organization of the Mosu was the segmentary lineage with confederations, the latter continually rearranging because of feuding over "subjects," household attendants, and slaves. The tripartite grouping into subject or White Lolo, attendant, and slave was the result of contact with Chinese. Slaves were recently captured Chinese who after several generations came to regard themselves as Lolo and became household retainers. Their master then would permit them to become White Lolo with the status of subject. Members of the lower caste were apparently not integrated as kin with the Black Lolo but formed their own patrilineages. The relations between Black Lolo and subjects and retainers were superficially relaxed and egalitarian. They ate the same food, slept in the same house, and observed the same rules of hospitality. One might say that the greatest concern of a Black Lolo was for the welfare of his subject, because, if he acted wisely, kindly, and generously, he could attract followers. This became, of course, a point of discord between himself and other Black Lolos, as the men he might attract already had ties as subjects, attendants, or slaves with other Black Lolo.

Socialization and education. East Asian traditional society was dominated by middle-aged men and women. Within the joint family it was the wife—a former daughter-in-law and later a mother-in-law—whose voice was paramount. In the outside world the middle-aged man directed affairs after years of apprenticeship. The transitions from one stage of life to the next were all gradual, except for retirement. Adulthood was not achieved by passing some initiation rite, taking a head, or getting married—and certainly not by reaching a certain age—but only by acting responsibly. Retirement marked a sharp break: within a short time a powerful and authoritative adult might be converted into a gentle grandparent.

Children were closely integrated into the household. In China, Korea, and Japan especially it was believed that they were never able to repay the debt to their parents, and thus they strove to succeed and by this means win parental approval. This was especially true of brothers who had to cooperate in the joint family but who were also competing individually for parental approval.

Economic life. *Settlement patterns and housing.* There are basically two house types in East Asia. One type is exemplified in its simplest form by the Ainu hut. Uprights support a ridgepole connected by rafters to the corners. Reed matting is used both for walls and roofing, as well as for flooring. The essential point is that the uprights, not the walls, support the roof. The other type may be

Dwelling types

exemplified by the Lolo house, which is constructed of wooden planks that, together with uprights, support the roof. The roof is made of planks held down by stones.

The North Chinese house is fundamentally the Ainu hut with elaborations. Large pillars may replace the simple uprights of the Ainu, fired bricks the mat walls, and tiles the mat roof, but the plan of construction remains the same. It is possible for an entire wall of a Chinese house to fall away without affecting the stability of the remaining sections. This is also the traditional house type of the Manchus, Hui-hui, Min-chia, Koreans, Taiwanese, Miao, and Yao. Significant elaborations include the construction of a raised sleeping platform heated by flues from the stove. The Koreans prefer a heated floor instead. The house may be enclosed within a courtyard among Manchus, Chinese, Min-chia, Hui-hui, and sometimes Miao and Yao. Or a smooth, hardened earth area, a *matang,* may be found in front of the house, as among the Koreans.

The Japanese house may be seen as a highly original synthesis of basic Ainu hut and Lolo house. Uprights are used, not necessarily pillars, so that the walls carry some of the weight. The mat flooring again suggests the Ainu type.

In the rural areas, houses may be grouped in tight clusters or as isolated farmsteads. By and large the Chinese, Manchus, Min-chia, Koreans, Ainu, and Hui-hui live in tightly clustered settlements, while the other peoples have both. On Taiwan the Chinese-derived population tends to live more often in isolated farmsteads reminiscent of the settlement patterns of the United States Middle West.

Agriculture and trade. Every one of these societies is to some extent dependent upon agriculture. The Ainu have perhaps been mostly dependent upon fishing, gathering, capturing falcons for sale, trapping, and hunting—especially to obtain trade goods from the Japanese.

Most Chinese, Min-chia, Koreans, Taiwanese, Hui-hui, and Japanese practiced a labour-intensive/land-scarce type of agriculture during the traditional period in which the primary beast of burden was man.

For South China, Korea, Japan, Taiwan, and among the Min-chia rice is the dominant crop. The traditional method of growing rice requires the use of a water buffalo for preparing the soil, in addition to tremendous amounts of human energy for transplanting, weeding, and harvesting. In order to grow paddy, or wet rice, the farmer must have more control over his water supply than nature generally assures. To satisfy this need, countless small irrigation systems have been constructed throughout East Asia. Only some of the marginally arable land of the Hui-hui and adjacent regions in China's Northwest are totally dependent on irrigation. Here waterwheels, placed along the riverbank and canals, are used.

The Miao, Yao, Manchus, and Lolo made more use of animals and undomesticated plants. The Lolo depended as much on pastoralism as on agriculture. The Manchus definitely depended primarily upon agriculture, but at nearly every point a beast of burden assisted man in working the soil. Although the Manchu today grow the same crops as the Chinese farmers of the North China Plain (corn [maize], kaoliang, wheat, oats, millet, and buckwheat), their fields are worked by oxen or horses. Goods are transported in carts pulled by animals. Animals are fed grain rather than left entirely to scavenge. Only in the kitchen garden does Manchu agriculture resemble the intensive cultivation of the Chinese.

In all of East Asia during the traditional period a commercial and monetary economy was manifest. Farmers produced more than was necessary for subsistence, selling their surplus to buy goods at a market or to pay taxes and rents or repay loans. It is difficult to arrange the traditional economies of East Asia along a scale, but certainly by the beginning of the 19th century Japan was the most commercialized and industrialized. The power of the Chinese gentry and the centralized state, always jealous of their prerogatives, stifled industrial development in China. The same was apparently the case in Korea, whereas Japan's postfeudal society was in this respect more like that of Europe.

Degrees of economic development prior to modernization

Religion and art. The syncretic religions found throughout the Far East developed through the interweaving of several belief systems. Ancient contributions were shamanism—that is, belief in an unseen world of gods, demons, and ancestral spirits responsive only to the shaman, or priest—and a belief in an impersonal supernatural force. In anthropological literature this force is often called *mana,* but in the Orient the term used is *tao,* the "way," or "lode" as its more fundamental and ancient meaning. This force has two aspects, an active and a passive, in Chinese known as *yang* and *yin,* respectively. Perhaps such beliefs derive from the areas to the south, but such a conceptualization of nature could also have derived from the philosophical underpinning of shamanism. A third ancient element is ancestor worship. Important later ingredients have been Buddhism and, in Korea, Christianity, which has been closely associated with Korean nationalism.

Features of East Asian shamanism

Twentieth-century Tungus (including Manchu) shamanism is probably as close to the ancient shamanistic foundation of East Asian beliefs as can still be found. There is a tendency for a shaman to represent his own lineage, although a shaman does not have to limit his services to his kin group. A feature of shamanistic practice is the séance, in which the shaman sends one of his souls out to meet with spirit helpers whose job it is to induce the lost soul, or fraction of a soul, of a patient to return. The séance demands both control and ecstasy. It is a dangerous occupation, psychologically demanding. In one variation, the shaman, instead of sending out one of his souls, becomes "possessed" by the spirit he has contacted. Much of the popular religious Taoism of China uses shamanistic séances of the latter type. Sometimes, a priest—that is, a person primarily concerned with ritual and doctrine rather than with personalized contact with spirits—is prominent, as in Japanese Shintō.

In religious terms *yang* is decidedly better than *yin.* Chinese saints such as Kuang Ti (supposedly a general who lived during the period of the Three Kingdoms, about the 3rd century after Christ) had an abundance of *yang.* Graveyard trees, the ghosts of those who died in childbirth, or those without male heirs have too much *yin.* The latter, called *kuei* in Chinese, supposedly caused much of the misfortune in traditional China. It therefore behooves people to propitiate the *kuei* at the same time that they take care of their own ancestral spirits. If the *kuei* cannot be placated by offerings made at crossroads or by a special festival celebrated at midsummer, then they may hopefully be frightened away by "hot noise," a literal translation of the Chinese word for happiness, with firecrackers, red-coloured paper, and by chanting Classical literature.

The three-part soul of one's own ancestor requires care after death lest it become a *kuei.* Generally, one's own ancestors are remembered by burning incense at the family altar. Ancestor worship is very ancient among the Chinese and probably among some of the other peoples in Asia, but it apparently became important in Japan only after the importation of Chinese culture. Before that time the Japanese worshipped only the deceased founder of the lineage. The reverence bestowed upon the emperor may be seen as deriving from this custom.

The social and intellectual life of the peoples of East Asia has affected their interpretation of the nature of religion. The Chinese learned gentleman has tended to be an agnostic, whereas his Japanese counterpart is generally a Buddhist of the Zen sect. In China and Korea there has been both a Confucian activism and a Taoist passivism. In Japan, on the other hand, Zen encouraged its *bushi* adherents to find their way within the world rather than by withdrawing to some lonely and beautiful spot, as in the Taoist ideal.

Aesthetically the Chinese gentry and their Japanese counterparts expressed themselves along similar lines. Fundamental to their visual art is the skill perfected in practicing calligraphy with brush and ink stone. The novel, a literary form dependent on a "middle class" having leisure time and an interest in details of personality rather than simply in characters of epic proportion, developed much earlier in both China and Japan than it did in the West. Stylized operatic theatrical productions, the *Ching-hsi* of China and the Nō of Japan, were enjoyed by the literati but were in form and content simply the most polished of a series

of musical theatrical productions found throughout the area in the form of shadow plays, puppet shows, and tales of itinerant storytellers. It is in pornography that a difference may be found, the Japanese celebrating an exuberant phallicism, the Chinese being more sensual and playful.

Modern developments. For the peoples of East Asia the 20th century has been a time of quickening change. The Taiwanese have emerged as a distinct ethnic group and nation, whereas the Manchus, once the rulers of China, have become eclipsed. By the time of the fall of the Ch'ing dynasty in 1911/12, most Manchus had become so Sinified that they were simply absorbed into the Chinese population, though some pockets of Manchus (the Bannermen) who had developed a strong sense of national identity and who resisted absorption by the Chinese remained. The establishment and sponsorship in Manchuria (Northeastern Provinces) of Manchukuo by the Japanese in 1932 gave such groups a respite. Now three generations removed from the fall of the Ch'ing dynasty and one from the unification of China under the Communists, there is probably little left of the Manchus as a distinct people.

The Ainu of Hokkaido have in a sense been fighting a losing battle against the better organized Japanese since the 1st millennium of the Christian Era. The Ainu, until Russian penetration of the Far East, were valued by the Japanese as a market. The Russian advance led to closer Japanese control over the Ainu and, after the Meiji Restoration (1868), to much closer supervision of Ainu life. The Ainu were encouraged to become sedentary farmers rather than moving on from plot to plot. This often led to their neglect of fishing and other traditional occupations. Large numbers of Japanese have since settled on Hokkaido, which has led to intermarriage, and for over 100 years now the Ainu have been encouraged to become successful within the framework of modern Japan. Despite discrimination by some Japanese, especially up to the time of World War II, the Ainu have accepted assimilation into Japanese society. To what extent the old ways will continue to be observed by some families cannot be ascertained, but as a distinct people the Ainu may soon disappear except in the ethnographic record.

The fate of the Min-chia was sealed much earlier, in the Yüan dynasty (1206–1368), when Kublai Khan conquered the Nan Chao state. The Min-chia were encouraged to become Chinese literati, and they have since then competed for Chinese honours with even greater fervour than the Chinese themselves. They achieved a high measure of success in Chinese society so that over the centuries they have become very Sinified. At the present time no political, social, or cultural situation of a nature that would slow the Min-chia's rapid assimilation into the Chinese population seems likely to come into being.

Ethnic awareness of the Muslim Hui-hui in China

One might assume that the Hui-hui are also on the way to losing their distinct identity, but in fact the victory of the atheistic Communist Chinese has heightened the ethnic awareness of the Islāmic Hui-hui, who have maintained their religious bonds. During the last quarter of the 19th century they were the victims of extreme religious persecution, in the course of which millions of them were killed. Now only tolerance and a real end of discrimination by the Chinese will lessen the Hui-hui's resistance to being assimilated. They have apparently been allowed to play a part as a nominally autonomous group by the Chinese Communists.

In one sense the Ta-liang Shan (independent) Lolo (called Yi by the Chinese) have created one of the more successful countercultures to the Chinese. They have maintained their independence and identity during 2,000 years of Chinese pressure. During this time they were able to elaborate a social and ceremonial organization of their own and to develop a non-Chinese and non-alphabetic writing system. But also during the years, other Lolo groups, which are closely related to the independent Lolo in speech and culture yet living outside the mountain fastness of the Taliang Shan and under Chinese jurisdiction, have increased in number until there are about 5,450,000 of them, not including the Nakhi, Lisu, Lahu, and Akha. There are, however, far fewer independent Lolo left: Chinese control is obviously gaining.

The so-called acephalous political organization (*i.e.*, lacking a governing head or chief) of the Lolo, Miao, Yao, and other groups in southern and southwestern China has given them certain advantages. The primary advantage lay in the inability of the Chinese to be done with them by one decisive battle or campaign. But it has proved a disadvantage as well, as it has left them unable to deal with the gradual encroachment of the Chinese onto their lands, first as individuals, later as officials. Only by enslaving the Chinese and destroying their settlements were the Lolo along the Szechwan–Yunnan border able to hold their own. Once deprived of the strength to carry out such drastic measures and forced to deal with the powerful government of the People's Republic of China, the Lolo, like the other non-Chinese peoples of southern and southwestern China, will probably find within a generation or two that all that remains of their ethnic uniqueness are some folk-art motifs and a few officially sponsored folk festivals.

Paradoxically, the 50,000,000 non-Chinese of southern and southwestern China have been able to maintain their own culture and language for thousands of years; yet, now that a government of China exists that pays special attention to them and that is sometimes sympathetic to them, they are likely to become totally acculturated. The reason for this is that they live scattered in small enclaves among the Chinese population.

Relations between the Taiwanese and the "mainlanders"

As pointed out already, the Taiwanese control most aspects of their lives and, through their domination of the provincial government, the careers of many "mainlanders" on the island. The latter are attempting to re-Sinify the Taiwanese by teaching Mandarin Chinese in the schools, by the cultural-restorations movement exemplified by the building of a beautiful new national museum, and by other government services and programs. But the fact that the Taiwanese own the land and control the resources, plus their greater numbers, suggests a different direction of development. If Taiwan is allowed to settle its own affairs, it is more likely that the Taiwanese will absorb the "mainlanders" than vice versa.

China and the West. Traditional East Asian culture, based on the tribute system and the respect for Chinese intellectual values, ended when it proved impossible to ignore the West and when the West would not agree to be just a somewhat different type of barbarian to be incorporated into the established order. Had the demands of the West been only those of imperialistic barbarians, they could have been dealt with in the classical manner. As it was, the dominant powers of the Far East had to see themselves as students instead of teachers. To accept this meant the destruction of the mythology underpinning the tribute system. The Westerners refused to kowtow and insisted on recognition at the national level. They defeated the Chinese militarily but did not thereupon attempt to conquer China and set up a new dynasty. Within a short time China lost suzerainty over Korea, over the Tungus on the northern watershed of the Amur, over the Khalkha Mongols, the Tibetans, the Vietnamese, the Ryukyuans, the people of Tannu Tuva, and the majority of Kazakhs and Kirgiz.

The response of the Manchu court to this situation was an attempt at antagonistic acculturation; that is, an attempt at learning the technology of the West without accepting Western values or allowing any changes in Chinese social, political, or economic organization. This response allowed the court to survive during the initial stages of the revolutionary period, which began in China partly because of the weakness of the court and its foreign origin but also because of a growing awareness that the gentry system was inadequate to supply services felt to be essential to a modern state. The use of modern arms at first enabled the court to put down rebellion, but by the 20th century the buildup of foreign mercantile and industrial centres in the major coastal cities and the activity of Western missionaries in the interior had awakened a growing desire for modernization and a reassertion of China's identity. It was only natural that the original leaders in the anti-Ch'ing movement should have come from those areas in closest contact with the West. The government tried to meet their demands by changing from the Classical system

of education to one stressing Western thought, by allowing the development of industry, and even by accepting the idea of a constitutional monarchy—in fact, abandoning the earlier attempt at antagonistic acculturation. But the anti-Manchu sentiment that had been released was too strong to halt the fall of the Ch'ing dynasty in 1911/12.

China from the fall of the Manchus until 1950 found itself bound by a number of severe contradictions. Those who had given the revolution intellectual force stressed individualism, Western intellectualism, science, universal education, democracy, national reconstruction, and an improvement in the peoples' livelihood, as well as other liberal and popular measures. These measures were all suited to the needs of the times. During previous periods of Chinese history, such as the Ch'in (221–206 BC) and the Sui (AD 581–618), the Chinese had been quite able to rework their institutions drastically. The new reformers, however, had no real power base in the countryside. Their strength lay in the Western-influenced treaty ports and was suspect to the majority of Chinese. Those who soon came to have real power, the warlords who from 1912 to 1949 controlled most of China independently of the central government, did not understand the nature of the changes that were needed. Most of the warlords had abandoned Classical ideals, it is true, but they had not replaced them with those of the West. In fact, they were primarily opportunists or chauvinistic nationalists. The head of the Kuomintang, Chiang Kai-shek, reflected the desires of the nation to reconcile the different factions, and he might have been successful in doing so had the Japanese not sought to prevent the emergence of a modern China.

In some areas modernization did occur. There was limited industrialization; a socially conscious literary movement exemplified especially by the *pai-hua* (vernacular-language) movement led by Hu Shih and the bitter essayist Lu Hsün; the building of railroads and motor roads; the organization of one of the more efficient postal services in the world; and a nationwide, although not compulsory, school system. But the basic problem was that the leaders lacked the power to deal with the patron–client relationship that had held the traditional society together. As the warlords of China increased services, as industrialization continued, as classical principles well adapted to the maintenance of good *kan-ch'ing* were abandoned in favour of individualism, and, most importantly, as the old-fashioned gentry lost their ability to intercede for the welfare of the community, the traditional institutions became means of exploiting the farmers. This was the case because, as the power of the government increased, the restraint on the landlord that had made it essential that he keep the respect and goodwill of his neighbours and tenants was removed. The government gave the landlord protection from bandits of whatever derivation and assured the collection of rents. It became no longer necessary for the landlord to maintain a good standing in the community in which he owned land. In the area near major cities in which this process had evolved furthest, landlords, now often urban industrialists, divorced themselves entirely from the problems of the land and appointed an estate manager who collected a set amount of rent and kept whatever else he could squeeze out of the farmers.

Breakdown of traditional relationships

The client–patron system had functioned amazingly well in traditional China. Few such holistic systems, encompassing political, educational, artistic, social, ethical, and economic goals, have lasted so long and with such vigour. Its major weakness—its lack of consistent service (in regard to the provision of regular admittance to schools, for example) and justice—was also a major virtue, as it accompanied a government that lay lightly upon the citizenry. But it became impossible to retain only a part of the system—namely, privileges for the gentry and nepotism in government. These were not destroyed until the Chinese Communist Party had fled to the countryside to lead a revolution in the classical Chinese manner.

The Communists' main thrust was a reorganization of society from the village level upward, with the cooperation of all "progressive" elements in an attempt to do away with "feudalism." This involved the elimination of vestiges of the client–patron system, along with many other associated traditional values, such as the hierarchical position of men in Chinese family life. Emphasis was placed upon the services individuals could offer the community and state regardless of the status of the person and his connections (*kuan-hsi*) or how those of higher rank felt about him (*kan-ch'ing*). In summary, the personalized, traditional society was considered exploitive and inefficient. Thus, the old lines of structural personalization were discarded and new ones established, such as farm cooperatives, women's clubs, literary movements, and anti-Japanese militia and crop improvement committees.

Non-Communist East Asia. Korea became gradually weaker under the Yi dynasty (1392–1910). The Yangban (gentry) took over once-public lands and then did not pay taxes. To make up for the deficit the freeholders were taxed more heavily. They could not afford the taxes and lost the land, which thereupon passed into the hands of the bureaucratic landlord class. When famine occurred the central government had no reserves to alleviate the situation. Meanwhile, the intelligentsia, unaware of the real peril, indulged themselves in factional disputes more serious perhaps than those that had contributed to the fall of the Ming dynasty in China (1368–1644). The West largely bypassed Korea, which remained oriented along traditional lines after China had long been opened up to Western trade, missionizing, and interference and after the Meiji Restoration had set Japan on a path of modernization and industrialization. The Yi dynasty would probably have been swept aside in due time by a Christian- or Tonghak- (a syncretic revitalistic religious response to the West including elements of both Eastern religion and Christianity) led rebellion if Japan had not invaded Korea first and conquered it.

The Japanese conquest increased the Korean sense of national unity and in many ways modernized the country. South Korea is one of the more industrialized states in Asia, and a land-reform program is in effect.

Japanese society at the beginning of the 19th century had the strengths of feudal values interwoven into a bureaucratic system of government. Its people were ready, however, to discard the remnants of the feudal class system. It seemed better to be a prosperous businessman than a penniless samurai.

Recent social changes in Japan

In recent times some novel social forms have emerged in Japan, in the development of which the Japanese have been able to draw upon a number of value systems. One of these has been Bushido, the hierarchical and authoritarian conception of the role and status of the *bushi* service bureaucrat, based on a mixture of classical Chinese, feudal Japanese, and bureaucratic values. With its strong emphasis on loyalty in a vertically ordered system of two-person reciprocity, this way of the samurai is one of the important values that have shaped modern Japan.

Another hierarchically structured model dominant in some parts of Japan has been the *dōzoku,* the combination of kith and kin with landlord and tenant. In other areas, relations between landlord and tenant, although influenced by the *dōzoku* model, tended toward abuse by the landlord after a strong central government had assured his rights. This led to the growth of unrest and militancy among the rural population, finally suppressed in the 1930s.

The paternalistic model of organization in Japanese industry, known as *oyabun-kobun,* is probably inspired by both the Bushido system of values and the *dōzoku* (segmentary lineage). The result has been security for the worker, full employment, and ordered advancement with a minimum of division along class lines and in the professions.

The third organizational model has been drawn from the local kindred kinship group and the cooperating local group, the *buraku* (hamlet). These local units have become especially important since a land reform that occurred after World War II brought an end to landlordism in rural Japan. These two sources have provided the inspiration for egalitarianism and a desire for consensus found among the Japanese.

The populace is also divided into associational frames of reference such as industrial firms and villages. Differences in rank and status are important, but class is not impor-

tant, except for former outcaste groups such as the Eta. Economically, the system probably works as well if not better than that of the West, because it does not tend to cause social alienation.

The interaction of these various concepts in the political arena has been a mixed blessing. Perhaps some unique adjustment—allowing for efficiency in government, the desire for consensus, and homage to rank while furthering egalitarianism—will be achieved in the future.

In the 20th century East Asia as a cultural entity, as defined and described at the beginning of this article, was badly shaken by the collapse of traditional China. At times it has seemed that Japan was destined to take China's place, but Japan's defeat in 1945 changed this. Up to the early 1980s China has directed its attention to its Inner Asian frontiers and to internal social, cultural, and economic change, while Japan, South Korea, and Taiwan have concentrated more on industrialization and on developing their international trade. Given a continuation of present circumstances, one might predict the end of East Asia as a psychological reality within another generation. (F.B.B.)

SOUTHEAST ASIA

Southeast Asia, in its broadest definition, comprises those continental margins and offshore archipelagoes of Asia lying south of China, north of Australia, and east of India. Its continental portion embraces the autonomous nations of Burma, Thailand, Laos, Kampuchea (formerly known as Cambodia), and Vietnam. Its insular portion consists of the island nation of Singapore, at the tip of the Malay Peninsula, and it can be defined to include the two sprawling archipelago nations of Indonesia and the Philippines. It also includes Malaysia, a political amalgam of two northern Borneo states and the Malay states of the peninsula. In addition, there is the independent sultanate of Brunei in northern Borneo. The region extends over 2,000 miles (3,200 kilometres) from north to south and 3,500 miles (5,600 kilometres) from east to west.

Most of Southeast Asia has a stable, homogeneous tropical climate. Its large, shallow archipelagic seas and extensive fringing bays help stabilize the region's average monthly temperatures at around 81° F (27° C).

Four vast, southward-flowing river systems shape continental Southeast Asia's topography and major settlement

Distribution of Southeast Asian peoples and cultures.

patterns. They are the Irrawaddy and Salween (Burma), Chao Phraya (Thailand), and the Mekong (marking the Thailand–Laos frontier and traversing both Kampuchea and Vietnam). The shorter, eastward-flowing Red River (Song Hong) reaches the Gulf of Tonkin farther north, near the Chinese border. All except the Salween flow through broad alluvial plains and fertile deltas, where intensive rice agriculture sustains dense population centres and large cities. No comparably large river systems exist in the islands. Closest in length are the huge meandering rivers of Borneo, the world's third largest island. The other major Indonesian and Philippine islands are, unlike the mainland, volcanic. Their topsoils support an intensive rice agriculture.

Language communities

Ethnic groups. Southeast Asia has one of the world's most ethnically diverse populations. In Burma alone, between 125 and 140 languages are spoken; in the Indonesian archipelago well over 200 distinct languages are in use. This complex ethnolinguistic mosaic, with all its cultural diversity, has been studied only in part, and major problems of descriptive ethnology and linguistics still abound. The political frontiers obviously have little to do with the human social boundaries in the region: the Laos–Thailand border along the Mekong River, for example, divides in two the large Lao ethnic community of the mid-Mekong Basin, and the northern Burma–Thailand border splits the Shan people into only nominally separate "national communities."

Many of the region's languages, though mutually incomprehensible, belong to three widespread language families—Sino-Tibetan, Mon-Khmer, and Austronesian (Malayo-Polynesian). The scattered languages of the Sino-Tibetan and Mon-Khmer families are indigenously restricted to the peoples of the mainland, while the languages of the Austronesian family are indigenous both to portions of the mainland and to the two archipelagoes. Two of the mainland's most important language groups—Tai and Vietnamese—cannot be conclusively placed in any of these families; like many of the mainland's smaller "independent" languages, their origins and ancient historical connections are still in doubt. Independent in another sense are the several important Chinese languages (Hakka, Hokkien, Cantonese, etc.) spoken by the largely urban Chinese of both the mainland and the archipelagoes; while forming a subgroup of the Sino-Tibetan family, these languages are not indigenous to the region, having recently been spread there by Chinese immigrants.

Chinese, Indian, and Islāmic influences

By the early Christian Era, Chinese and Indian merchants had already extended their trading networks well into Southeast Asia. Lured by the area's natural wealth, they had opened both overland and maritime trade routes to tap its minerals, spices, and jungle products. Into the Neolithic villages and port settlements, they had peacefully begun to introduce the skills, social patterns and religions of two major civilizations. From about AD 100 to 1500, Indian, rather than Chinese, civilization became imprinted most explicitly in the region. Chinese influence spread overtly only among the people of Vietnam. Throughout "farther India" (Burma, southern Thailand, Kampuchea, and southern Vietnam), on coastal west Malaya, in eastern Sumatra, and on Java's north coast, pioneer Indian settlements grew up along the trade routes. As they waxed commercially, they expanded their religious and political influence. Indian Sanskrit scholars, monks, and priests settled and helped consolidate these communities into regional, theocratic, and socially stratified "Indianized states," economically based upon irrigated rice agriculture. These Indianized states were variously and at different times either Buddhist or Hindu-Brahmin in their dominant religious orientation. Until the mid-14th century they struggled among themselves for political supremacy and the control of regional trade. In doing so they implanted Indian religions, legal codes, social patterns, cosmologies, syllabic writing systems, classical literature, graphic arts, drama, and music throughout much of Southeast Asia.

A second major historical influence has been that of Islām. Between the 13th and the 15th centuries Arab and Indian Muslim traders began to dominate the maritime trade with the coastal peoples of Sumatra, Malacca, the Malay Peninsula, and many of the islands eastward to Mindanao. Often settling and intermarrying among the local ruling families of these scattered principalities, they imparted a pervasively Islāmic way of life. Islām embodied a monotheistic philosophy of religious brotherhood, revealed and concretely chartered by the Qur'ān. It contrasted with the already widespread polytheistic Hinduized religion that emphasized caste-segmented ritual and a variety of beliefs. By the 17th century, missionaries and converted chieftains (sultans) had widely extended Islām along the trading routes to Indonesia's Spice Islands (the Moluccas), into the southern Philippines, and, on the mainland, even among the Cham population of southern Vietnam. While also helping to undermine Java's Indianized state of Majapahit, Islāmic culture had begun by this time to penetrate deeply—mainly in insular Southeast Asia—into the belief systems, languages, literatures, and domestic law of many indigenous peoples.

Indigenous culture traits

Despite the spread of Indian, Islāmic, and Chinese civilizations into the region, Southeast Asia's interior, less accessible peoples have retained certain widespread, indigenous culture traits from the pre-Christian era. These include slash-and-burn (shifting field) agriculture, the terrace irrigation of wet rice, and the domestic use of chickens, pigs, and water buffalo. Other such traits are a highly diversified use of bamboo (to construct houses, tools, weapons, musical instruments, containers, etc.), separate premarriage dormitories for men and women, megalithic monuments, pile-supported dwellings, and multifamily longhouses. Associated patterns have included trial marriage, the blood feud, headhunting, strong local patterns of etiquette and custom, weak institutions of political leadership, and the practice of making decisions through rituals of ordeal and divination. Most of the technologically simpler peoples of even the remote interior have acquired the use of iron and the blacksmith's skills. Yet, such has been their generally ingenious, careful, and effective use of all available plant and animal resources that the French geographer Pierre Gourou was able to describe Tonkinese peasant life in northern Vietnam before World War II as a "civilization of the vegetable kingdom."

The Burmese, Thai, Khmer, and Laotian peoples. The lowlands of Burma, Thailand, Kampuchea, and Laos are each occupied by one politically predominant ethnic group—respectively, the Burmese, Thai, Khmer, and Lao. They probably total about half of mainland Southeast Asia's population. The Burmese inhabit the valleys of the Irrawaddy, Sittang, and Chindwin rivers, the Arakan Coast, and the Tenasserim panhandle; they extend northward to about 25° N. The Thais live on the alluvial plains of the Mae Nam Chao Phraya (Chao Phraya River) and its tributaries in central and south Thailand. The Khmers inhabit Kampuchea's interior plain drained by the Mekong and Tonle Sap rivers. The Lao occupy both banks of the Middle Mekong River and extend into Thailand. Although separated by language, ethnicity, and nationality, these peoples are alike in their agrarian economy, social organization, and adherence to Theravāda Buddhism.

Settlement patterns

These mainly rural peoples live in settlements that vary from compact house clusters to the more common linear communities extending along rivers, irrigation canals, roads, or paths. The size of a settlement may range from a few families to over 3,000 people. Houses are rectangular, gable-roofed, and usually built on piles two to six feet above the ground; they range in width from 10 to 25 feet and in length from 20 to 40 feet. Most have plaited bamboo walls and floors, with thatched roofs; houses built of planks with roofs of tile or sheet iron usually proclaim wealth and high status. Better houses may contain several rooms and have shuttered windows; poorer houses often consist of single compartments opened only by doors. Kitchens, often essentially wood-framed earthen hearths with stones to support cooking pots, may be separately partitioned, sometimes occupying rear verandas. Roofed front verandas, reached from the ground by short ladders, serve among the Burmese and Thai for other household chores and for sociability. Furnishings are usually simple but may include low eating tables, bed mats, storage bins, baskets, water jars, shelves, Buddha images, household al-

tars, and spirit houses. Almost every settlement includes a Buddhist temple compound and, additionally, among the Burmese, a brick pagoda containing a religious relic. Temple compounds in the larger settlements of all these peoples may also contain monks' quarters, meeting halls, and tomb monuments.

Rice, the major crop and staple food, is grown in many different varieties, usually selected for their productivity and endurance under local conditions. It is generally grown in permanent, diked fields watered by rainfall or—as in Burma's dry zone—by irrigation. Both annual cropping and multicropping are widely utilized. Draught buffalo (water buffalo) and bullocks are commonly used to plow and harrow the fields for rice planting; animal dung, ashes, and paddy stalks are frequently, if erratically, used as fertilizer. Common planting techniques include broadcasting, dibbling, and transplanting from seedling nurseries, depending upon local soil and water conditions. Hand sickles are generally used for rice harvesting. Threshing and winnowing techniques are quite varied and include trampling by buffalo.

Supplementary vegetables, tree crops, and other crops are also widely grown, both in house-plot gardens and in rotation on multicropped rice fields. Vegetables include beans, eggplants, tomatoes, onions, garlic, chillies, cucumbers, yams, sweet potatoes, white potatoes, gourds, cassava, peppers, ginger, mint, turmeric, and other exclusively local plants. Tree crops include bananas, coconuts, mangoes, durians, betel nuts, citrus, jackfruit, papaya, guavas, tamarind, and sapodilla. Other crops, sometimes raised commercially, include rubber, cotton, corn (maize), millet, sesame, sugarcane, and various condiments. No single farming household, however, will normally raise more than a few of these items. Burma's Tenasserim area specializes in tree crops, while in Kampuchea the sugar palm (*Borassus flabellifer*) has become a national symbol because of its local importance as a source of thatching material, fruit, sugar, and juice.

A major animal food is fish—fresh, dried, or salted. Although particularly important in the Burmese diet, it is found everywhere in local trade and, especially in its dried and salted form, readily reaches even communities distant from fishing areas.

Social organization typically is based upon the nuclear-family household—*i.e.,* essentially parents and their unmarried children; there is no tradition of extended-family groups or clans. Traditional society is largely stratified in hereditary ranks according to degree of royal ancestry, aristocratic background, and even religious or military achievement. The variety of such ranking has allowed individuals to move from one rank to another and occasionally even from the bottom of the traditional social ladder—*i.e.,* from among formerly indentured people and slaves.

Religious practices among mainland Buddhists

Religion is a blend of Theravāda Buddhism and local spirit worship. Buddhist theology posits that man undergoes a cycle of separate existences and that all animate life is endowed with individual souls. The moral and behavioral guidelines of his faith enable a devout Buddhist to progress toward his goal of Nirvāṇa by acquiring merit during his lifetime. There are many merit-earning acts. They include becoming a monk, financing construction of a temple, having one's son ordained as a monk, visiting Buddhist shrines, contributing to the repair of a temple, giving daily food to monks, entering the monkhood as a novice, attending the temple on all Buddhist holy days, and generally obeying Buddhist precepts. Conversely, by flouting Buddhist precepts, one can diminish one's stockpile of merit and even risk being reincarnated as a less desirable form of life in the next existence. Devout Buddhists are enjoined from taking life and usually even refrain from killing mosquitoes. Yet, this does not prevent most from eating meat or fish. If another person kills an animal, no merit is lost by eating its meat. The fisherman loses no merit because his catch is thought to die from leaving the water and not from the fisherman's action.

The worship and propitiation of local spirits provides the complementary, non-Buddhist religious dimension. Known as *nat*s in Burma and *phi* in Thailand and among the Lao, these spirits are ubiquitous, are thought to exercise great influence over daily human fortune, and are susceptible to approach and manipulation. The Burmese formally recognize 37 major *nat*s and a horde of lesser ones; the Thai, Khmer, and Lao recognize comparably large numbers. There are spirits of one's house, village, and rice field; there are many others of lakes, rivers, forests and of objects and periods. Equally various are the rituals necessary to keep them placated and amicable. Accidents, sickness, epidemic, drought, floods, and domestic problems are generally blamed upon the malevolence of spirits. Individual households and villages maintain shrines for placating their respective guardian spirits, usually through offerings of food and flowers. Male and female mediums can, when in trance, communicate with certain spirits and ascertain from them the shape of future events.

Because entering the monkhood is a major merit-generating act, most men spend part of their lives as monks. A young man will be initiated at about 14 years of age as a novice monk and remain in the monkhood temporarily, for several months or half a year. His initiation, a ritual re-enactment of Lord Buddha's own renunciation of material wealth and assumption of monastic discipline, is a major festive occasion; the initiate's head is shaved, and he receives a new name, recites the monastic vows, and dons the monk's orange robe. During his monkhood the novice obeys the same rules governing diet, celibacy, and material possessions that discipline the senior monks. He lives in the temple monastery, begs his morning food in the community, and recites the sacred texts. He usually leaves the monastery without taking the examinations for advancement in the monkhood. He may return later in life either for another short period or to enter the monkhood permanently.

The full-time Thai Buddhist monk usually assumes an active pastoral role in the community of his home temple. The role commits him to officiate at the weekly temple services and at the monthly and seasonal rituals. Normally, it also requires that he conduct individual family ceremonies, as invited, for weddings, house blessings, cremations, and ordination of novices. He is a locally influential religious counsellor on important matters such as determining auspicious dates for personal activities. He may even study *sajjaad,* or magical healing, and thus extend his leadership into matters of health.

Since illness is often attributed to hostile spirits, its cure is not an established responsibility of the professional Buddhist monk. There are many sorts of locally recognized healers. Among the Thai, they include the spirit doctor (*maw phii*), who, through therapy such as text chanting, body massage, and the use of herbal medicines, is able to coax out, or extract, the hostile intrusive spirit. A *maw khan* specializes in recalling a patient's wandering personal soul and thereby restoring his health. Mediums can often diagnose illness by singling out the responsible *phi* and even, when in trance, contacting it directly for therapeutic advice. Such healers and diagnosticians may employ explicitly Buddhist symbols and ritual acts, while the pastoral monk, on the other hand, may utilize "magical healing" among his congregation; yet, healers and monks seem seldom to conflict. Using different techniques, they both provide some "mental-health" therapy; in addition, the healers are frequently astute both as physical therapists and as physicians.

The Vietnamese. The Vietnamese are heavily concentrated in the northern Red River Plain of Tongking and in the southern Mekong Delta area. They are settled only thinly along the coast between. In the north, where population density is very high, rice cultivation is supplemented by root crops such as manioc, taro, and sweet potatoes; in the south, more exclusively rice-growing areas predominate. Villages tend to specialize in pursuits such as metalworking, silk production, and coastal fishing; trade between such specialized communities has traditionally been important.

Chinese cultural influence in Vietnam

Chinese influence has profoundly affected Vietnamese society, setting it apart from the mainland societies farther west. Vietnamese use paternal surnames and, for purposes of ritually honouring their paternal ancestors, congregate

into patrilineal clan groups. The oldest male of the clan's senior branch usually conducts the ritual and maintains the ancestral clan shrine, its altars, and its commemorative tablets. Since this vital responsibility descends only among males, the importance of having male children is stressed more than in Thailand or Burma. Vietnamese Buddhism, formally of the Mahāyāna school, is also of Chinese origin. Far fewer men enter the monkhood in Vietnam than in the westerly Theravāda mainland countries. Those who do usually become full-time professionals; they follow primarily scholarly pursuits, do not beg for daily food in their communities, and play little or no pastoral role. Some become divination specialists, making use of Taoist ritual techniques, another Chinese cultural influence.

Vietnamese society has also incorporated traditional Confucian social doctrine binding subject to ruler, son to father, wife to husband, younger to older brother, and friend to friend in ideally respectful, permanent relationships. These ties have closely supported Vietnam's traditional hierarchical systems of civil and religious authority extending from the supreme emperor down to the village-level "council of notables." The intervening bureaucracy, mainly educated Mandarin landholders, has traditionally represented the emperor and is split into a number of ranks, each with its special privileges, powers, and status trappings. One could attain or better one's rank in the Mandarin bureaucracy through education and by passing examinations in Chinese characters and Classics. While this fostered some mobility both in rank and in residential location among the bureaucracy, the peasantry remained closely rooted to its villages. This "rootedness" is religious in nature, exemplified in the villagers' attachment to local clan shrines and also in the Vietnamese *dinh*. The *dinh* is a walled compound containing a sanctuary for the village's patron deity—often a legendary or historical personage—together with sanctuaries for lesser local spirits. While the *dinh* does not contain monk's quarters, it often provides a meeting place for the discussion of village affairs. Above all, it is a focal point for the discharge of the villager's crucial ritual obligations to the locally powerful deities and to the numerous spirits that, in traditional belief, control his world.

Montagnard cultures

The Indochinese hill peoples. In Vietnam, as throughout Southeast Asia, the lowland peoples tend to be culturally distinct from their upland neighbours, even when they belong to the same ethnolinguistic group. Vietnam's own interior hill peoples, the many groups collectively known by the French word Montagnard (Mountaineer), have traditionally been despised by the surrounding Vietnamese, Khmer, and Thai as "savages." Like the other scattered hill and mountain peoples of northern Thailand, Laos, and Burma, most are economically dependent upon shifting field cultivation and upon supplementary trade in the products of forest and of local craftsmanship. Mostly they live in scattered, impermanent settlements often located on steep slopes at high altitudes. They are predominantly and elaborately animistic in their religious beliefs and ritual practices, and their community leaders are usually endowed with some shamanistic ability as curers, diviners, or trance mediums.

Rice is the most desired food and usually also the staple crop among the mainland hill peoples. Some groups acquire it mainly by trading other crops for it, such as tea and opium. In Laos the hill peoples are generally self-sufficient in rice but also produce cane, wax, Spanish peppers, shellac, cardamom, and many other trade items; these often pass through the ethnically Laotian, river-mouth trading villages along the Mekong tributaries to their ultimate lowland buyers. In return, the mountain farmer receives such essential items as cloth, steel tools, and salt. Today, he also receives an increasing flow of Western consumer goods and even cash.

Varieties of Indonesian societies

Indonesia. *Rice farmers of Java and Bali.* The peoples of Indonesia may be divided into several categories. The first includes the irrigation rice farmers of the strongly Hinduized inland areas of Java and Bali. These two peoples comprise about half of Indonesia's total population, and their densely settled homelands reflect Indonesia's severest population pressures. But the Balinese have escaped the Islāmization experienced by the Javanese; their expressively rich religion, Ugama Bali, shows the widespread influence of classical Indian cosmology, ritual organization, and litany. Such influence may also be seen in the ancestral shrines, village temples, and courtyard structures adorned with Hindu art motifs carved into wood and volcanic stone. While both Balinese and Javanese share a common, pervasively Hinduized cultural heritage, the Balinese retain certain old Hindu caste distinctions and stress status differences more formally than the Javanese, especially those between aristocrat and commoner. By his simultaneous membership in a number of groups based on kinship and occupation, the typical Balinese villager also participates in a formally structured local community.

Islāmic peoples of Indonesia. The second category of societies includes the trade-oriented Islāmic peoples scattered along the coastal and riverine portions of the archipelago. Economically, these peoples range from full-time merchants to petty traders who supplement their livelihood by occasional gardening and cottage industry. In religious orientation, they are Muslims and share a fundamental respect for Qur'ānic tradition and law. While linguistically facile in Malay, which serves as their commercial lingua franca, they also speak a variety of native languages. The three largest constituent societies are those of the Malays of lowland eastern Sumatra and western Borneo and the mutually similar Makasarese and Buginese societies composed of peoples who stem from southwestern Celebes but are today scattered widely along the coast of the archipelago. Other small societies include hamlets, villages, and port-town enclaves of mixed Javanese, Arab, Portuguese, Indian, Chinese, Dutch, and other ancestries.

Marginal peoples. A third category of Indonesian societies comprises mainly isolated peoples unaffected by Indian or Islāmic influence. Their economy has consisted mainly of shifting-field rice cultivation or, as in extreme eastern Indonesia, the gardening of sago and various root crops. By the late 19th century, several of these peoples had experienced Christian-missionary influence. Until the 20th century, many retained their traditional customs of sporadic intervillage warfare and headhunting. Their traditional religious beliefs were richly animistic, with conspicuous emphasis upon ancestor-revering ritual. Included in this category are the Toradja of central Celebes, the scattered Dayak peoples of Borneo, a number of interior peoples of Nusa Tenggara (principally the islands of Lombok, Sumbawa, Flores, Timor, and Sumba), several smaller Sumatran groups (Gayo, Rejang, Lampung), and certain peoples of Seram and the Moluccas in extreme eastern Indonesia.

Major Sumatran peoples. Three other major peoples inhabit Sumatra: the Achinese of the island's northern end, the interior Batak somewhat farther south, and the upland Minangkabau of west central Sumatra. None has experienced the extensive Indian cultural influence felt in Java and Bali. Yet, historically, all have adopted foreign religions: the Achinese and Minangkabau have long been Islāmized, while the Batak have been significantly and increasingly influenced by Christian missionaries. All three peoples depend heavily upon rice agriculture (both wet and dry) and share western Indonesian patterns both of language and of material culture.

The matrilineal Minangkabau

The Minangkabau have traditionally occupied handsome, multifamily wooden longhouses (*rumah gadang*). Together with the historically cognate Minangkabau Malays of the Malay Peninsula (Negeri Sembilan state), they probably constitute the world's largest matrilineal society. Each Minangkabau man and woman, even after marriage, maintains permanent membership in his or her maternal, natal longhouse. Hence, Minangkabau fathers often split their residence between their wife's and their own natal longhouses. Their children, growing up in the mother's longhouse, tend to experience their fathers as "visitors," whereas their frequently co-resident uncles (mother's brothers) of the same longhouse often become equally or more familiar. In his wife's longhouse a man holds no property and, at best, only exercises supervisory control over his wife's property affairs. In his own natal longhouse he inherits property through his mother and

exercises influence, control, and even formal leadership both in the property and the ritual affairs of his maternal "corporation." He tends to assume a partial "father role" toward his sister's children, while, with his own children, he partially relinquishes it to his wife's brothers. By Islāmic law he is entitled to four wives simultaneously, but polygyny among the Minangkabau has been relatively rare. So also has been the observance of Islāmic property law, which favours males as heirs over females, for it conflicts sharply with Minangkabau customary laws of inheritance.

Although matrilineal organization is unique among the Minangkabau in Indonesia, it may once have been more widespread. It also occurs in Negeri Sembilan and even among certain Montagnard groups in Vietnam, such as the Mnong. The traditional Minangkabau clans have steadily subdivided into more numerous, locally autonomous subclans and have lost some of their traditional ways. In Negeri Sembilan the introduction of smallholder rubber cultivation in the early 1920s began to divert labour away from the traditional wet-rice cultivation, which had been a major source of economic strength for the matrilineal clan.

The Javanese. Three culturally similar ethnolinguistic groups predominate on Java and the nearby isle of Madura: the Javanese, the Sundanese, and the Madurese. The Javanese, occupying central and eastern Java, are most numerous.

Sedentary farm communities in Java

Most Javanese are lowland sedentary farmers, tilling small irrigated rice fields and household gardens. A large Javanese population also inhabits the mountainous interior, growing cassava as a staple crop. Upland villages consist mostly of dispersed hamlets; in the lowlands relatively nucleated communities predominate, although usually strung along roads or rivers. Rural houses are small, rectangular, and usually built close together on the ground. House frames are of wood, walls of plaited bamboo, roofs of thatch, and floors of tamped earth. Small gardens, rice granaries, and livestock yards (for chickens, goats, and buffalo) frequently augment the rural family homestead. The countryside is interlaced by many hundreds of densely populated, contiguous rural settlements. The average "landed" farmer owns perhaps 1.24 acres (0.5 hectare) of cultivable land; many others own far less, and millions subsist entirely by renting or sharecropping land. Most rural Javanese live at a bare subsistence level.

Javanese society divides into two major strata: a small upper stratum and a large lower stratum of rural villagers and manually employed townspeople. Javanese also recognize a vertical division into two religious groups: the *santri* and the *abangan. Santri* Javanese seriously observe Islāmic principles in their daily life; they conduct the five daily prayers, fast during the *pasa* month (Ramaḍān), refuse to eat pork, attempt to make the Mecca pilgrimage, pay their religious tithes, and obey other Islāmic obligations. *Abangan* Javanese proclaim no such devotion and ignore, or even flout, most of these observances. Most Javanese are *abangan,* but *santri* are found widely throughout Javanese society. In the ancient, princely court areas of central Java and in many towns, there are conspicuous localized *santri* communities of merchants, craftsmen, and entrepreneurs. Scattered *santri* families are widely influential also among the nobility, in urban intellectual circles, and in rural villages.

The Javanese life cycle is strongly influenced by non-Islāmic religious beliefs about cosmic and social order. While one's social position and ultimate fate are foreordained, correct family-focussed ritual helps prevent accidents, illness, and misfortune. The central ritual, or *selamatan,* is basically one of distributing sanctified food; this explicitly generates spiritual awareness and emotional equilibrium within the family, particularly during crises. Major family *selamatans* are performed during pregnancy and at childbirth, circumcision, betrothal, marriage, and death. An elaborate series of protective *selamatans* also follows childbirth until, on the 245th day, the postnatal crisis period ends and the baby ritually "descends to Earth" during a major family *selamatan.* Another *selamatan* series also follows a death; it usually symbolizes the deceased's gradual metamorphosis into an idealized, sacred forebear who will continue, it is believed, supernaturally to assist his living descendants. Participants in such family *selamatans* include one's kinsmen, neighbours, and—to recite the appropriate Qur'ānic verses and prayers—local religious leaders.

Child rearing and marriage ceremonies

Social order and emotional self-control are also dominant values in Javanese child rearing. Whatever his social stratum, a Javanese child begins early to learn the complex language etiquette and other behaviour denoting respect toward his parents, other senior kinsmen, and—more widely—all senior people. Discipline begins at about two years of age and variously assumes the form of scolding, bodily punishment, supernatural punishment (by ghosts, monsters, evil spirits), or invidious comparisons with older siblings or unrelated age peers. The shy and withdrawn personalities of many adult Javanese seem to be partly the result of this conditioning. Among preadolescents, boys tend to expand their outside social contacts more widely than girls; the latter, even when in school, maintain responsibilities at home in cooking, paddy threshing, and care of younger siblings. For boys, circumcision between the ages of 10 and 15 marks the transition to adolescence and, among *abangan,* to the ideally proper "Muslim" status; for girls the transition is observed by a *selamatan* at first menstruation. Many Javanese parents avoid imparting sexual knowledge to their children. Chaperonage of unmarried women is fairly common, especially in upper Javanese society, and premarital pregnancy brings great shame to a girl's family. Abortion is considered sinful.

Marriage celebrations involve the most intricate and important rituals of the Javanese life cycle. Weddings are occasions of great display, rich in Hindu–Buddhist symbolism. While most marriages are probably no longer parentally arranged, parental approval remains an important factor. Young men often send intermediaries to gain the approval of a girl's guardian, who, following Islāmic law, is usually her father or brother. Getting married is a protracted process of betrothal (negotiation, gift exchanges, formal announcement), setting an auspicious wedding date, holding the ceremony, paying the bride price, celebrating the subsequent wedding *selamatan,* and possibly carrying out postwedding familial exchanges. The usual marriage age, especially in rural areas, is 15–18 for girls and 17–20 years for young men, and a newly married couple is usually not expected to be economically or residentially independent. The couple frequently lives for several years with the bride's parents.

During and after middle age, many Javanese become increasingly preoccupied with the relationships between natural and supernatural phenomena. Comprehending them requires sensitivity to *kebatinan,* problems of inner experience, as well as observation and knowledge. These are sought through individual and group meditation (*sudjud*), with the aid of systematic exercises, recitations, prayers, devotional offerings, abstinence, and visits to ancestral graves. The primary devotional symbol tends to be "one god," who may also be identified with a particular angel, ancestor, or spirit. Javanese meditation groups differ in their concepts of life, death, the afterlife, and the spirit world, but they usually claim that their philosophy represents *ilmu kedjawen,* or "Javanese science." While much of the imagery is indigenous, it also conspicuously includes both Hindu–Buddhist and Islāmic symbolism.

The Western tendency to make a sharp dichotomy between the natural and the supernatural is generally alien to most Javanese. Their religious philosophy contemplates, instead, the fundamental continuity of man with nature within the known and sensed universe. It stresses the inexorable dependence of man upon superior forces and teaches the value of constant awareness of—and mystical surrender to—these forces. As such, it evokes the expression of the pervasive Javanese values of harmony, balance, and aesthetic sensitivity.

The Philippines. Some 70 ethnolinguistic groups are found in the Philippine archipelago. The dominant, aggregated society, however, consists of the mainly lowland-dwelling Christians on the few large islands, and the label Filipino usually refers to these people. They include a number of large ethnolinguistic groups, which, while strikingly

similar in overall culture, differ in languages, food habits, art forms, dress, housing, and ritual practices. Roman Catholics constitute most of the national population. The non-Christian minority peoples are either Muslim or pagan. Those professing Islām, derogatorily labelled Moros, constitute a small percentage of the national population.

Family relationships among Filipino Christians

Filipino Christians. Christian Filipino society centres upon the nuclear-family group. The father officially heads the family, but women occupy a high social position and exercise much authority, as suggested by their control of 80 percent of Filipino business firms and their fairly dominant position in the scientific–academic community. Within the nuclear family, respect for elders is inculcated early in life; honorific pronouns and articles are obligatory in speech to parents and elders. Collective responsibility within the family is highly valued; older siblings must care for and, if necessary, make considerable sacrifices for their younger siblings, especially in educational opportunity. Stinginess toward close relatives is a serious breach of fundamental Filipino social values.

Social relationships outside the nuclear family are distinctively patterned. The individual recognizes relatives on both the maternal and paternal sides. By carefully bestowing gifts or services upon individual members of his kindred group, a Tagalog develops enduring ties of reciprocal obligation called *utang na loob,* literally "debt of the inside." Two persons become linked through a long period of alternating indebtedness that leads to a complementary relationship wherein obligations often cease to be calculated. While seldom explicitly discussed, such partnerships can involve deep feelings of mutual gratitude and friendship. Normally, an individual's social network includes many such dyadic ties in different stages of development. Even a child's obligation to care for his parents is thought ultimately to rest upon basic indebtedness to the parents for their gift of life. To renege upon one's *utang na loob* obligations is a grave breach of personal trust; it causes shame (*hiya*) and can even lead to violent revenge by the aggrieved.

Filipino society is stratified according to wealth, occupation, and education. Wealthier upper-class people usually live in small towns or regional capitals, in non-Malay-style houses; they tend to have professional occupations. People of different status levels frequently establish ritual kinship ties with one another through the widespread Spanish Catholic custom of ritual co-parenthood (*compadre*). This can facilitate an even wider extension both of *utang na loob* relationships and of entire patron–client networks, the latter being an essential ingredient of political power. Socially important family alliances may also be consolidated by marriage. While national law and Ilocano custom both prescribe equal property inheritance, Ilocano parents, like other Filipinos, will often bestow all their property on only one heir. While concentrating family resources effectively, this custom also perpetuates parental control over "key" marriages and family linkages among many groups.

Rural settlements among the Christians are made up of Malay-style houses (bamboo, pile-supported, thatch-roofed) with their associated rice fields, gardens, and groves of coconut and bamboo. Domesticated animals are mainly chickens, pigs, and water buffalo. In areas of shifting agriculture, the farmer will often move away seasonally from his home to small guardhouses near the crops in distant fields. Where plantation-scale sugar or copra is grown, as in the Visayan Islands, Filipino worker settlements may be conspicuously large and compact.

Islāmic peoples

Other Philippine societies. The Islāmic peoples of the Philippines fall into several categories. The interior Maranao of Mindanao's Lake Lanao area, while primarily wet-rice farmers, also maintain vigorous cottage industries in mat making, weaving, woodwork, bronzework, and brasswork. Less conspicuous as craftsmen but depending both on fishing and rice farming are the Magindanao, the Philippines' largest group of Islāmic people, who inhabit the Cagayan River Valley inland from Cotabato. The Tau Sug, mainly occupying the northern Sulu Archipelago, depend upon extensive maritime trade, pearl diving, handicrafts, dry- (and some wet-) rice agriculture, and a varied horticulture and arboriculture. The Samal of the southern Sulu islands maintain homes, mosques, and agricultural plots ashore but live much of the time on boats, where they engage in fishing and maritime trade. Only their coastal dwellings and Islāmic community ties separate many of these Samal from the pagan Bajau, impoverished fishermen of the same area who live exclusively aboard small family houseboats. The Yakan, of Mindanao's Zamboanga Peninsula and Basilan Island, are fishermen, craftsmen, and gardeners but only occasionally Islāmic.

Social organization among these groups remains significantly traditional. The larger societies are composed of locally powerful social pyramids usually headed by hereditary aristocratic chiefs, or *datus.* A pyramid includes the *datu* and his close kinsmen (including in-laws) of highest social rank, the usually related nobility of lesser rank, and a large subordinate group of dependent local retainers and commoners linked by ties of kinship, fealty, or debt bondage. A *datu*'s title will often proclaim his allegedly direct descent from the Prophet and symbolize his considerable local authority in both religious and civil matters. In practice he usually shares this power with an Islāmic religious adviser (*kadi*) and a council of ranking noblemen. Both locally and in their external ties to the greater society, such sociopolitical pyramids are sustained by intricate combinations of leadership charisma, kinship solidarity, respect for hereditary rank, family size, alliances by marriage, and control over resources, such as rice land and slave labour. Polygyny, cousin marriage, and hypergamy (female marriage upward) are three enduring customs by which, particularly among the nobility, kinship ties are used to bind groups together and become major channels of alliance between families of equal rank within different local pyramids.

Power of the *datu*

Most small sultanates and principalities among the Maranao and Magindanao have this pyramidal organization. The Maranao are more politically fragmented and more recently Islāmized than the Magindanao. The Tau Sug, ethnically the core population of the traditional sultanate of Sulu, have been Islāmized since the 15th century. Historically vigorous in resisting all foreign political and religious domination (Spanish, U.S., and Christian Filipino), their small *datu*-led principalities maintain a regionally marked tradition of political coordination. Today, the Tau Sug are still widely regarded as conspicuously proud, aggressive, and devotedly Islāmic. For their close neighbours, the less Islāmized Samal, the Tau Sug have constituted an ethnically distinct superior social class. While a Tau Sug man may marry a Samal, rarely if ever does a Tau Sug woman marry "down" with a Samal. The Samal maintains a similar class position toward the Bajau, who, as they have begun to settle ashore and adopt Islām, are increasingly allowing their women to marry "up" among Samal men.

In their social organization, economy, technology, dress, art, and proud religious identity, the Islāmic peoples of the southern Philippines are most immediately similar in culture to the coastal peoples of southern Celebes and of western and northern Borneo, such as the Banjarese, Makasarese, Brunei Malays, and others. The similarities stem largely from their common experience of Islām, their traditional dynastic interrelationships, and their collective antipathy toward Christian colonizers. Today, the Philippines' Islāmic groups constitute an increasingly self-conscious political minority with distinct problems of national assimilation.

The many pagan minority peoples of the Philippines live mostly in the highlands of Luzon, where they are collectively called Igorot ("Mountaineers"), and of Mindanao. They range from the small hunting bands of pygmy Negritos (on Luzon and Palawan) to the physically "Malay" terrace rice farmers (such as the Ifugao) of northern Luzon, the Visayan Islands, and Mindanao.

Modern developments. The traditional customs have changed greatly. Except for Thailand, the nations of the region have all experienced the colonial imposition of western legal and administrative codes. Burma and Malaysia knew the Pax Britannica, and Indonesia experienced prolonged Dutch rule. The Philippines were subjected to

successive Spanish and U.S. administrations. Laos, Kampuchea, and Vietnam underwent French administration.

The persistence of old ways

Localized patterns of customary law remain. Many pagan and Islāmized peoples of the southern Philippines and northern Borneo have intricate conceptions of proper interpersonal behaviour and impose sanctions, especially fines, for violations of standards. These often involve questions of sexual and marital propriety and give rise to constant village-level litigation before local headmen or "native-court" adjudicative bodies. Below the level of litigation, sensitivity to minor local custom, or folkways, remains exceedingly widespread, especially throughout insular Southeast Asia. Few topics are more common in the folklore of the Malay languages area than themes of "other villages having other ways."

Divining is still common. Sarawak peoples such as the Iban (Sea Dayak) predict future events with the aid of a rich lore about the flying formations and habits of particular birds. They and other Borneo peoples also often read the entrails of recently killed animals to determine future events. Trial by ordeal is probably still common among many mainland hill peoples and in some interior regions of the archipelagoes; in the 1960s ordeal by water immersion was in occasional use before certain "native courts" of eastern Sabah (Malaysia). The reading of horoscopes is widespread throughout the region, even among educated urban people. The Burmese and Javanese have rich traditions of horoscopy, drawing judgments of the future from an enormous array of temporal, chromatic, auditory, and physiological cues.

The major religions have been dealt with above. In addition to the creeds introduced from outside, indigenous beliefs in natural and ancestral spirits are exceedingly common. Often they are fused with Buddhist, Islāmic, and Christian elements, but they are found in plain form among the smaller, more isolated, and still essentially pagan peoples. The widespread practice of revering ancestral spirits expresses the generally sacred significance of family continuity. The Vietnamese clan shrine is but one major example. The practice occurs in various forms among the Balinese, Batak, Burmese, Thai, Makasarese, and other peoples.

Ethnic images and regional patterns

Ethnic styles and stereotypes have a subtle but important influence on personal judgments and social relationships in Southeast Asia. Among Filipinos, the Ilocano is stereotyped as industrious, tough, frugal, and without a sense of humour; the Bicolano is supposedly even-tempered and religious. Among Islāmic Indonesians, the Makasarese is thought to be excessively proud, quick to take offense, and implacably vengeful. Such stereotypes are likely to be even cruder among culturally less similar peoples. Among the Burmans (the major ethnic group of Burma), for example, the Shan peoples of northern Burma tend to be regarded as uniformly "untrustworthy." Throughout much of Southeast Asia, immigrant Chinese businessmen are considered unfailingly industrious, penurious, and shrewd to the point of dishonesty. The Chinese often reciprocate by stereotyping others, as in Malaysia, as lazy, untrustworthy, and superstitious.

With all their diversity, Southeast Asia's people have several cultural patterns in common. Women command substantial social and economic authority; only among recently immigrant Arabs and Indians are they distinctly excluded from any social sphere, although on formal occasions there is often a physical separation of the sexes. The prevalent type of family organization is that of the nuclear, monogamous family linked to both maternal and paternal kinsmen; unilineally organized large family groups are found only among the patrilineal Vietnamese, the unilineal peoples of Sumatra and the Malaysian state of Negeri Sembilan, and among several hill groups of the mainland and of eastern Indonesia. Widespread sensitivity to, and preoccupation with, social status is another conspicuous pattern. This is visible in the still-traditional court etiquette of Thailand and central Java, but it is also apparent in the widespread sensitivity to language "levels" and to the implications of making a "proper" marriage. Such sensitivity goes with an equally widespread concern for maintenance of self-esteem, or "face"; there is an elaborate personal etiquette for maintaining outwardly harmonious social relations and thereby avoiding the infliction of shame or indignity upon another person.

Population movements

The growth of population and the expansion of cities are among the most important of recent trends in Southeast Asia. Manila, Ho Chi Minh City (formerly Saigon), Kuala Lumpur, and Bangkok are among the world's fastest growing cities. This urban growth opens up wide opportunities for the intermingling of different groups and cultures.

Some population shifts have resulted from government programs. Vietnam's numerous postwar relocation and reeducation camps and resettlement schemes have forced vast numbers of uprooted Vietnamese civilians and Montagnards to settle among new neighbours and cope with new economic circumstances, often under very adverse conditions. The Indonesian government has resettled substantial numbers of Javanese in new communities in central Sumatra. Similarly, the Philippines government has resettled Christians from the crowded Visayan Islands to the northern and western sectors of Mindanao. Rural-development projects in Malaya have relocated substantial numbers of land-poor Malays to pioneer areas cleared for future rubber plantations.

In the 1960s newly opened rubber and palm-oil plantations in eastern Sabah (Malaysia) drew substantial numbers of Indonesian immigrant labourers from the Makasarese area of southern Celebes and from Indonesian Timor. Timber-cutting operations attracted many Christian Filipino labourers south to interior Mindanao. Among the Lawa hill peoples of northern Thailand, the increased costs of maintaining their trained timbering elephants has forced many to sell their animals and seek wage-labour jobs in the lowlands. Families and whole villages of Meo, Lahu, and Yao hill peoples have moved into northern Thailand from upper Burma and Laos in search of new, high-altitude farming areas for the cultivation of opium poppies. Thousands of Vietnamese families, dislocated since 1965 by the massive destruction of their agricultural resources through saturation bombing and herbicide spraying, have had to leave their home communities in search of new livelihoods.

Rise of national languages

The cities and suburbs of Southeast Asia have become commercial centres for the sale of foreign manufactures; they symbolize a transnational culture of salaried economic prosperity, literacy, white-collar employment, and recreation. As centres of government for new, multi-ethnic nations, one of their most urgent tasks is to develop that basic vehicle of national culture, the national language. Bangkok is energetically fostering early school education in the national language (Thai) and literacy in its Indian-derived writing system. The Indonesian language, developed from the literary Malay language of Sumatra, is universally taught in Indonesian schools and has come to be widely spoken and read (in a Romanized alphabet) by younger Indonesians. Malaysia's new national language, another dialect of literary Malay written in a Romanized alphabet, is being vigorously promoted as a common medium to link the large resident Chinese population with the Malays and other indigenous peoples of Sarawak and Sabah. In Manila a version of Tagalog written in a Roman alphabet has been declared a national language and is gaining acceptance. Burmese and Khmer, both written in Indianized scripts, are routine school subjects in Burma and Kampuchea. Vietnamese, mostly written in the Romanized Quoc-ngu script, is increasingly being taught among southern Vietnam's hill peoples.

New nonlinguistic skills are also spreading. Road building in Thailand, Malaysia, and the Philippines has taught local workers the use of motorized equipment ranging from bulldozers to passenger vehicles. In northern Thailand, Mindanao, and northern Borneo, the chain saw has enormously accelerated local timber-cutting operations. In Borneo this tool, together with growing access to inexpensive outboard motors, is revolutionizing the economy of many communities along the rivers. In Vietnam and Thailand the U.S. military forces trained thousands of Vietnamese and Thai in such skills as helicopter operation, ordnance logistics, aircraft maintenance, and the use of electronic computing and tracking systems.

Other tendencies to cultural "homogenization" include an increase in inter-ethnic marriages. In the cities of Java they even occur on a significant scale between ethnically different Christians and Muslims. Another aspect is the displacement of traditional codes of behaviour among urban immigrants: in Sabah (Malaysia), for example, rural Dusun women coming as nurse trainees to the capital city discover that their "native law and custom" does not apply to social or sexual behaviour in the city.

Movements reacting against cultural uniformity have also developed, usually as efforts to reassert traditional cultural values. Since Indonesia obtained its independence, the Javanese have experienced a great proliferation of meditative sects, or *aliran kebatinan,* each usually led by a local guru, or teacher. Many *aliran* meet regularly to examine the detailed ramifications of "Javanese science" and to repudiate materialistic secular values such as efficiency, individual success, and Western scientific attitudes. Nativistic local leaders have also been conspicuous in Borneo, Sumatra, Burma, the Philippines, and peninsular Malaysia; to rally and hold followers, they often invoke the symbols, episodes, and folk heroes of regional history. In some instances these groups have developed into armed revolutionary movements. (P.R.Go.)

SOUTHWEST ASIA

Southwest Asia—or the so-called Middle East, if Egypt is included—may be defined as the region extending from Egypt to Iran. Countries included in the area are thus Egypt, the Sudan, Saudi Arabia, Yemen (Ṣanʿā'), Yemen (Aden), and Oman; Kuwait and the small Persian Gulf states; and Israel, Jordan, Lebanon, Syria, and Iraq. The total land area is approximately 2,700,000 square miles (7,000,000 square kilometres). The area has, in the 20th century, been given other appellations, including the Levant, the Near East, and the Mashriq (Arabic for "the East"), as well as the Middle East.

From the perspective of its cultural geography, Southwest Asia may be divided into two types of region: the arid, infertile desert areas that are the home of the nomadic Bedouin and the fertile, cultivated river valleys of the settled agriculturalists.

The great plateau is a notable physical feature of the Arabian Peninsula and constitutes most of the desert area of the peninsular landmass. Barren, except for occasional oases and the Najd, a pastoral tableland in the north, the peninsula is fronted to the west by mountains rising in the Hejaz, the holy land of Islām, to some 9,000 feet (3,000 metres) and further south, in the Yemen, to some 14,000 feet. The south central portion of the peninsula, an area of some 40,000 square miles (100,000 square kilometres), descriptively named, is known as ar-Rubʿ al-Khali (the Empty Quarter). Here, the average summer temperatures are in excess of 120° F (49° F).

The Fertile Crescent

The outstanding geographical feature of the remainder of Southwest Asia is the Fertile Crescent, an arc of territory stretching from the Nile Valley of Egypt through present-day Israel, Lebanon, Jordan, and Syria, into Iraq. At each horn of the Fertile Crescent there is a river valley—to the east that of the Tigris and Euphrates rivers, in Iraq, and to the west that of the Nile, in Egypt. These valleys are composed of rich and fertile masses of alluvium deposited year after year on a substratum of sand and rock. In Egypt this deposit is almost 40 feet deep. As they are annually replenished, these valleys are continually refertilized. They also provide irrigation and drinking water.

Ethnic groups. Southwest Asia has the longest cultural history of any region in the world and was in fact the birthplace of human civilization. Thee monotheistic religions—Judaism, Christianity, and Islām—originated there. As a land bridge connecting Asia, Africa, and Europe, it has been traversed by numerous peoples, many of whom left small groups in the region. Its cultural evolution has been such that, until recently, minority groups were permitted to retain separate and distinct entities, so that the most conspicuous fact about the Middle East has been its demographic heterogeneity; its population is a mosaic of peoples.

Among the earliest inhabitants of the region were the Sumerians, believed to have been of an Armenoid physical type—roundheaded, of moderate height, and of heavy build. They spoke a language with certain agglutinative features suggestive of the Mongoloid family of languages. The ancient Egyptians were of the Mediterranean physical type—longheaded, dark-haired, relatively slender, and of moderate stature, speaking a separate language classified as Hamitic.

Additional early peoples in the area were the Akkadian, Babylonian, and Assyrian invaders and conquerors of Mesopotamia, the Amorites of northern Syria, the Aramaeans of inner Syria, the Canaanites of the Levant coast, and the Hebrews of the hill country. All of these peoples are generally classified as Semites—not because they possessed any unique physiological feature but because they all spoke languages assigned to the North Semitic family and because they are generally believed to have originated somewhere in the central Arabian Peninsula. Arabic, the primary contemporary South Semitic language, was apparently originally spoken by a small group of traders, townsfolk, and desert nomads in the district of Mecca and Medina. Arabic was the language of Muḥammad and his early followers, and, with the rise and expansion of Islām in the 7th century, it quickly replaced most other languages of the Mashriq. Within the region, Egyptian, Palestinian, Syrian, Iraqi, and Yemeni dialects of Arabic are spoken today. Some of these differ only in pronunciation, while in other instances the colloquial forms may be quite wide apart. Other languages spoken in the region are Hebrew, which is one of the official languages of Israel (the other is Arabic); Aramaic, which is reputed to be still spoken in a few villages of Syria; Kurdish, an Indo-European language spoken by the approximately 10,000,000 Kurds resident chiefly in Iraq, Turkey, Iran, and Syria; and Fārsī, spoken by the majority of Iranians.

Linguistic basis of the Semitic ethnic classification

Settlement patterns and economic organization. *The fellahin ("villagers").* Agriculture has long been and still continues to be the predominant way of life of the majority of the peoples of the Middle East. Spread along the Fertile Crescent and in the isolated oases of the deserts, the overwhelming majority of the people were—and still are—peasants. Despite widespread urbanization, more than two-thirds of the people may still be classified as fellahin.

Family solidarity and the need for security have always precluded a pattern of isolated farmsteads. Instead, the peasants gather together in villages, ranging from modern and well-off communities in Lebanon (particularly among the Christians) to the miserable mud and kerosine-can huts of seminomadic tribesmen. Despite such disparities, however, a broad cultural pattern can be discerned, especially in the Arabic-speaking regions.

Landownership is complicated by the fact that landed property, or the right to work other people's landed property, is divided among the children, with boys sharing equally but with a girl's share equalling only half that of a boy. Complex fragmentation of agricultural land results, with the dispersed small plots, sometimes termed "dwarf holdings," rarely economically feasible. High-interest loans are often not met, and as a result the accumulation of large tracts of land in the hands of a few wealthy landowners is quite common; many debt-ridden peasants are consequently forced into sharecropping. In virtually all of the newly independent nations of the Middle East, land reform is one of the basic tenets of revolutionary programs. Much, however, still remains to be accomplished.

The traditional village

Throughout the region village layout is strikingly similar. At the centre is the mosque, or church, bordered by a few small shops of the grocer, blacksmith, and cobbler. Some five to 10 twisted alleys, just wide enough for a donkey with its side baskets, wind between the walls of low, flat-roofed houses and then narrow into even smaller lanes that generally end where the fields begin. The village square functions as a marketplace for itinerant peddlers and is the scene of village ceremonial events.

The homes, whether of stone blocks or sunbaked bricks, consist of a few rooms and a courtyard. When a son marries, a new room is added along the side of the house, on the roof, or in the courtyard. Stables are uncommon, and the fellahin share with their animals the same dirt-

and dung-littered courtyard and, in winter, the same room for sleeping.

Furniture among the peasants is limited and strictly utilitarian. Since squatting on the floor is the normal position, whether doing the housework or eating with the fingers or with a piece of bread dipped into a common bowl, there is no need for tables and chairs or forks, knives, and spoons. A chest, a pile of bedding stored away during the day in a recess in the wall, a cradle, a few straw mats, a kerosine lamp, and perhaps one or two cheap carpets constitute the essential furniture. Kitchen needs are met by a few clay jars for keeping the drinking water, a large grain bin coated with clay and sealed at the top but open at the bottom for access to the daily need, a copper kettle, a few handwoven baskets, some locally made earthen jars, and a variety of imported bowls and basins. A few coffee cups on a tray for occasional visitors, a few photographs of the sons (daughters are rarely photographed, and their pictures are certainly not on view), some religious prints or calligraphic verses of the Qur'ān, and a newspaper photograph of the local national leader complete the inventory of the typical village home.

Water is generally a scarce commodity. If the peasants are fortunate, they may have a well in the courtyard. Generally, however, water is obtained from common village wells, springs, or ponds. When these dry up during the long arid summer months, animals and humans share the same distant, often polluted, ponds or streams. Piped water to the courtyards or to the village centre is not uncommon among wealthy villagers or in areas in which the government has embarked on special rural development programs. In some regions of Iraq and Egypt and along the rivers of the Levant, although water may be readily available, it often is a source of malaria and other diseases. Flush toilets are very rare, as are outhouses. Urine is generally disposed of in the courtyards, along the paths, or in the fields. Fecal matter is spread in the fields.

Animal dung, in treeless regions, is the prime source of fuel. Women and children carefully gather the dung, shape it, dry it in the sun, and store it on the roof or in the courtyard. The family oven, in which bread is baked, is generally in the courtyard and is a dome-shaped mud affair with a low opening.

The diet is simple and generally monotonous. Its mainstays consist of unleavened bread, accompanied by rice and a few vegetables, usually onions and garlic, and, depending on the region, olives or dates. In season, small amounts of fresh fruit and green vegetables may be added. Tea and coffee are luxury items paid for out of a limited budget. Meat and fat are eaten, by most, only at special ceremonies such as a wedding, birth of a son, or great holidays or in honour of a distinguished visitor.

Health and education among the fellahin

Standards of health and education well demonstrate the poverty of the peasants. Malnutrition is chronic, and infectious diseases are endemic. It is estimated that perhaps nine-tenths of the fellahin suffer from a variety of dysfunctional or debilitating diseases such as trachoma, typhoid, dysentery, intestinal worms, and tuberculosis. Western-type medical aid is still not widely available among peasants, and their main recourse continues to be the traditional village practitioners. Schooling, despite the fact that most countries of the Middle East have enacted laws requiring compulsory education, is uneven and sporadic. Illiteracy among rural adults, particularly women, is still the rule.

The way of life for the early Zionist farmers working the marginal lands of Palestine often was as difficult as that of their fellow Arabs. It was the attempt to break out of that pattern that led to the development among the Jews of collective forms of agriculture—variants of the kibbutz and the moshav—designed to maximize their financial and labour input, to raise their agricultural returns, and to improve their collective standard of living.

The Bedouin. The Bedouin, as one observer has put it, are more glamorous than numerous. They have rarely constituted more than 15 percent of the total population of the Middle East, and in the late 20th century their numbers were rapidly decreasing, even in the Arabian Peninsula, as they voluntarily migrated to the towns to work for the oil industry or to settle on government-sponsored tracts of land.

The homeland of the classical, traditional Bedouin is an area of the northern Arabian Desert, stretching into Iraq, Syria, and Jordan. The Bedouin are nomads in varying degrees, but the true Bedouin are those who possess only camels and who spend their summer months camped around wells or streams and the rest of the year ranging the desert. In kinship they trace their origin back either to Qaḥṭān, an ancient patriarch who lived before Abraham, or to Ishmael, son of Abraham and Hagar. Collectively they constitute the *aṣīlīn,* a closed circle of numerous tribes or tribal confederations.

The material culture of the Bedouin is limited. The chief possession is the home—a long, low, black tent of woven goat hair. Men and women have separate divisions within the tent. The male side, away from the wind, is left open; it has a carpet, spread before a portable stove, or coffee hearth, with a mortar for grinding the coffee, a coffeepot, and cups. This is the social centre of the household, the meeting place for guests, and the site for the transaction of all tribal affairs.

Clothing is simple: a long robe of thick material, a large headcloth wound around the head and held in place by a black goat-hair cord and rawhide, and heelless sandals.

Water is generally too scarce to be used for washing, and sand or animal urine is substituted for use in daily ablutions. The diet is limited, consisting of a variety of milk products such as curds, buttermilk, and cheese, dried fruit, particularly dates, and, when obtainable from the fellahin, grains such as wheat, barley, or rice. Meat, as among the peasants, is a luxury eaten only when an animal has died a natural death or for a special ceremony. The meals are served first to the older men of the tent, then to the younger men and boys, and finally to the women and the girls. The sexes do not intermingle during mealtime. The inadequate diet and low standard of nutrition result in a people who are small, even stunted, and slightly built. By the age of 40, particularly among the women, old age has begun.

Bedouin raiding

Raiding was the traditional means of supplementing the deficiencies of life in the arid zone. The Bedouin took by force from the peasants what they lacked in foodstuffs, material goods, and even women and children. Successful leadership in raids could be a most effective means of developing reputation and power, a practice that to this day has not been completely curtailed.

Social organization. In certain respects the fellahin and the Bedouin share a similar social structure, and thus the pattern traditionally common to both can be broadly sketched. The model described is that of the classical, idealized, Arab way of life.

The family. The family is the most fundamental and important feature of Arab social structure. The family is extended—that is, it is headed by an elderly man and consists of his wife, or wives, his married sons and their wives and children, his unmarried sons, and his unmarried daughters. Daughters, when they marry, leave their own extended families and become incorporated into the extended families of their husbands, even though their moral conduct continues to be the responsibility of the original family.

The extended family may total a dozen or more persons, all of whom live in one house or in a number of adjoining houses or, as among the Bedouin, in a number of tents pitched next to each other. The extended family functions as one entity in economic, political, and military ventures.

The family is patrilineal—*i.e.,* an individual belongs only to the family of his father—and patrilocal (young couples when they marry take up residence in or near the house of the bridegroom's father). It is also patriarchal: the father is the head of his family, and the elderly male who is head of the entire extended family is the absolute ruler over the entire group, traditionally. Even to this day, in certain respects, he has jurisdiction over life and death.

Pride in descent forms one of the most significant traits in the ethos of the peoples of the Middle East. Emphasis on noble ancestry and lineage is a matter of great importance among the Bedouin and to a lesser extent among the fellahin. Tribal and village genealogists trace an individ-

ual back from his individual extended family to his sub-*ḥamūlah* (a cluster of extended families) to his *ḥamūlah* (a large group of families tied together by an actual or assumed genetic relationship) and to his subtribe (*'ashīrah*) and tribe (*qabīlah*). Beyond his tribal affiliation, a man's lineage will be traced to his supposed progenitor dating back to early biblical sources. This tribal pedigree, even among the villagers, can play a fundamental role in determining and justifying positions of traditional authority and political power.

Marriage and divorce. There is a strong tradition of endogamous marriage (that is, of marriage within a very small social circle), ideally with the father's brother's daughter or, failing that, within the same village or tribe or having the same social status. A man may be married to more than one wife simultaneously (polygyny).

Prevalence of polygyny

The prevalence of polygyny among the Arabs has been greatly exaggerated by Western commentators, and the practice has always been the exception rather than the rule, especially in modern times. The simple biological problem of acquiring a surplus of young females plus the economic cost of paying their bride-prices and then adequately supporting them and their offspring all have prevented any more than 4 to 5 percent of the married men from having more than one spouse.

Traditionally marriages are arranged and consummated when the women are still in their teens. This is considered highly proper and beneficial. An old Arab saying holds that there are only three things in life a man must do quickly: bury the dead, serve a guest, and marry off a marriageable daughter.

Great prestige is attached to the male's procreative ability and to the regular delivery of live, preferably male, children. Status is achieved through the production of large families; a childless couple is regarded with contempt. Only the extended family with many children, preferably boys, can be a strong family, providing economic and political security. The injunction to "be fruitful and multiply" thus provides a constant goal.

Term, or temporary, marriage, usually referred to as *muṭ'ah,* dates back to pre-Islāmic days and still survives in certain parts of the Middle East. Term marriages were contracts made with women by soldiers, traders, and travellers under which the couple were legally married for the day, the length of the caravan journey, or for another defined time period. At the stated end of the period the contract was no longer binding. Children born of *muṭ'ah* marriages held all the rights of the offspring of ordinary marriages, but the term wife could not inherit from her husband nor he from her.

The announcement of the birth of a child to the father is based upon long-standing traditions. If the infant is a boy, the father will be told *bishārah,* meaning "good tidings" or "reward." The father will ordinarily give a modest gift to the individual who first brings him this news. Among those who can afford it, a sacrifice of a sheep or a feast is also common. If the infant is a girl, no reward will be given for the announcement, neither will there be a celebration.

Traditional divorce in Arab Muslim society is extremely simple. The male is the sole initiator of such an action. The wife can neither oppose a divorce nor initiate one on her own behalf. The husband pronounces the traditional formula, "I divorce thee," in the presence of two witnesses, and the divorce is effected. The wife must return to her father's (or brother's) family and can claim only that portion of the bride-price that was stipulated in the original marriage contract to be paid in the event of such an occurrence. If a wife should leave her husband and he later agrees to a divorce, she forfeits any part of the bride-price. After the divorce a man may marry immediately, but the woman must wait three months to make sure that she is not pregnant. If she is pregnant, she cannot remarry until she has given birth to the child and reared it, during which time the father (her former husband) must support her. All children belong to the father and, in the event of a divorce, must be delivered up to him.

Among the Arab Muslims no stigma is attached to divorce, and divorced women as a rule remarry as soon as they are permitted to do so.

Socialization. Shortly after birth, the infant will be rubbed with salt and oil, not only to cleanse him but also to enable him to grow up to be modest and courteous. The infant will then be wrapped in swaddling clothes for a time period ranging from 40 days to six months.

Nursing may last for 18 months to three years. A mother is traditionally prohibited from weaning her child before two years. The extended nursing period postpones another pregnancy (suckling periods are generally connected with an abstinence from intercourse), and it is believed to make the child strong.

To mothers it is said: "pampering a girl disgraces thee; pampering a boy makes thee rich." The boy will therefore be nursed longer than the girl. As soon as the child shows the slightest signs of restlessness, he will be breast fed. As a consequence, the female is seen as sympathetic and the woman's breast as a symbol of compassion.

If for some reason, such as illness, a child is nursed by a woman other than its mother, certain protocol must be observed. A woman may not give her milk to another child or have another woman give her child milk without her husband's permission; for the child belongs to the husband's patrilineage, and that important decision rests with him. Furthermore, a boy and a girl nursed by the same woman are considered to be brother and sister and can never marry. Nursing relationships can also be a method of adoption (especially of an orphan): if a woman wishes to adopt a strange child or even an adult, she does so by offering her breast.

The religious rite of circumcision

One of the most important events in a Muslim and Jewish boy's life, not only for himself but also for his parents, is circumcision. The rite of circumcision is a religious duty, and no male can enter heaven uncircumcised. The age at which circumcision is carried out varies: it may be done shortly after birth or in childhood, but it must be done before marriage. Fathers will ordinarily have their sons circumcised soon after birth, not only because it is less painful then but also because of fear that the boy might die before the father has fulfilled the commandment of circumcising his son.

Among Muslims, the rite of circumcision proceeds in a program similar to that of a wedding. It is preceded by an evening of joy in village festivities, henna (a dye) is put on by the woman, and a new outfit is provided for the child. Festival garments decorated with flowers and leaves are worn, and, in similar fashion to a bride, the boy is led in procession around the village in his new outfit. After the ceremony there is a feast at which presents are given. Several boys may be circumcised at once, or the circumcision may be combined with a wedding to save expenses. There is a further connection between circumcision and marriage, for, as it is the duty of the father to marry his children before his death, he attempts to find a suitable mate for them as soon as possible. A boy thus often acquires his bride at his circumcision ceremony, although he will not marry her until sometime after puberty.

The early socialization of a boy and a girl into the traditional family of the Middle East differs profoundly. The growing boy is trained to become an obedient member of his family, able to subordinate his wishes to those of his father and elder brothers. The fact is impressed upon him that the interests of his family always come first and that he must govern his actions to enhance the collective strength of his family. The girl, on the contrary, is not only weaned earlier than the boy but, by the age of five or six, is already being consciously prepared for moving out of the home of her parents and into the home of her mother-in-law. There, she will fill a subordinate and even a servant role—a role that will be improved only as she bears sons.

Religion and health. A discussion of the predominant religion of the area, Islām, can be found in the article ISLĀM. But some understanding of the function of religion in the traditional Middle East can be grasped from an analysis of its role in the peoples' approach to health, disease, and death.

Philosophy of illness. The essential philosophy underlying the traditional system of medicine in the region is that illnesses and injuries are subjective affairs arising from

one's own acts or omissions or caused by someone or something that is possessed with power. Illnesses or injuries do not just randomly occur—they befall a certain victim, at a given time and in a definite manner, because of specific causal actions.

Evil spirits and the evil eye

Two essential elements of this philosophy are belief in animism and animatism, or, more specifically, the belief in the existence of evil spirits (animism) and of the "evil eye" (animatism), or the power of certain persons or objects to affect and influence the human body—and nature as well, in certain circumstances.

Evil spirits abound in the environment, ready to pounce on the unsuspecting victim. Strong, healthy, mature individuals are the least susceptible to such attacks; the most susceptible are infants and children, the weak, the ill, the aged, and normally healthy individuals in certain circumstances (women during menstruation, pregnancy, or while giving birth, for example).

The presence of strong, healthy, mature individuals near the susceptible person is a strong deterrent to evil spirits. As one cannot rely on such persons to be constantly on duty, however, various inanimate objects with strong power to repel the evil spirits are called into play. Common objects of this nature are the Hand of Fāṭimah (beloved daughter of Muḥammad), which may have inscribed on it holy words in Arabic or Hebrew and which is generally worn around the neck (Cochin Jews may put it around the abdomen); the shield of David (a six-pointed star); and blue beads, pieces of jewelry, or bits of cloth that are worn around the neck or attached to the clothing (blue is particularly repugnant to the evil spirits and the evil eye). Sometimes a concoction of evil-smelling herbs will be placed in a bag and worn close to the body, or various religious phrases will be written on paper and sewed into the clothing or put into a bag and worn on the body.

The Bible also possesses the power to repel evil spirits, and thus some Jewish mothers place a copy of it beneath the pillow. Another practice, but less common, is the preservation of the foreskin that is cut off during the *ber-it mila* (the ceremony of circumcision conducted on every Jewish male child when he is eight days old), by drying the piece of skin and powdering it. It is then sewn into a piece of cloth and kept under the pillow or among the blankets of the child's bed.

Evil spirits fear the name of Allāh, which strikes terror into their hearts and weakens them and forces them to withdraw or repels them completely. Consequently, his name is uttered frequently while a person is engaged in the everyday routine of life. Otherwise healthy individuals must, under certain circumstances, be particularly careful to invoke Allāh's name. When, for example, a Muslim couple is about to engage in sexual intercourse, the male must first say a prayer: "I seek refuge in Allāh, from the accursed Satan, in the name of Allāh, the Beneficent, the Merciful." If this prayer is not spoken, the evil spirit will enter the woman, and the child she conceives will be evil, bad, or a devil, or the woman herself might fall ill. A prayer to Allāh also ensures conception—the mere physical act of intercourse is no guarantee.

Vulnerability of women. From the moment of conception until the last birth pang, the pregnant woman is especially sensitive to evil spirits. Each of her actions is carefully watched by them, and, should she commit a transgression or fail to constantly invoke the power of Allāh, retribution will be certain to follow. The afterbirth provides powerful protection for the newborn child and so must be saved. It may be left attached to the child for a period of some hours or overnight, and it then must be preserved in or near the house.

Restrictions on menstruating women

Women during menstruation are thought to be very dangerous. They are not only considered impure and unclean but also, if not actually possessed by a spirit, likely to be transmitters of the actions of evil spirits. They must therefore be separated from others, particularly from the ill and women in labour. Women during menstruation sometimes must leave the home and live in a menstruation hut or tent for the entire period, returning to their homes only after they have been purified. When a woman is permitted to remain in the home, she is subject to numerous restrictions: She must sleep on the floor or on a low bed, must have no sexual relations with her husband, must not even touch him or his bed, and should not prepare any meals or enter a home in which there is an ill person or a woman in labour. Such restrictions are only practicable within the extended family, and disintegration of the extended family leads to forced abandonment of many such practices. The beliefs themselves, however, may persist, as may anxieties when the taboos are broken.

Safeguards against the evil eye. The prevention of illness to the inner body by the evil eye—as distinct from the prevention of illnesses caused by evil spirits—is based on the principle of misleading, deceiving, and deluding the evil eye. The evil eye is particularly feared, and more than half of all deaths are attributed to it. It is attracted to the healthy, the beautiful, the happy, and to children.

In the Middle East, possessors of the evil eye are often women. Psychologists have argued that the evil eye is in reality an envious eye, and the entire corpus of preventive measures seems to be based upon the principle of not attracting its attention (or envy). The youngest are a particular attraction to the evil eye. Thus children, esteemed the greatest blessings, are kept dirty, ragged, and unkempt, particularly when out in public. The child may be called "Oh, dirty one" or "Oh, evil one" and similar names, in order to disguise the true feelings of the parent. The child may be given a false name, and its true name kept a secret, in order that the real name may not be overheard and utilized for negative purposes. The child will never be praised in public or boasted about; on the contrary, it will constantly be decried and complained about. A male child may be dressed as a girl and referred to in the feminine, since females have less prestige. Such practices against the evil eye are especially likely to be followed if there have been previous infant deaths in the family or if there is only one child.

To arouse disgust in the eye of the beholder is far healthier than to arouse admiration; and praise when given must be denied or deprecated. Inquiries about personal or family health, business, or status should be responded to with shaking heads and gloomy predictions. Boasting is considered the fool's way of courting disaster, as is the disclosure of one's future plans. It is possible, however, to accept praise or note good looks, good health, or good fortune if one is careful to constantly invoke the name of Allāh and to deny the force of the evil eye.

Particularly valuable as a defense against the evil eye are amulets. Blue beads are the most common type, and they may be worn on the person or placed in the house or on a dog, horse, cart, or automobile.

Curative practices

If, despite such precautions, evil spirits do gain access to the body and illness sets in, medicine is brought into play. Curative practices are clearly recognized, however, as being less successful ultimately than preventive measures, and their purpose is partly to provoke emotional comfort and security to the patient and his family.

The local practitioner gives the family his undivided attention, identifies and names the disease, makes a positive prognosis, and initiates certain measures to evict the evil spirits or draw away the evil eye. These include smoking, drinking, chanting, praying, burning, bloodletting, emetics, purgatives, and massages. A burning blue rag may be snuffed and the smoke inhaled to weaken or frighten out the evil spirits, especially during childbirth. Charms and holy phrases written on paper may be soaked in a liquid and then drunk in order to internalize the holy power. The spittle of a holy man may be applied to the disturbed organ of the body. The patient's name may be changed so that the evil spirit may somehow be misled and lose the patient or never find him.

Drastic, painful measures may be taken to force out the spirit. A red-hot nail may be pressed against the abdomen of an infant a number of times to force out the evil causing dysentery; it may be pressed against other parts of the skin to evict the evil spirit causing smallpox or rheumatism, or it may be pressed against a "boil" under the tongue to enable the baby to take the breast. The entire family may be required to chant special prayers, songs, or phrases or to fast or suffer other discomforts. Parents may have to

abstain from sexual relations, and members of the family may be required to travel on a pilgrimage. There may be bargaining on the part of the family, and precious animate or inanimate objects may be sacrificed to appease the evil spirit. The patient, recipient of this rich fund of strength, is as a result emotionally able to endure the physical discomfort during the prescribed course of curative treatment and is mentally prepared for possible death should the treatment not succeed.

If the patient should die it is because he or his family, consciously or unconsciously, committed offenses and attracted the evil eye or permitted evil spirits to enter the body to the extent that no power was able to avert the evil and save the patient.

It is clear that both patient and practitioner operate within a cultural framework of knowledge, beliefs, and values that explain their respective actions. The patient and his family search their thoughts and actions to ascertain how the misfortune could have occurred or by whom it could have been inflicted. The practitioner, on his part, listens to the family's statements and the patient's complaint and then formulates his diagnosis and treatment. The success in such treatment in a certain number of cases inspires further faith in the system. There are few cynics among the patients and few charlatans among the practitioners.

Types of medical practitioner

Within this medical system may be distinguished three or four types of local practitioners specializing in particular areas and methods of treatment. Their sex, role, or status may vary, or they may specialize in cures for external complaints such as sores and wounds; or they may practice preventive and curative medicine for the internal body. The latter is generally a male of senior years and of religious–medical standing. To fulfill his role he must be well versed in the concepts of animism and animatism. Such knowledge alone, however, is not sufficient; he must also have a personality that will inspire confidence and elicit information from the patient and his family. In addition, he must be aware of the significance of symptoms and complaints, certain of which may or may not be expressed. He must possess to a high degree the quality of sensitivity and intuitiveness, for many of his diagnoses will be based upon implicit understanding and tacit agreement. He must have the ability to quickly analyze and evaluate the relations between the patient, his family, and his community.

Modern developments. A process of Westernization, modernization, and development has increasingly come to dominate Southwest Asia. The evolution of the traditional folk culture of the fellahin and the Bedouin is being radically altered, and even rural peasants want to live the supposed good life that is made possible by Western technology. The town in the region is the focal centre of this Westernization process, and from there it is dispersed throughout the countryside and into the desert. It is useful, therefore, to examine the traditional town and the ways in which it is changing.

Classical urban pattern. Urban life began in this area of the world, and it is not uncommon to find towns still extant that have been occupied continuously or intermittently over the past 4,000 years. While the Bedouin and the fellahin have long remained rather constant in their culture, the towns have constantly reflected the effects of political changes, military invasions, and population movements. Thus, whereas the nomads and villagers were comparatively homogeneous in both population and occupational structure, the towns contained separate quarters (*harat*) for various ethnic and religious populations and also contained a range of occupations and social classes.

The tradi-ional Arab town

The typical Arab town of the Middle Ages consisted of a mosque, a palace, a bathhouse, a school, a *khān* (hostelry for foreign merchants), a hospital, and a *maydān* (an open field for the horse and cattle market). Two main thoroughfares ran through the town at right angles to each other, leading from the four gates of the thick and heavily fortified walls. From the centre a network of narrow streets and alleys radiated into the residential quarters. Each of these quarters was also walled, frequently with its own bathhouse and school, as well as a mosque, church, or synagogue. The cemeteries and the *maydān* were outside the town walls.

Such was the typical town during the classical Arab period. With the conquests of the Ottoman Turks in the 16th century came urban poverty, exodus, and decay. Only within the 20th century has there been a redevelopment of the urban centres.

20th-century urbanization. Rapid, unplanned, and uncontrolled urbanization is a marked feature of the contemporary Middle East. The 3 percent annual rate of population growth (the population of the region is apparently growing faster than any other major region of the world, except for Latin America) is causing great pressure on available agricultural land and is pushing the excess population into the urban centres. The growing difficulty in finding additional land for cultivation and the chronic weaknesses in the implementation of agricultural reforms are aggravating the already serious situation of large numbers of a redundant, underemployed rural labour force.

Other factors are also at work in the acceleration of the urbanization process: improved roads and transportation have made the towns increasingly accessible; increased communications, particularly through radio and television, have spread the attraction of the cities as centres of education, health, employment, recreation, and a totally different way of life; and the discovery and exploitation of oil have brought thousands of workers into the cities in search of employment. The flow of capital from oil revenues has revolutionized the cultures of some of the countries of the area, particularly Saudi Arabia and Kuwait, and it has had a direct impact on the economy of cities in nearby countries as well.

Israel holds third place in the world in proportion of urban dwellers to total population. In Kuwait, some half of the total population lives in localities of 100,000 and more inhabitants, and, in Lebanon and Syria, the proportion is one-third. This large-scale urbanization has been primarily a 20th-century movement, particularly since World War II.

Urbanization has resulted in a series of physical changes and problems. The uncontrolled growth of what were for the most part medieval towns has led to their mushrooming over surrounding areas at a rapacious rate. Most cities have doubled in size within the past few decades, and some—Beirut, for example—have incorporated surrounding towns. Greater Baghdad and Kuwait have spilled over their city walls. In Israel, almost the entire coastal plains from Haifa in the north to below Tel Aviv-Yafo in the south is gradually evolving into one continuous metropolitan area. The megalopolis, as well as the metropolis, is becoming a significant term in the Middle East.

Problems accompanying urbanization

As in many other regions of the world, acute urban problems have accompanied this movement. Traffic in urban land, with building for quick, speculative profit, has become a major economic activity and has often had the result of greatly modifying traditional values and behaviour. Slums and shantytowns, a hallmark of metropolitan growth in underdeveloped countries, are extensive. In the Middle East, rural migrants build simple one-room shacks. Lacking sanitary amenities, these are crowded next to each other and quickly deteriorate into unhealthy slums. Sections of greater Baghdad, for example, are ringed by slum dwellings mainly of the *ṣarīfah* type—a one-room house constructed mainly of reed matting. It is estimated that there are more than 44,000 *ṣarīfah*s in greater Baghdad—nearly 45 percent of the total number of houses.

The emergence of a large and rapidly growing unskilled (and therefore underemployed or unemployed), illiterate, and unhealthy peasant population in the towns may create a new version of the lumpenproletariat. With little to lose and much to gain, these people may profoundly and radically alter traditional religious and familial values and behaviour. Juvenile delinquency, petty thievery, prostitution, and mob violence frequently accompany such changes. Violence, of course, may be aggravated by political figures. Acute xenophobia, directed particularly at traditional "imperialist" powers, and the brutal elimination of incumbent leaders are two of the more obvious by-products. During the July 1958 revolution in Iraq, for example, street mobs ran amok in Baghdad, pillaging and burning at will. Scores were slaughtered, including King Faysal II and Prime Minister Nuri as-Said.

In certain respects the general life of the majority of the people has been materially improved. The traditional scourges of the region—disease, poverty, and ignorance—are slowly weakening their grip. Clothing, housing, household effects, kitchen utensils, communication equipment such as radio and television, and transportation facilities are becoming increasingly available from the cities of the coast to the towns and villages of the interior and even to the peoples of the desert.

New social classes have developed, and the traditional cleavage of a small wealthy class supported by a mass of poverty-stricken peasants and nomads has been somewhat modified. The upper classes are frequently educated in the West, and their financial resources, their educational interests, and their way of life all orient them away from their own country's traditions.

The new middle class

A new and increasingly numerous middle class has developed. The traditional class of craftsmen, merchants, artisans, and professionals has been augmented by a rapidly growing white-collar population of educated persons who go into law, politics, journalism, clerical and higher administration, and religious and teaching positions. This class frequently manifests an aversion to rural life and earning a living by one's hands.

As a result, the emerging middle class of the Middle East, it has been argued, is lopsided when judged by Western standards. It has a profusion of white-collar workers and "intellectuals," many of whom are chronically unemployed or underemployed and underpaid. Conversely, this class lacks a sufficient number of doctors, engineers, architects, chemists, and technicians.

Disenchantment with many of their own cultural values and with many aspects of their way of life and disillusionment with much of what the West has to offer have resulted in pronounced dissatisfactions and frustrations. To the extent that such general discontent prevails, one might predict a continuation of the social and political instability characteristic of much of the area for the greater part of the 20th century. (An.S.)

The economy

RESOURCES

Continental immensity and geological diversity explain the mineral wealth of Asia, which includes reserves of almost every important mineral. Abundant reserves of coal, oil, natural gas, and uranium, iron, bauxite, and other ores are either being exploited or await development; much wealth also remains to be surveyed. Difficulty of access, however, sometimes constitutes a barrier to exploitation.

Mineral resources. *Coal.* Asia has enormous reserves of coal, amounting to more than half of the world's total, but they are unevenly distributed. The largest reserves are found in China and especially in the Asian part of the Soviet Union; Taiwan, Japan, North Korea, South Korea, Vietnam, Indonesia, and India have smaller but economically important reserves. Burma, Thailand, Malaysia, and the Philippines have only insignificant amounts of poor coal. In Southwest Asia both Turkey and Afghanistan have small economic reserves.

Chinese coal reserves are chiefly high-grade coals. Every province has at least one coalfield, but the largest reserves are in Shansi and Shensi in the north. Szechwan, Shantung, and the Northeast (Fu-shun, in Liaoning Province) are old coal-producing regions with good reserves, and a coal-mining region with large reserves has been developed in central Anhwei, north of the Yangtze River. Mines in Ningsia and Kansu supply northern industrial plants, but their reserves are not clearly known. The long-known reserves in western Hopeh are being exploited.

Coal reserves in Siberia

In the Soviet Union are found the world's largest proven coal reserves, but the extent and quality of Siberian deposits are not fully known. There are more than 200 fields that are worked, but as new economic developments occur, the regional mining picture shifts somewhat according to the quality of the coal and the cost of transport. The brown coals of the Moscow Basin and the higher quality coals of the Donets Basin of the eastern Ukraine continue to be important in the west, and the Vorkuta field west of the Urals and south of the Arctic coast helps supply the western zone. The Ural Mountains are not rich in coal, but there are some small fields of lower grade coals. The Karaganda fields in Kazakhstan in the southeast have huge deposits, but the coal is high in ash, and mining there is not expanding since newer sources of better coals exist in Western Siberia. The Ekibastuz field, north of the main Karaganda fields, is a producer of high-quality coal.

Most of the known coal supplies of the Soviet Union lie in Siberia. The Kuznetsk Basin in Southwestern Siberia has become a giant producer. The Minusinsk Basin in the central region of Western Siberia, the Kansk region to the north along the Trans-Siberian Railway, the Cheremkhovo area west of Lake Baikal, and the Bureya Basin in the southeast are the major areas of production. Many smaller deposits are worked to supply local regions, such as the small and scattered fields north of Vladivostok, on Sakhalin Island, or in the hilly valleys of southeasternmost Turkistan.

Petroleum and natural gas. At least two-thirds of the world's known oil and gas reserves are found in Asia; the proportion may prove higher with the continued exploration of Siberia and the seas of southeastern Asia. Many of the island chains bordering eastern Asia have geological formations favouring petroleum accumulation, and oil fields are in production in Sumatra, Java, and Borneo. Western Asia has the largest known oil reserves, located in Iran, Iraq, and the Persian Gulf area of Arabia. Other regions in Southwest Asia have only small amounts of oil, and known petroleum reserves on the Indian subcontinent are small.

Malaysia is the only important oil-producing area on the mainland of Southeast Asia, although offshore waters may yield production after further exploration. The Philippines are negligible as a producing region, and the petroleum production of Japan is also small. North and South Korea appear to have virtually no prospects of production, but China has a number of oil-producing fields in Szechwan, Kansu, Sinkiang, and the Northeast. The Tsaidam Basin in northwestern Tsinghai Province is also a producing region. Some oil has been produced regularly from oil shales found in the Northeast, and natural gas is exploited in Szechwan.

The Soviet oil and gas fields

In the Soviet Union, Siberia has overtaken Southwest Asia in the production of oil and natural gas. The older fields lie in the southern Volga Basin and in the margins of the Caucasus Mountains. The flanks of the Ural Mountains have a number of large oil fields and small gas fields. The northern Volga Basin, along the western flank of the Ural Mountains, once contained the leading producing regions for oil. Major gas fields are located in the northeastern Ukraine south of Kharkov and in the Carpathian foothills of the western Ukraine. Uzen, on the Mangyshlak Peninsula on the eastern shore of the Caspian Sea, is a major gas-producing field that also yields oil. Another major field is that of Gazli in the Kyzyl Kum Desert south of the Aral Sea, and the rich gas field in the northern Ob River Basin at Berezovo indicates that the entire Ob Basin may yield natural gas. In the Lena River Basin, north of Yakutsk, there are large proven gas reserves.

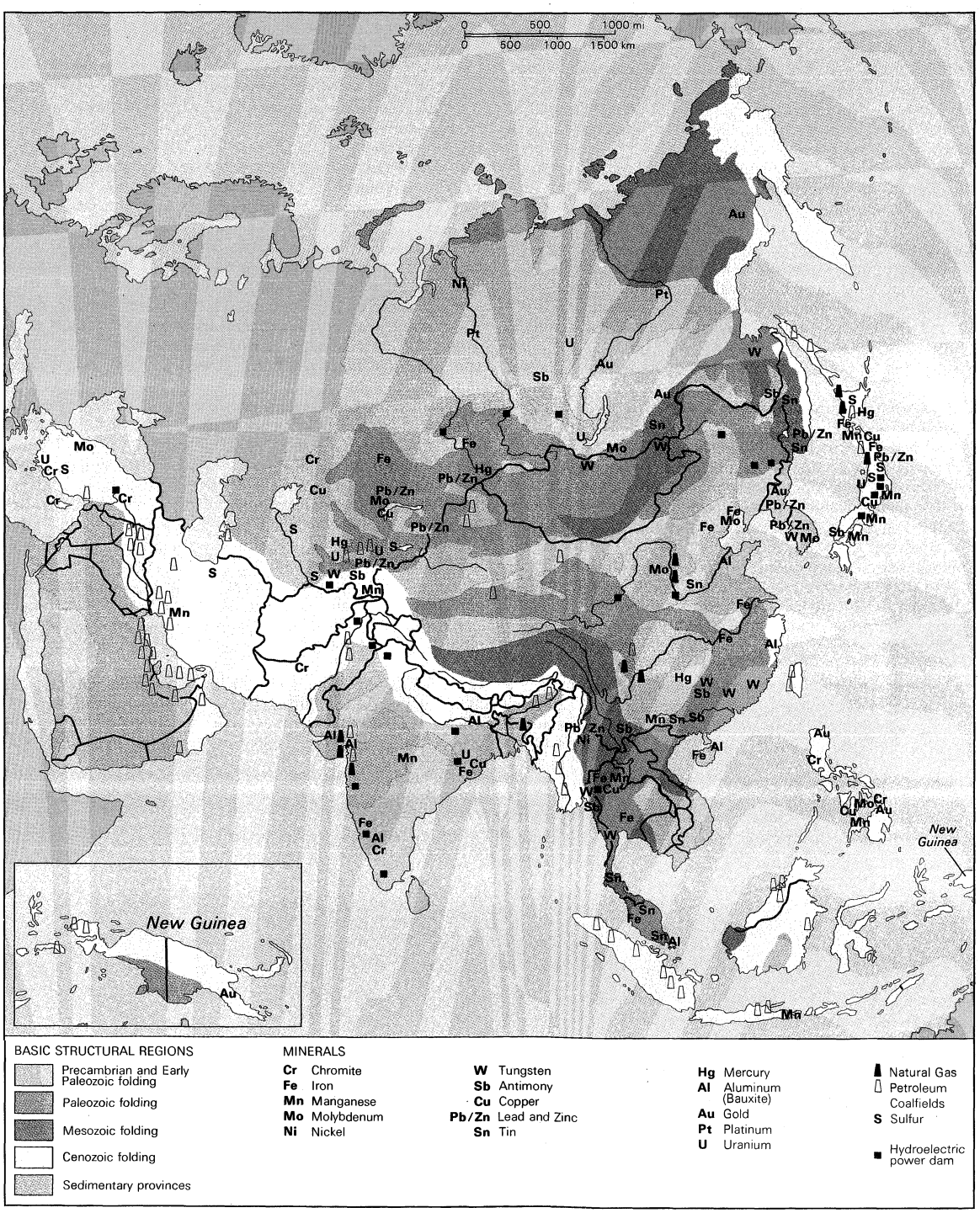
Uranium. Reserves of uranium ore are found in Asia's ancient crystalline rocks. The Soviet Union, China, and India have their own supplies of uranium. The Soviet Union, in particular, has rich ore fields in Kirgizia, between Osh and Tuya Muyun. Chinese uranium resources are probably in northern Sinkiang and southern Hunan.

Iron. All portions of Asia have deposits of iron ore, although not every political state has its own private supply. South Korea, Taiwan, Sri Lanka, and several smaller countries in Southwest Asia appear to have only small iron-ore supplies. Japan has far less than needed by its large iron and steel industry and depends largely on imported supplies. The Philippines has more ore than needed by its modest industrial needs and is an ore exporter. Malaysia produces a considerable volume. Thailand, Burma, and Pakistan have fair amounts of relatively low-grade ores, and Vietnam and Turkey have good ores in substantial volume. Indonesia and India both have large deposits of good iron ores that are reasonably distributed.

Although formerly regarded as deficient in iron ores, China contains huge quantities of varying grades of ores that are widely distributed and often located close to coal supplies. Regional centres of ore mining, smelting, and fabrication are located at An-shan, a southern area of the Northeast; near Peking; in southern Anhwei, west of Shanghai; in central China, east of Wu-han; in southern Inner Mongolia, north of Pao-t'ou; in central western Kansu; and on Hainan Island, off the southern coast. Large iron ore deposits also occur near Chungking in Szechwan. Iron ore in small local volumes is widely located in Kweichow and Yunnan in the Southwest. China ranks among the world's major producers of iron ore.

The Soviet Union formerly depended largely on iron ore from the Krivoy Rog and Kerch basins of the southern Ukraine, but huge deposits of magnetic ore have been found near Belgorod. Iron ore has long been extracted from the Ural Mountains, and further deposits have been found at and near Magnitogorsk in the Southern Urals. There appears to be a virtually unlimited supply of low-grade ore in the Kustanay Basin east of the Southern Urals in southwestern Siberia. A major iron ore range, at Kachkanar, west of the Northern Urals, contains low-grade ores. Large deposits of medium-grade iron ore have been found northwest of Lake Baikal, close to the Cheremkhovo coal deposits. Smaller deposits have been located in the Murmansk Peninsula and at several locations in Eastern Siberia. The Soviet Union has become the world's largest producer of iron ore.

Ferroalloy metals. Asian resources of nickel are not extensive. There is a notable Soviet nickel-ore field at Norilsk (in Northern Siberia); Indonesia also possesses reserves. Japan and the Philippines both produce considerable quantities of nickel.



Basic structural regions and principal mineral and hydroelectric sites of Asia.

Asian countries with reserves of chromium include Turkey, the Philippines, India, Iran, Pakistan, and Cyprus; reserves are also found in Soviet Asia.

Manganese is found in abundance. The Soviet Union has large reserves in Central Asia and in Siberia, and India also possesses large quantities. Chinese reserves are considerable, but those of Japan are limited.

Southern China has exceptionally large reserves of tungsten. Tungsten reserves of Soviet Asia are also important, as are those of molybdenum.

Copper, tin, and zinc

Nonferrous base metals. Asia is not richly endowed with copper ore. In Soviet Asia the principal fields are Almalyk, southeast of Tashkent; Dzhezkazgan, west of Karaganda; Kounrad, Lake Balkhash; and in the Kuznetsk Basin. Japan's widespread copper-ore reserves are of medium importance, and the Philippines have limited reserves. China has deposits in Kansu, Hopei, Anhwei, and Hupeh, but production is insignificant. Turkey, Israel, India, and North and South Korea have small reserves.

Significant reserves of tin exist along a north–south axis running from southwestern China through the Malay Peninsula to Indonesia. Thailand, Burma, Vietnam, Laos, and Yunnan Province in China also have deposits of tin. Soviet Asia has substantial reserves in Transbaikal and also in the Sikhote-Alin Range.

Soviet Asia's lead and zinc reserves—the largest in Asia—are located in the Kuznetsk Basin and in central and eastern Kazakhstan. China also has abundant reserves of zinc and lead ores, and North Korea has important lead resources.

Asia has enormous reserves of bauxite. In Soviet Asia bauxite fields are located in Kazakhstan and in the Sayan Mountains. There are also large reserves in India, Indonesia, the Philippines, and Malaysia, as well as significant reserves in China.

Important quantities of mercury occur in south central China and in the Soviet regions of the Ukraine and Siberia. Magnesite is common in Asia. There are large reserves of antimony in central China; Turkey and Thailand also have substantial reserves.

Gold and silver

Precious metals. Many Asian countries have produced gold from alluvial stream deposits in past centuries, and some continue to do so. Small volumes of alluvial gold are produced in Burma, Kampuchea, and Indonesia, and the headwaters of the Yangtze River in the Tibetan border region yield some gold. India formerly was a large producer of gold from lode mines, but the best ores appear to have been exhausted. Japan, North and South Korea, Taiwan, and the Philippines have significant gold-ore reserves and periodically produce gold from small lode mines.

The Soviet Union has produced gold from lode mines in the Central Ural Mountains for centuries, and in the 19th century there were several gold rushes to work alluvial stream deposits in Siberia on the Lena and Yenisey rivers. In the late 20th century, Soviet gold production was considerable, and lodes were worked in several Siberian locations, centring on the upper reaches of the Kolmya River. The lode at Auezov in eastern Kazakhstan, south of Semipalatinsk, potentially allows the Soviet Union to rival South Africa in rank as a gold producer.

Platinum is mined near Norilsk in the Central Siberian Plateau in Northern Siberia.

Nonmetallic resources. Reserves of asbestos are localized; it is abundant in China, South Korea, and Soviet Asia. Mica is abundant in Soviet Asia and is also found in large quantities in India. Asia has abundant reserves of rock salt, but the hills and "glaciers" of salt in southern Iran have not been exploitable. Reserves of sulfur and gypsum are abundant in Central and West Asia. Japan has large reserves of sulfur. In Vietnam phosphates are obtainable from sedimentary fields and from apatite. The Soviet Union has large deposits of phosphates in the Mangyshlak Peninsula on the eastern shore of the Caspian Sea and other scattered deposits of lesser value. Diamonds are produced in east central Siberia.

Water resources. Asia's water resources constitute a vast potential, both for generating hydroelectricity and for irrigation. In the arid parts of the continent, water is primarily used for irrigation.

Siberian hydroelectric power

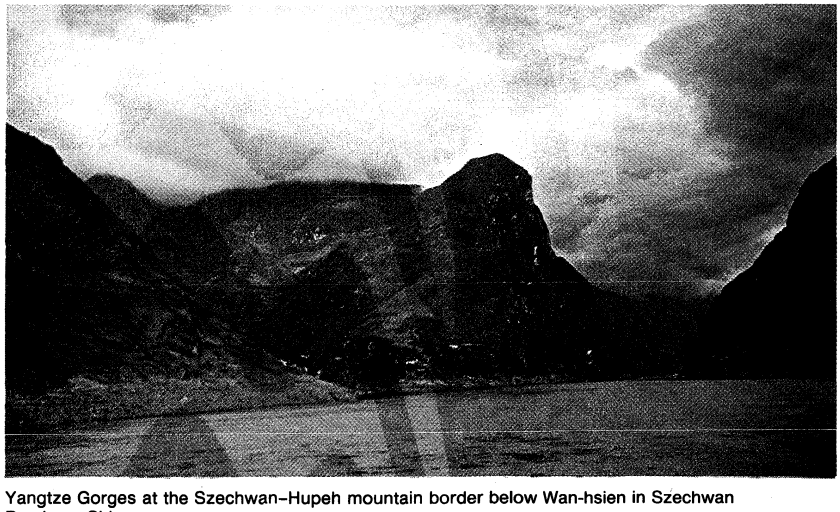
Siberian rivers have an excellent hydroelectric potential, for when dammed they provide low falls with an enormous volume of flow. Extreme cold and low winter water levels, however, hinder their exploitation. Thanks to abundant precipitation and great differences in water level, the Soviet Far East has an immense potential for hydroelectricity, although the remoteness of Eastern Siberia discourages industrialization.

Japan, a country of high man-made waterfalls but relatively small volumes of water flow, has already harnessed almost all its rivers that have a hydroelectric potential; this potential, however, is increased by heavy rains, particularly in summer.

The waterpower potential of northern China is extremely limited because the flow of the Huang Ho and other northern rivers is erratic, and all carry heavy volumes of silt. The hydroelectric potential of China south of the Tsinling Mountains, however, is great. The Yangtze River has a considerable waterpower potential, particularly near I-ch'ang, although this site would be expensive to develop.

The hydroelectric potential in the Indian subcontinent is subject to regional variations. The Western Ghāts, which slope down abruptly to the western maritime plains, permit high waterfalls; unfortunately, the rivulets that rise on the summit have an insignificant volume of winter flow. Rivers of the eastern slope of the Deccan, such as the Mahānadi and the Godāvari, lend themselves to the construction of dams with low falls and great volumes of flow, as also do the Himalayan rivers entering the Ganges Plain. The Himalayan ranges offer rich possibilities for the utilization of high waterfalls for generating hydroelectric-

Emil Schulthess—Black Star



Yangtze Gorges at the Szechwan–Hupeh mountain border below Wan-hsien in Szechwan Province, China.

ity, but in winter their waters are very low, and most of the sites would be expensive to develop.

Biological resources. Widely varying climatic conditions, particularly in the distribution of rainfall, have produced terrains in Asia ranging from tundra and desert to forest and alluvial plain, each supporting its appropriate plant cover, on which, in turn, typical animals and birds subsist. The Arctic north of the continent and large areas of the central mountain massif—known as "the roof of the world"—are practically uninhabitable. In addition, even where there is water—and nowhere is water conservation pursued more carefully than in Asia—there are still many areas of undrained swamp. Much else is desert. By far the greater part of Asia remains uncultivated. Prime resources are the extraordinarily intensive agriculture made possible by irrigation of the alluvial soils of the great river deltas and courses; the forests, with commercially valuable species of trees; the flocks of sheep and goats supported by Asia's semiarid deserts and grasslands; and the produce of the intensively fished surrounding seas of South and East Asia.

Timber resources. Much of Northern Siberia, south of the Arctic Circle, is covered by coniferous and mixed forest, which is commercially exploitable. The great deciduous forests of northeast India, Burma, Thailand, and Malaysia contain teak and other important hardwoods, as well as bamboo. Mangrove forests line the waters of the Ganges and Irrawaddy deltas and many small stretches of coast along the Malay Peninsula, Indonesia, and the Philippines. But in the Indian subcontinent lowland, forest has yielded place to cultivated land, as a result of population expansion; agriculture has similarly reduced the natural forest areas of China to insignificance, except in northern Manchuria. Japan, on the other hand, is relatively heavily forested in relation to its area and population, although much of the present cover is planted forest. More than half of the Philippines still carries heavy forest, but good commercial forests cover only one-third of the country. These forests produce valuable hardwoods and the soft "Philippine mahogany."

Crops. In Soviet Asia, the black-earth belt across Southern Siberia is cultivated with grain crops, of which wheat is the most important, as also are areas of the Soviet Central Asian republics. Grain crops, chiefly wheat, are cultivated in North China—where soybeans are also grown—and in Japan. Intensive use of water resources from wells, as well as from irrigated rivers, has enabled grain crops to be raised in Iraq, Iran, Pakistan, and northern India. The great staple of South Asia is rice. It is the chief food crop of Japan, South China, Taiwan, Southeast Asia, the Indonesian islands, the Philippines, Burma, Sri Lanka, and parts of India and Pakistan, and is found in Iran, southwestern Asia, and elsewhere.

Of plantation crops, rubber, from a Brazilian plant imported in the 19th century, is cultivated in Malaysia and Indonesia and also in India and Sri Lanka. Tea is grown on commercial plantations in the uplands of northern India and Sri Lanka for export and in China and Soviet Asia on small holdings for domestic consumption. Sugarcane is harvested in Java, the Philippines, India, and Central Asia; and tobacco is grown widely, notably in Turkey, Soviet Asia, China, and Indonesia. Citrus fruit is produced in the Mediterranean lands, in the Soviet Central Asian republics, and in China and Japan. Date palms are cultivated, particularly in Arabia. Licorice is grown in Turkey. Asia is also a producer of opium from the poppy.

Seminomadic pastoralism

Livestock. The uncultivated steppe lands and deserts of Central Asia and Mongolia support flocks of sheep and goats. Seminomadic pastoralism is the rule there, as it is in parts of Afghanistan, Pakistan, Iran, and Arabia. In Central Asia, the horse and the yak are the riding animal and the beast of burden, respectively; in Arabia, the camel is both. Cattle are raised in agricultural areas. Hides, wool, and other animal products are important economically. Reindeer herds are kept in the northern tundra of Siberia, where they feed on mosses and shrubs. In Siberia, valuable furbearing animals have long been hunted. In India, Burma, and Thailand elephants still work as draft animals in the lumbering industry; particularly in Southeast Asia, the water buffalo is an important draft animal as well as a milk and butter producer. Angora goats are herded in Anatolian Turkey to provide the silky mohair for which they are noted. Silkworms are reared for silk in China, Japan, India, and Soviet Central Asia.

Game birds. North of the Himalayas, such game birds as ptarmigan, grouse, plover, and various kinds of waterfowl are found. South of the Himalayas, pigeons, pheasants, and other game birds are taken. Various kinds of hawk and falcon, trained to hunt, have their habitat in Arabia and other parts of Asia.

Seafood. Fish and other sea creatures and various kinds of crab and shrimp are intensively fished off the coasts of China, Japan, and Southeast Asia. The sturgeon, prized for caviar, is fished commercially, particularly in the Caspian Sea and the rivers of Siberia. (P.Gu.)

Factors in resource development

Resource development. The utilization of Asia's natural resources has depended, to a large extent, not only on the development of technology but also on political circumstances. Thus, until the end of World War II and the beginning of the process of decolonization in Asia, most Asian countries were not free to develop their own natural resources independently and without reference to the economic interest of a metropolitan power. Cultural attitudes also affect the utilization of resources. Cattle, for example, which are a source of immense wealth in many developed countries, are a drain on scarce resources in India, where cultural taboos prohibit the slaughter of cattle either for food or for the conservation of resources when the animals are no longer productive.

The value of natural resources also varies with the prevailing technology. For example, with the application of new technology to the production of cereals, the same area of land can give greatly increased yields. The application of modern technology has also produced improvement in many other areas, such as in Japan for the production of silk or of cultured pearls. Technology may also make it possible to exploit mineral wealth that was previously unusable because of problems of accessibility or of juxtaposition of other minerals.

INDUSTRY

Mining. Asia produces a variety of minerals. Many of these are mineral fuels, such as coal and petroleum. The largest Asian producers of coal are the People's Republic of China and Soviet Asia, followed by India and Japan. Very small quantities of coal are produced in a number of other countries. The Arab countries of Southwest Asia are the principal producers of petroleum in the world. The biggest producer among the Asian countries is Saudi Arabia. Soviet Asia was second, followed by Kuwait, Iran, the United Arab Emirates, and Iraq. The biggest producer of natural gas is Soviet Asia, followed by China, Saudi Arabia, Iran, and Indonesia.

The largest producers of iron ore and ores for ferroalloys are China, Soviet Asia, India, and North Korea. Together these four account for almost all the production of the region. India and China are among the major world producers of manganese ore and between them account for virtually all of Asia's output. Asia's biggest producer of chromite is the Soviet Union, followed by the Philippines, Turkey, India, and Iran. There is also some production of tungsten in China, the Soviet Union and North and South Korea, while nickel is mined in Indonesia, the Soviet Union, and the Philippines. Soviet Asia has become an increasingly important producer of many of the ferroalloys.

Asia is one of the world's main producers of tin-in-concentrates (tin ore that has been partially processed to increase the concentration of tin), providing more than half of the world's total production. Malaysia alone accounts for about half of Asia's production, followed by Indonesia and China. There is also considerable production of copper ore in the Soviet Union, the Philippines, China, and Japan.

The bauxite produced in Asia represents only a small part of total world production, although production in Soviet Asia has increased. Development of the Eastern Siberian gold mines has given Soviet Asia a leading position in the world's production of gold. Asia produces more than

one-tenth of the world's sulfur, principally from Japan and China. Asia also accounts for much of the world's production of graphite, from North and South Korea and from China.

Heavy industry and engineering. Despite the fact that the continent has such a variety of mineral resources, metallurgical industries have not been fully developed, except in Japan and in Soviet Asia. The major producers of steel in the region are Japan and China. Japan, China, and India are the major steel consumers, although the consumption of steel is increasing in Soviet Asia, Pakistan, and the Philippines; Hong Kong is one of the main consumers on a per capita basis. Japan, China, and India are also the region's leading producers of metallurgical coke.

The production of aluminum is concentrated in three countries—Japan, Soviet Asia, and China. India has a relatively well-developed aluminum industry. There is also some production of copper, zinc, lead, and tin in Asia, with Japan and Soviet Asia leading in the production of zinc and lead and Malaysia in the production of tin. Japan, China, and India are leading consumers of tin.

Japan produces every variety of engineering goods, from tankers and locomotives to miniaturized electronic equipment. Since World War II India has also gradually diversified its engineering industries and now produces heavy capital goods (machines and tools used in the production of other goods), various types of industrial machinery, prime movers (engines and other sources of motive power) and boilers, diesel engines, sewing machines, machine tools, agricultural machinery, and all types of electrical equipment. India also produces radio receivers, metal manufactures, railway rolling stock, automobiles, bicycles, and precision instruments. China has also made considerable progress in the field of engineering industries. Other Asian countries have primarily concentrated on the production of durable consumer goods.

Asia's engineering industries

Chemical and petrochemical industries. The consumption of nitrogenous and phosphatic fertilizers has greatly increased in Asia, largely because additional countries have begun to use the advanced techniques and improved seeds that have now become available. The major consumers of fertilizers, on a per acre basis of arable land, have been Japan, South Korea, and Singapore. Because of their vast size and the increased use of fertilizers, India and China are, in absolute terms, among the major consumers. India has greatly increased its production, especially of ammonium sulfate, and has also experimented with fertilizers that have a much higher nitrogen content, such as urea. Production of phosphatic fertilizers has also been increased in Asia.

Asia also produces and consumes basic chemicals, such as caustic soda, soda ash, and sulfuric acid; Japan and China are the leading producers of these, followed by India.

The consumption of pulp and paper throughout the continent has grown steadily, largely because of higher standards of living. The major consumers are Japan and India, and the major producers are China, Japan, Soviet Asia, and India.

Various surveys undertaken under the auspices of the United Nation's Economic and Social Commission for Asia and the Pacific (ESCAP) have shown that there is considerable scope for the manufacture of petrochemical products in Asia.

Manufacturing and textiles. The textile industries, particularly cotton, have expanded greatly in Asia since World War II. Japan and Hong Kong are among the world's largest exporters of cotton textiles, and China, Taiwan, Singapore, Pakistan, and India have also entered the international market. The industry produces cotton yarn, cloth, and finished garments. There is also some processing of wool (both yarn and woven fabrics) in the region. China, India, and Turkey are among the main producers and consumers; China is Asia's chief producer of woollen fabrics. Japan and India have also become major producers of woven rayon and acetate fabrics. Japan has also turned to noncellulose synthetic fibres, especially nylon, acrylic, and polyester fibres.

The expanding textile industry

Industrial development in the region has made relatively significant progress, though in absolute terms progress has been limited; in relation to its size and vast population, the contribution of Asia to total world industrial output has been small. There is, however, a discernible trend toward a transition from light to heavy industry in many countries of the region; the development of Soviet Asia is dramatic, and it is likely that the continent will have an increasing share of world production.

Timber, fisheries, and animal husbandry. Logs are exported from China, Soviet Asia, Malaysia, Indonesia, and the Philippines to industrialized and timber-deficient countries, especially Japan. Thailand and Burma produce special varieties of timber such as teak. Thai teak is also exported to Europe.

Soviet Asia has an enormous forest area. The wood ranges from pine around the Bratsk area to a mixture of pine, larch, aspen, birch, and other species in the region south of Lake Baikal. Logging and transport operations are highly mechanized and have been facilitated by a road-building program.

Bamboos are an important component of wet evergreen, moist deciduous, and dry deciduous forests in the tropical parts of Southeast Asia, principally in Burma, Kampuchea, Sri Lanka, India, Indonesia, Laos, Malaysia, Papua New Guinea, Pakistan, the Philippines, Thailand, and Vietnam. At higher altitudes and in temperate climates in Asia, as in Bhutan, China, Japan, and Nepal, many of the genera found in tropical parts are represented by different species, and other genera are common in China and Japan. Pure bamboo forests are common on slopes where temporary cultivation has been carried on in Burma, Bangladesh, and other parts of Asia.

Asia has a considerable potential for increased development of its fisheries. Japan has shown how far afield a well-organized fishing fleet can go in search of fish. In general, the problems of the fishing industry stem from lack of adequate capital and advanced technology, which tend to restrict fishing to coastal and offshore areas and make it difficult to extend the fishing to the deep seas. Because of a widespread lack of refrigerated transport and storage, there are also the problems of preserving fish after the catch and of transporting the catches to centres of consumption. In some countries freshwater fish are also an important addition to the diet of the local people; the raising of fish in culturally controlled ponds is important in southern China, Indonesia, and the Philippines. While the dairy industry is important in a few countries such as India, Pakistan, Soviet Asia, and Turkey, there is not much large-scale development of beef-cattle farming; Soviet Asia, however, has attempted to develop such patterns. Both China and Japan are discarding their traditional taboos against the use of milk products, and both countries have growing urban dairy industries. China, the world's largest producer of pork, is the only producer in Asia, with Japan as a distant second. The poultry industry has made rapid strides during recent years, and the production of both eggs and broiling chickens has gathered considerable momentum. The availability of feed for poultry is one of the major limiting factors in growth and development of the industry. Straw, obtained from rice crop, is the primary fodder for livestock in southern Asia. Cattle feed is usually supplemented by concentrates, such as oil cake.

The raising of livestock

The raising of sheep and goats for meat and wool is especially important in China, India, Pakistan, and Iran; these animals are also raised in practically all the other countries of Asia. The sheep population of Southeast Asia is small.

In spite of the large number of cattle and sheep in the region, the hides and skins industry has not been fully developed. Technological problems in connection with both the flaying of skins and their curing have not been overcome.

Handicrafts. Traditional cottage industries and handicrafts continue to play an important role in the economies of all Asian countries. They not only constitute an important manufacturing activity in themselves but also are often the only available means of providing additional employment and of raising the level of living for both rural and urban populations. In view of the growing world market for products of traditional Asian cottage industries and for Asian handicrafts, there is room for considerable

expansion, especially in standardizing production and in marketing products in advanced countries.

Other industries. Asian countries are at different stages in developing their pharmaceutical industries. The progress of the industry in Japan is comparable to that achieved in western Europe and the United States. Great progress has also been made in India and, in some respects, in Pakistan, but these two countries have not reached the stage of being self-supporting in technology, raw materials, or equipment. China has begun to develop an industry based on a distinctive blending of Occidental and native pharmaceutical manufacturing. In most of the other countries the pharmaceutical industry is only a processing industry based on basic drugs, imported in bulk, which are then marketed as capsules, tablets, and injectibles. In many Asian countries traditional medicinal products and treatments are still popular, especially in rural areas.

International tourism has developed significantly. The most visited places include Hong Kong, Japan, Thailand, China, Singapore, India, Pakistan, Turkey, Syria, and Israel. Hong Kong and Singapore each have a big entrepôt trade and attract visitors partly because they are duty-free ports. With the gradual lifting of the "bamboo curtain" beginning in the 1970s, the number of tourists visiting China has increased. There is considerable scope for further development of tourism in some Asian countries, for many of them have ancient monuments and natural attractions of great beauty. Unfortunately, many of the places that are most interesting to tourists are not easily accessible.

POWER

The per capita consumption of power in Asia outside the oil-producing countries of the Middle East tends to be very small compared with the world average. Japan is the biggest producer of power in Asia, and its generation capacity amounts to nearly half the region's total. The electricity output by public utilities and similar bodies in the Asian countries includes thermal power, hydroelectric power, and nuclear power. All these types of production have progressed rapidly in Soviet Asia.

Thermal and hydroelectric power

Thermal power has become the most important source of supply among the various types of electric power generated in the region, particularly in the Middle East. There is an increasing tendency to operate thermal power stations for meeting regular demand and to build—or plan—hydroelectric power stations to serve during peak demand periods. In countries such as Afghanistan and Sri Lanka, however, hydroelectric power generation is well developed and is several times greater than the generation of thermal power.

Nuclear power plants have been developed in a number of countries. Japan completed its first atomic power plant in 1967, with a capacity of 160 megawatts, and has constructed more and more such plants. India commissioned its first nuclear power station at Tarapur in early 1969 with a capacity of 200 megawatts, and others were subsequently constructed.

In the construction of steam thermal power stations, there has been a tendency toward a larger unit capacity operating at very high pressure and temperature. The Soviet Union has made extensive use of geothermal power in Asia, with geothermal plants at Makhack Kala, Tashkent, Lake Baikal, and Kamchatka. Japan has two small geothermal plants. The only other Asian country to use geothermal power is the Philippines. Small gas-turbine generating stations have also been installed in many countries. Pakistan uses natural gas from the Sui gas field for both thermal and gas-turbine generation.

AGRICULTURE

The most important modern development in Asian agriculture is the evolution of new high-yielding strains of cereals. This development is being utilized in several Asian countries and has had a marked influence on the yield per acre of cereals since the late 1960s. Rice is the staple food crop among most Asian countries. Asia produces some 90 percent of the world's total supply of rice. Except in the Middle East, India, Pakistan, Afghanistan, Soviet Asia, and Malaysia, rice occupies more land area than any other single crop. In the Middle East, Afghanistan, India, Pakistan, and Siberia wheat is the dominant crop, while in Malaysia rubber occupies the greatest land area, with rice ranking second. The total percentage of land under rice cultivation, as compared to total arable land, is highest in Vietnam, Bangladesh, and Sri Lanka; it varies between 25 and 50 percent in most Asian countries outside the Middle East.

High-yielding cereals

In spite of the fact that the region is the world's largest producer of rice, most countries (among them Sri Lanka and Indonesia) are not self-sufficient in rice. Thailand, the Philippines, China, and Japan are notable exporters of rice. Barley is grown in China and India, among other countries. Corn (maize) is grown in China, Soviet Asia, India, the Philippines, Thailand, and North Korea. India, China, Pakistan, and Soviet Asia also grow sorghum and millet.

Asia produces several plantation crops, of which the most important are tea, rubber, coconuts, sugarcane, and pineapples. The major producers of rubber are Malaysia, Indonesia, Thailand, India, and Sri Lanka. India and Sri Lanka predominate in plantation tea production, and China, Taiwan, and Japan produce several types of tea on small holdings. Coconuts are an important crop in the Philippines, Indonesia, India, and Sri Lanka. India, the world's leader in sugarcane production, produces primarily for domestic use, whereas the Philippines, Indonesia, and Taiwan produce both for domestic consumption and export. The Philippines, Taiwan, and Malaysia produce pineapples, which are canned for export.

The continent produces a variety of tropical and subtropical fruit, mainly for domestic consumption. Transport facilities, where available, can be used only for limited distances. In view of the climatic conditions and the general lack of refrigerated transport, consumption tends to be seasonal and confined to areas close to centres of production. Among the main varieties of fruit produced are bananas, mangoes, apples, oranges and other citrus fruits, pineapples, papayas, and some specialities such as mangosteen (a dark reddish-brown fruit) and durian (a large oval fruit with a prickly rind, a soft pulp, and a peculiar odour). Taiwan, the Philippines, and Malaysia export bananas to Japan.

Except in a few countries, the canning of surplus fruit has been developed only to a limited extent. In view of the tremendous potential for the production of fruit, there is vast scope for increased canning for export, both of fruits and of fruit juices. The fruit canning industry has markedly increased in Taiwan.

The same factors affect the production of vegetables. Vegetables are produced mainly for local consumption, and only tubers can be transported over distances and stored for any period of time. In Taiwan successful efforts have been made in the canning of mushrooms and asparagus, both of which have come to be among the country's leading exports.

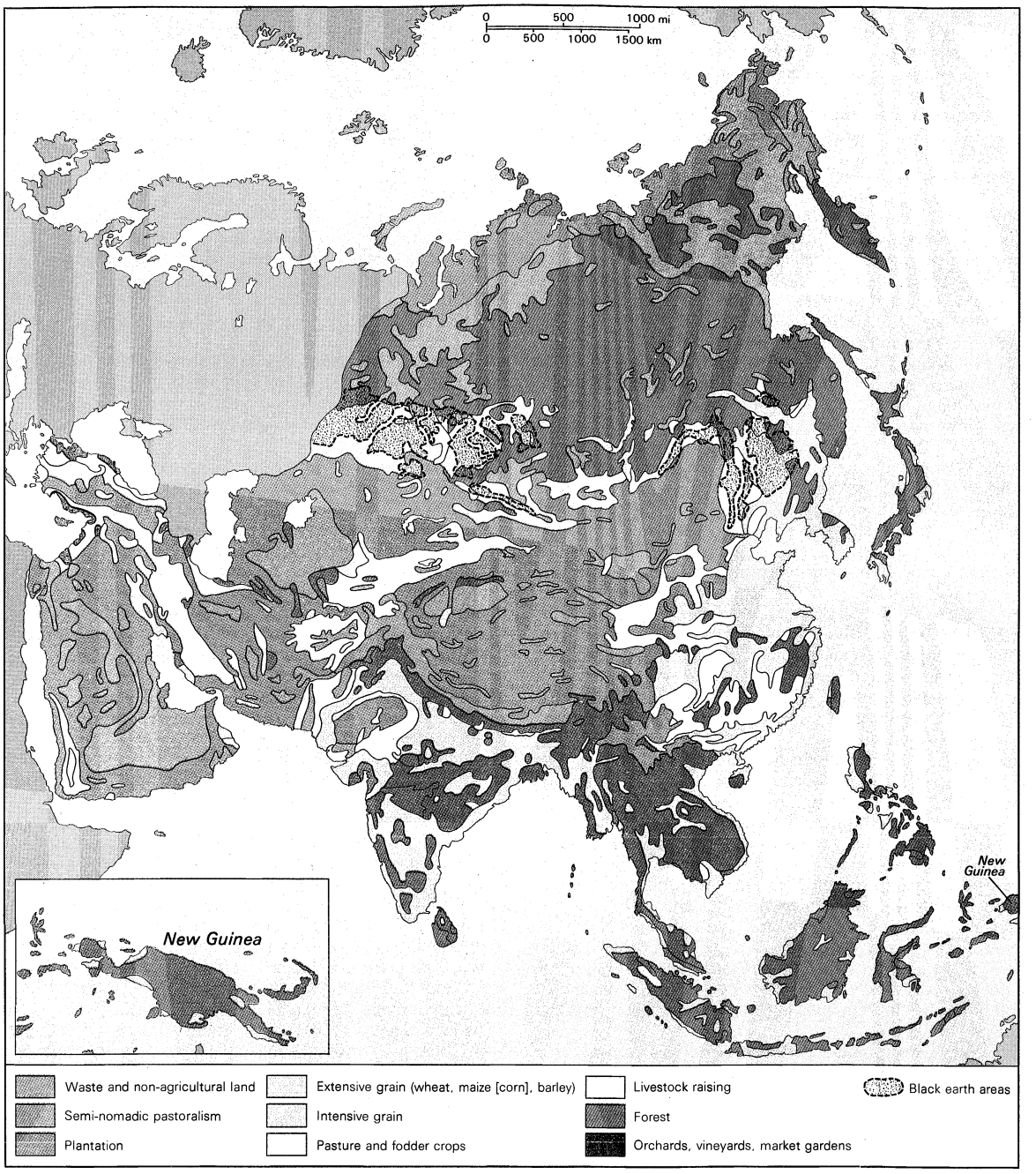
The traditional method of irrigation in Asia is by gravity water flow. The water from upstream storage reservoirs or diversion dams is carried through canals to field distributaries. The fields adjoin one another, and the water is able to flow from one field to the next; it may, however, take some time for the water to move across the terraced fields back to the canal system. The disadvantages of this system include the loss of water by evaporation and seepage and the possibility that the continuously flowing water will carry with it soil nutrients, fertilizers, and pesticides. In Japan and Taiwan water is moved by small electric pumps, which operate continuously during the growing seasons.

Methods of irrigation

By the late 20th century more attention was being given to the use of underground water by lift. The use of ordinary pumps as well as of deep-bore well turbine pumps has become common, especially in India, Pakistan, and Iran. Such irrigation avoids some of the disadvantages of flow irrigation and allows for easier drainage.

Of the crops cultivated in the region, rice, sugarcane, and, in Soviet Asia, sugar beets need the most water. Cereals other than rice, legumes, and root crops can be grown even on rain-fed land.

The availability of an assured source of water supply



Agricultural regions of Asia.

is an important element in the new technology, which also requires the use of fertilizer in conjunction with the improved cereal seeds that have been developed. Huge irrigation projects in Southern Siberia, southwestern Soviet Asia, and Pakistan are rapidly altering traditional agricultural patterns.

TRADE

Several centuries before the Christian Era, Asian countries had commercial relations among themselves as well as with the countries of the West. In the earliest days nomadic tribes carried on this trade over considerable distances, using barter as the medium of exchange. Particularly important in such trade were fine textiles, silk, gold and other metals, various precious and semiprecious stones, and spices and aromatic products. There was a considerable expansion of trade during the Greek era around the 4th century BC, by which time various land routes had been well established connecting Greece, via Asia Minor, with the northwestern part of the Indian subcontinent. Further development of land and sea routes, especially to South India, occurred during Roman times. This East–West trade flourished in the first four centuries AD but was subject to considerable vicissitudes in later centuries. During this period there was also a great expansion of trade to Southeast Asia and to China through what are now Malaysia and Kampuchea.

Ancient trade patterns

After Spain, in the 15th century, became interested in discovering a direct sea route to Asia—an interest that led to the discovery of the Western Hemisphere—the era of the great circumnavigators arrived in the 16th century. Portugal was one of the first countries to dream of establishing a monopoly over the lucrative spice trade with Asia. The Dutch and the British started similar enterprises at the turn of the 17th century, each country establishing its own East India company. The British began by centring their activities on the Indian subcontinent and extended their interest to Burma, Ceylon (now Sri Lanka), and present-day Malaysia. The Dutch first concentrated on Ceylon but later expanded into and concentrated on Southeast Asia. The French were able to establish only minor footholds on the Indian subcontinent, but their 19th-century

penetration of Indochina was more successful. Spain was content with the control it had established over the Philippines. Over the course of time these European trading companies developed into colonial empires.

The export trade of Asia was formed by these historical circumstances. In the days before the East India companies, the products of Asia were those demanding the exploitation of human skills, such as silk and textiles, and such precious commodities as spices and aromatic products, which depend especially upon climatic and soil conditions and careful human labour.

The colonial trading system

With the development of the East India companies and the further assumption of colonial rule, a new pattern of trade emerged. Generally speaking, the colonial countries became the exporters of raw materials and imported the finished products from their metropolitan rulers. During this period tea and tobacco also entered into international trade, and jute became a monopoly product of the Indian subcontinent. Until the end of the 19th century Japan had trading relations mainly with Korea and China and generally remained somewhat aloof from world trade. China was the first country of Asia to have any significant trade relations with the Western Hemisphere; these were chiefly with the United States.

The latter half of the 19th century and the early part of the 20th constituted the heyday of colonial rule. By the first decade of the 20th century, Japan had emerged as a major military and naval power and gradually developed into an important trading partner with the rest of the world. The era that followed was that of the colonies' struggle for political independence, which reached its climax immediately after World War II. Fifteen years after World War II the great British, French, and Dutch empires had virtually ceased to exist in Asia.

As a result of the achievement of independence by the former colonies, the emergence of the People's Republic of China, and the development of Japan into one of the major producing and trading countries of the world, the old established patterns of trade have changed considerably. Most of the newly independent countries have planned their economic and social development and have tried—with varying degrees of success—to diversify their production and to develop new industries, even if they cannot wholly overcome the handicaps of the colonial past.

Internal trade. In view of the division of labour that existed between the colonial countries and the metropolitan powers in colonial days, it is not surprising that the economies of the independent countries of Asia are often more competitive than complementary. In the case of certain countries, such as the Philippines, Sri Lanka, India, Afghanistan, Iran, and Pakistan, their intraregional exports have amounted to only a small fraction of their total exports. Iran, Afghanistan, Taiwan, South Korea, the Philippines, and India imported only about one-tenth of their requirements from other Asian countries by the early 1980s. Japan, the most developed of Asian countries, exported only about one-third of its products to developing ESCAP countries, and Japan received from these countries only about one-fourth of its imports.

Competitive economies

Asia is the biggest producer of rice in the world; rice is indeed one of the most important commodities of intraregional trade, and it is the most important export item of such countries as China, Pakistan, and Thailand.

There has been an effort on the part of Asian countries to improve their trading position by joining in commodity agreements. Malaysia, for example, is a member of the International Tin and Rubber Agreement. Other Asian countries are members of the International Sugar Agreement. The Asian Coconut Community was established in 1968, with Sri Lanka, India, Indonesia, Malaysia, the Philippines, Singapore, and Thailand participating. Bangladesh, India, Indonesia, and Sri Lanka were instrumental in the formation of the International Tea Promotion Association in 1979. Participation in these commodity agreements is not designed so much to promote intraregional trade as to help stabilize the prices of the primary products produced in Asia that then enter into world trade.

China was one of the main trading partners of the Socialist bloc of countries in Europe, especially the Soviet Union, in the early days after the revolution. China has since sought trade contacts with other Asian countries. In many cases trade agreements between China and other Asian countries are in the form of barter agreements. China's most important trading partners in Asia have been countries such as Japan, Burma, Sri Lanka, India, Indonesia, Malaysia, Pakistan, and Singapore.

In 1958 a special effort was made to promote intraregional trade by the establishment, under the auspices of ECAFE, of intraregional trade talks. An important intraregional organization has been the Association of South East Asian Nations (ASEAN), which was established in 1967 to promote the economic and social development of the region. The ASEAN has fostered joint economic ventures among member states (Indonesia, Malaysia, the Philippines, Singapore, Thailand, and Brunei), and has worked to reduce trade barriers.

There has, however, been little effort at trade integration on a regional or subregional basis in Asia. In this respect Asia lags behind Latin America and, to some extent, Africa. Trade between India and Pakistan, which could be of great mutual benefit, is hampered by political relations between the two countries.

Many Asian countries are engaged in diversifying their internal production. Their goal is generally self-sufficiency rather than specialization and division of labour among a number of countries. As Asian countries increasingly become industrialized, however, and as they make greater use of their individual natural advantages, trade among the countries of the region could become more complementary.

External trade. The external trade of Asian countries has been considerably affected by the political ties that existed in colonial times, when important trade connections were established and when a certain trade pattern developed that proved convenient for both trading partners to continue. This convenience extended even to payments arrangements; the former British colonies were part of the sterling area, while the former French colonies found the franc a convenient medium of exchange. In addition, a system of Commonwealth preferences existed within the sterling area in Asia.

Several factors have tended to interfere with this pattern since the 1960s: the European Economic Community (EEC) was formed, with a new system of preferences; Japan became a major producer of both consumer and capital goods and a major market for the commodities exported from other Asian countries; and China began to take increasing interest in trade with countries outside the Socialist bloc.

Major export items

The main items exported from the developing countries of Asia are rubber, tea, crude petroleum and petroleum products, rice, sugar, copra, coconut oil and palm oil, cotton and cotton fabrics, jute and jute fabrics, tin-in-concentrates, tobacco, wood products, iron ore, wool, and hides and skins. Spices, which were such an important part of the trade with Western countries in earlier centuries, now form a very small part of the total exports of Asia.

Since the major exports of the developing countries of Asia are primary products, both their volume and price depend upon external demand, which fluctuates according to the level of industrial activity in developed countries. As a result there are great fluctuations both in the volume and price of exports. Further, there has been a tendency on the part of developed countries to replace some of these primary products with synthetic products, such as synthetic rubber and nylon and polyester fabrics. The developing countries are faced, therefore, with the problem that, even when they are able to increase their volume of exports, there is no corresponding increase in the value of their exports. In addition, the price of the consumer and capital goods exported by the advanced countries to the developing countries has been steadily rising since the end of World War II.

The main imports

The main imports of the developing countries have been machinery and transport equipment (including trucks, automobiles, and tractors); other manufactured goods; chemicals, including fertilizers; food, beverages, and tobacco, especially cigarettes; mineral fuels; and oils and fats. There

has been a decline in British exports to the sterling bloc countries of Asia in the postwar period. The major trading partners of France in Asia have been the countries of Indochina and the Middle East, while The Netherlands trades extensively with Indonesia. West Germany has developed considerable trade with Asia in the postwar period.

Since 1959 the Soviet Union has increased its trade with Asian countries, especially Burma, Sri Lanka, India, and Indonesia, which have been pursuing a neutralist foreign policy and have accepted military and economic aid from the Soviet Union as well.

The United States has been the biggest provider of economic aid to many Asian countries in the postwar period; there has also been a concurrent increase in trade between the United States and Asian countries. The dominant Asian trading partner of the United States is Japan; others include Saudi Arabia, Taiwan, South Korea, Hong Kong, Indonesia, Singapore, Malaysia, and the Philippines.

In its efforts to trade with countries outside the Socialist bloc, China has increased its trade with such countries as the United States, the United Kingdom, Australia, Canada, and France; it has also entered into a series of trade agreements with Japan. Trade between China and the United States came to a virtual standstill in 1950, when the United States imposed an embargo on hundreds of commodities; while the ban continues to apply to traffic in certain strategic goods, in 1971 it was lifted for other commodities. As a result, in the late 20th century the United States became one of the largest exporters to China. China has also been working to extend its trade relations with the newly independent countries in Africa.

During the 20th century petroleum has become an important part of the trade of West Asian countries of West Asia with the rest of the world. Iran and the Arab countries are among the chief beneficiaries of this trade. Originally, the Western powers had considerable economic superiority, which they used to negotiate terms that were less advantageous to the producing countries. The producing countries, in their turn, began to work closely together to protect their common interest; negotiations have resulted in much better terms for the producing countries, as shown in the strength of the Organization of Petroleum Exporting Countries (OPEC) formed in 1960.

TRANSPORTATION

Reference has already been made to the main transport systems that linked Asia and the Western world. Until the 19th century the land, or caravan, routes, supplemented by oceangoing vessels, were predominant. In the latter half of the 19th century there was a major shift to seagoing vessels. Rail transport has begun to play a progressively more important role, mainly in the internal movement of passengers within individual states and in the transport of heavier goods over longer distances. Concurrently, there has been considerable development of ports and harbours, linked to their hinterlands by rail and road. Air transport has proved to be not only the speediest but also often the cheapest means of transport, especially for costly items of relatively small weight and bulk. Air transport plays a particularly important role in landlocked countries—such as Afghanistan, Nepal, and Laos—and in the opening up of relatively inaccessible areas.

Within Asian countries animal transport remains the main means for the local transport of goods from one village to another or from the villages to a central marketplace. The animals, used for plowing during the cropping season, are also used for transport of goods at other times. The diesel truck, however, is rapidly replacing draft animals for internal traffic, and there is a simultaneous development of roads and highways in most countries.

Inland navigation is important in certain countries; a good river and canal system is capable of carrying goods and passengers at small cost over considerable distances. Among the countries having a well-developed inland water transport system are Bangladesh, Burma, Thailand, the former Indochina countries, and China. There are also great riverine ports such as Calcutta, Rangoon, Bangkok, and Ho Chi Minh City; oceangoing ships can navigate the Mekong River to inland ports such as Phnom Penh (Kampuchea) and can sail up the Yangtze River to Wu-han (China). Ultimately, it may be possible to connect even Laos with the sea by an extension of inland navigation facilities on the Mekong. The Yangtze, Sungari, and Hsi rivers of China provide a wide network of routes for motorized barges, supplementing traditional water transport.

Means of internal transport

For the movement of petroleum products there has been some development of pipelines, especially in West Asia and western and southwestern Soviet Asia. Pipelines have considerable advantages, such as economy and speed, but they also have the disadvantage of being subject to political vicissitudes when they cross international boundaries. Meanwhile, there has been considerable development of large oil tankers, which, through economies of scale, can compete with the most efficient pipelines and can also effect point-to-point delivery. (C.V.N.)

Administrative and social conditions

HISTORICAL DEVELOPMENT

The Pre-European era. The first sophisticated organization of space, people, and cultural system is customarily attributed to ancient Southwest Asia. In the earliest times there emerged the concept of a normal political state ruled by a god-king to whom all resources belonged and who held total power over the human population. Whereas some early states were city-states, the long-term trend was toward the spatial state that included rural hinterlands. States rose and fell in the ancient Orient as the conceptual system spread both westward into the Mediterranean Basin, as well as into the peninsula of Europe, and eastward into India and China. Historically, Indian political systems dominated South and Southeast Asia, Chinese systems controlled East Asia, and variations on the original model continued in Southwest Asia, while steadily changing variations developed new patterns in the Mediterranean Basin and the European peninsula.

Ancient state systems

At the end of the 15th century, at the time of Vasco da Gama's voyage to India (which signalled the dawn of European influence in Asia), the situation on the continent was approximately as follows: East Asian systems were relatively stable; the political situation in Southeast Asia was in a state of flux, as several states were in decay, and Islāmic political missionaries from Arabia had not yet succeeded in consolidating their influence; South Asia was similarly in a state of flux as the Mughal Empire struggled to achieve spatial hegemony over the Indian subcontinent; Southwest Asian political systems were experiencing a period of readjustment as the Turkic peoples penetrated the region and began to assume control; Central Asia was experiencing the last phase of the expansion of the Mongols, as well as the spread of Islām; no formal political states had so far evolved in the Siberian zone.

The evolution of European contact. The evolution of Europe's political relationship with Asia may be conveniently divided into two phases. The earlier phase that characterized the modern political geography of Asia was one in which some of the states of the European peninsula were able to introduce political controls into the unstable parts of the remainder of the Eurasian continent. The second, and more recent, phase has been characterized by the withdrawal of these controls from the whole of the southern zone, despite the fact that varying economic and political links remain operative. In the northern zone, however, where at an earlier stage no political states were in existence, the current phase is marked by the integration of the territory into the Soviet Union, a state centred on the eastern part of the European peninsula.

South and East Asia. The earliest European contacts with parts of southern and eastern Asia aimed at trade in the exotic products of the East; Europeans used their seapower to establish control over the trade routes. The European countries fought each other for the monopoly of the Eastern trade and established trading posts ("factories") at various points extending from Persia to South China and to the East Indies. The Portuguese, who were the first to arrive in India (1498), Malaya (1511), and southern China (1514), began to trade with these areas and established the first European trading posts there. The Dutch and the

Trade and politics

British followed not long after and the French and the Danes a little later. By 1700 the coasts of southern and southeastern Asia were dotted with European-controlled trade ports. Gradually these ports became points of territorial expansion, as Europeans increasingly intervened in the hinterlands so as to extend their control over production and trade. Some areas, such as Burma, Thailand, and Vietnam, were of little trading interest to the Europeans, and others, such as China, Korea, and Japan, declined to deal with the Europeans freely. By 1700 Portuguese power was in decline and the primary division of trading territories resulted from wars between the Dutch, English, and French, the Eastern conflicts often mirroring those in Europe itself. As a result, the Indies became a Dutch preserve, India became a British zone, and Spain held the Philippines. Only in the latter had the Christianizing of Asians as well as the acquisition of territory been an initial objective.

Political settlements in Europe after the conclusion of the Napoleonic Wars (1815) affected Southeast Asia. The Indies became totally Dutch, Britain acquired control over Malaya, and France gave up its claims to large territories in India in favour of Britain. The trading companies of The Netherlands and Great Britain ceased to operate in 1798 and 1858, respectively, and territorial political administration was assumed by the two governments. France returned to Southeast Asia somewhat later, taking over weak political states unable to resist encroachment, in order to establish power in Indochina. During the 19th century the United Kingdom took over Burma by stages, finally annexing it to India, and eventually assuming political control over a portion of western Borneo never effectively occupied by The Netherlands. The United States took over the Philippines from the Spanish in 1898. The German effort to gain control over the Chinese Shantung Peninsula was thwarted by other European powers, so that Germany had to be content with a long-term lease on the territory of the city of Tsingtao in southern Shantung, and a railroad-building concession. Germany did, however, become a political power in the Pacific islands in the late 19th century. During the last half of the 19th century, European countries pressured China into granting small holdings, called treaty ports, which were guaranteed by treaties, and these dotted the China coast and extended up the Yangtze Valley as far inland as Chungking in West China.

Thailand (then called Siam) remained free from political encroachment, and both Korea and Japan spurned all European effort to establish trade or to obtain political privilege. Late in the 19th century, however, Japan began its own political expansion, first in the form of establishing a sphere of influence, and then by conquest, taking over the Kuril Islands to the north and the Ryukyu Islands to the south, in 1875, and gained Korea and Taiwan in 1895. In 1905 Japan obtained southern Sakhalin and then gradually gained control of Manchuria, finally establishing the puppet state of Manchoukuo there in 1932. The expansion of the Turkish Ottoman Empire in Asia Minor left much of Arabia outside organized political control, and numerous small sheikhdoms continued to exist independently around oases in southern Arabia and along the Persian Gulf.

Japanese expansion

Central Asia. In AD 1400 the pastoral empire of the Golden Horde ruled all of Central Asia and much of eastern European Russia. During the 15th and 16th centuries Russian agricultural colonization spread eastward, and the exploitation of furs and forest timber products began; by the end of the 16th century, Russian explorers had crossed the Urals. In 1639 the first Russian explorer reached the Pacific Ocean at the Sea of Okhotsk, and in 1650 the Russians and the Chinese reached their first impasse over the control of trade and territory in the Amur River Valley. As in Siberia there were only small and fragmented ethnic territories lacking formal political organization, Russian political control was rapidly and easily achieved there. Expansion east of the Caspian Sea, however, proceeded more slowly, and it took until the end of the 19th century to bring the many Islâmic pastoral societies under control and to extend Russia's boundaries to the frontiers of Persia and Afghanistan in the south and to China in the Central Asian zone.

THE CONTEMPORARY PATTERN

The end of colonialism. During the late 19th and early 20th centuries the European powers were educating Asian peoples in methods of modern political administration and economic development. Political and cultural nationalisms gained in strength after 1900, focussing on traditional culture systems in the various regions. Political independence came to parts of southwestern Asia between 1920 and 1926, after the collapse of the Ottoman Empire, as a result of a deliberately planned separation of various ethnic groups. Conflicts and changes continued in Southwest Asia as the century progressed, heightened by the setting up of the State of Israel in 1948.

The Japanese wartime confrontation with and subsequent defeat of Western military power in Southeast Asia in 1942 gave an enormous psychological stimulus to nationalistic movements for political independence even during the Japanese occupation, and each colonial ruler was faced with demands for the end of colonial status as World War II came to an end with the defeat of the Japanese.

The United States in 1935 had promised the Philippines independence in 1945; this promise was fulfilled in 1946, after World War II, and this first release from externally administered status in the East was the forerunner of other moves. Political independence came to India and Pakistan in 1947; Ceylon (now Sri Lanka) and Burma in 1948; Indonesia in 1949; Cambodia (now Kampuchea) in 1954; nominally to both Vietnam and Laos in 1954; and to Malaya in 1957, with the Borneo colonies being added in 1963 to form the Federation of Malaysia. Singapore withdrew from Malaysia to become an independent state in 1965. After Indonesia annexed Portuguese Timor in 1976, British Hong Kong and Portuguese Macau remained the only European colonial holdings in Asia. The Soviet Union has largely avoided charges of imperialism in Central Asia and Siberia because territories there have been given the status of autonomous republics within the federal union.

Independence of colonial states

Problems of Asian nationalism. Three major conflicts occurred after World War II in which Western powers and their allies, Asian or other, were drawn into wars with Communist-supported Asian forces. The first of these was the French war in Indochina (1946 to 1954), which ended with the withdrawal of the French and the nominal independence of the territories of the region. The second was the Korean War (1950 to 1953), in which North Korean troops, later supported by the Communist Chinese, invaded South Korea, whose government was supported by United Nations forces; the result was a perpetuation of the division of the country into North and South by a cease-fire agreement. The third conflict was the war in Vietnam in which the United States supported the South Vietnamese government against Communist North Vietnam; fighting later spread to Laos and Cambodia. The government in South Vietnam collapsed in 1975, and on July 2, 1976, the two Vietnams were formally united as the Socialist Republic of Vietnam.

By the beginning of the 1970s Asian nationalism had recovered internal political and cultural control from the Western powers in all the larger countries of Asia, except in the Asian part of the Soviet Union. Within the Soviet Union, Russian strength has been sufficient to prevent outright rebellion, although the flight of some pastoral ethnic groups has continued back and forth across the boundary between the Soviet Union and China. Political independence has been much more completely achieved than has economic independence or social and economic advancement. Internal stability has also sometimes been threatened by regional or ethnic imbalances, as in East Pakistan (now Bangladesh). Japan has emerged as the most ethnically unified nation-state in Asia, and its great economic development from 1945 has resulted in relative internal stability. Elsewhere in Asia, political and cultural conditions have been in flux, as nationalistic groups, ethnic minorities, and political parties strive for the attainment of different objectives and as different nation-states attempt

Problems after independence

to pursue their national interests. Contrasts between traditional rural societies and modern urban industrial populations raise new kinds of problems. The establishment of the People's Republic of China in 1949 represented one of the major political developments of the century in Asia and is fraught with long-term consequences for South and Southeast Asia. In Southwest Asia the Arab–Israeli conflict threatens to create worldwide repercussions if the conflicting interests cannot be reconciled.

Continuing European linguistic influences. Europe has imprinted many permanent marks upon Asian cultures. Perhaps this is most noticeable in the continued use of European languages. Russian, English, and French are the dominant languages used, but German, Spanish, Dutch, and Portuguese are also employed in particular regions. Thus, German ethnic minorities in Soviet Central Asia continue to use German; older Filipinos speak Spanish; Indonesians use Dutch; and Portuguese is employed in Macau, Timor, and former Portuguese India. Vietnam, Kampuchea, and Laos continue to use French as a lingua franca, and French is taught in many schools. Dutch is the lingua franca of the educated class in Indonesia and is still taught in some schools. Pilipino is the official language of the Philippines, while English serves as a lingua franca and is taught throughout the schools. In Malaysia, India, and Hong Kong, English has an official status, is taught in schools, is spoken widely among the educated classes, and is the language of parliamentary debate. In parts of Southwest Asia, French and English are commonly spoken and are often taught in schools. Within Soviet Central Asia and Siberia, Russian is becoming more widely used and is the only common language. English, the most widely used non-Asian language, is a lingua franca throughout East, South, and Southwest Asia. (Jo.E.S.)

HISTORY

Prehistory

The traditional picture of Asia as the cradle of mankind provided the impetus for Western scholars to search this part of the world for the vestiges of man's biological and cultural beginnings. The romanticism of this notion has been tempered by scientific discoveries indicating that man had inhabited Europe and Africa as long as Asia, but Asia was nonetheless important in its broad spectrum of climatic zones, from Arctic to temperate and tropical, its mountain ranges and grassland steppes, its inland seas and Arctic tundra, plus other ecological settings to which prehistoric and modern man had to adapt both biologically and culturally. Since the formation of the present geological features of Asia during the period of mountain-building activity in the Middle Tertiary (about 35,000,000 BP [Before Present]) the continent gave rise to major taxa of primates that included the dryopithecine apes of the Miocene–Pliocene age (21,000,000 BP) and the earliest known hominid, *Ramapithecus punjabicus*. With the onset of the Pleistocene (about 2,000,000 BP), Asia was one of the habitats of the australopithecines, who, by middle Pleistocene times (500,000–200,000 BP), were replaced by *Homo erectus* forms. Evidence of Neanderthal man and later *Homo sapiens* (modern man) appeared during the upper Pleistocene, which began about 150,000 years ago. Thus, the major events of human evolution are represented in the paleontological and archaeological record of Asia.

THE FOSSIL RECORD OF PREHISTORIC MAN IN ASIA

Areas of human occupation. Human occupation in Asia from 100,000 years ago until about 35,000 years ago, a period that includes the geological-climatic events of the third interglacial and the initial glaciation of the Würm (Last) Glacial Period, was restricted to a considerable degree by the extension of ice sheets that covered portions of eastern Europe, the Tibetan plateau, and the diagonal mountain chains of Central Asia during periods of peak glaciation. These conditions blocked passage between Europe and eastern Asia save for narrow corridors not covered by ice. For this reason, high-altitude glaciated regions in western Asia and in the Himalayas are poor in prehistoric sites. The Bosporus remained dry land during much of the Pleistocene and formed the main avenue of communication between the Near East and the Black Sea. During periods of maximum glaciation, the Caspian Sea rose 250 to 300 feet (76 to 91 metres), and in this condition of flooding by waters fed by glacial melt it formed with the Volga River a spillway into the Black Sea. The Aral Sea enlarged in rhythm with the Caspian, and to the south and the east of the Ural Mountains a vast swamp marked the limits of glacial ice. With the onset of interglacials, of which the third is the one relevant to the period of time under consideration, these landlocked bodies of water became lowered while the Black Sea was elevated, since it was fed by oceans expanding in size from water liberated by melting marine glaciers. During these peak glaciation periods, portions of Asia were isolated from Europe.

Effects of glaciation

Thus, with the onset of the Würm glaciation some 70,000 years ago, climatic conditions effectively separated the Neanderthal populations of western Europe from those Neanderthals of western Asia whose occupation sites are recognized today in the Zagros Mountains of Iran and just north of the Elburz Mountains and Hindu Kush in Iran, Soviet Central Asia, and Afghanistan. With the onset of the second phase of glaciation (Würm II Stadial) 40,000 years ago and during a short period of milder climates in Europe and western Asia, the Neanderthal occupations were replaced by the sites of early modern hominids.

The Bering Strait formed an effective land bridge between northeastern Asia and the New World during the glacial epochs but ceased to serve as such with the elevation of sea levels during the Third Interglacial and again after the Würm glaciation. The former continental landmasses of the Sunda Shelf and Sahul Shelf in Southeast Asia were separated at the close of the Pleistocene as sea levels rose. These areas were dry during the Riss and Würm glaciations and provided a passageway for populations moving from Southeast Asia to Australia. The geographical boundary of Wallace's Line, running between Bali and Lombok, Borneo and the Celebes, Mindanao and Sangi, isolated the Australian fauna, including man, from Southeast Asia during the upper Pleistocene. South Asia was affected by glacial activity in the Himalayan region but in the south that area retained a tropical ecological setting throughout most of the Pleistocene. The upper Pleistocene of China is characterized by deposits of yellow earth, or loess, upon which man settled. The climate remained cold and dry for much of this period in China.

Remains of Neanderthal man. More than a dozen sites with fossils of Neanderthal man have been located in the Near East. Prehistoric research began there in 1864, when Louis Lartet (who later discovered Cro-Magnon man) rediscovered in Syria an ancient settlement observed some 30 years earlier. In 1878 an inventory of Stone Age artifact discoveries was compiled, but the first fossil find of early man in western Asia was made in 1900 at Grotte d'Antelias, in Lebanon. The remains of a seven-year-old child were recovered in 1938 in Lebanon at Ksar 'Akil, a rock-shelter near Beirut dated to 43,750 years BP. Discovery of an adult Neanderthal skull was made in 1925 at el-Zuttiyeh (Robbers' Cave) at Lake Tiberias in Israel—the "Galilee Skull" now datable to 70,000 years BP. In its vicinity is the cave of Har Qedumim (Jebel Qafzeh), where Neanderthal bones of comparable antiquity were recovered in 1933–35. The cave at Amud, also in this region, yielded in 1961 bones dated to the period of the Würm I Stadial of the last Ice Age (70,000–50,000 BP).

Neanderthaloids from western Asia

Near Jerusalem the Shuqbā (Shukbah) Cave contained human remains that had been deposited at the end of Neanderthal occupation of this part of Asia—*i.e.*, at about 35,000 years BP. The most critical discoveries in Israel are from two caves adjacent to one another in the Mt. Carmel range—Maghārat at-Tabūn, excavated in 1929–

Political divisions of Asia.

34, and Maghārat as-Skhūl, excavated in 1931–32. The age of occupation deposits at at-Tabūn ranges from 70,000 to 37,750 years BP. The occupation of as-Skhūl may have been slightly later, but both caves have yielded Neanderthal fossils with many physical features characteristic of later *Homo sapiens* hominids.

The original designation *Palaeoanthropus palestinus* assigned to the Mt. Carmel hominids was later dropped from use. The Turkish sites of Karain near Adala, excavated in 1949, and Musa Dağı have yielded teeth that may belong to the Würm I hominids, but the dating of these sites remains uncertain. Shanidar Cave in Iraq was excavated from 1953 to 1960. Its deposits range in age from 60,000 to 44,950 years BP and include the bones of a child and several adults. The two sites of Würm I date in Iran are the Kermanshah Cave, near Bisitun, and the Tamtama Cave, near Orūmīyeh (Urmia). Excavation of these fossil-bearing deposits began in 1949.

Fossils of Neanderthals were found in the Soviet Union in 1924 at the cave of Kiik-Koba, in the Crimea. Two skeletons, both missing skulls, appear to have been purposeful burials. The deposit dated to the early part of Würm I, while at another Crimean site, called Staroselye, excavated in 1952, the fossil remains date to 35,000 years BP. Teshik-Tash cave, in Uzbekistan, yielded in 1938–39 a child burial of the Würm I–II Interstadial, but the skeletal remains unquestionably belonged to a Neanderthal population. Less certain is the dating of an alluvial deposit containing some prehistoric human remains found in 1925 along the lower Volga River at Undory. Aman-Kutan Cave, in Samarkand, has the earliest dated fossil Neanderthal in Central Asia.

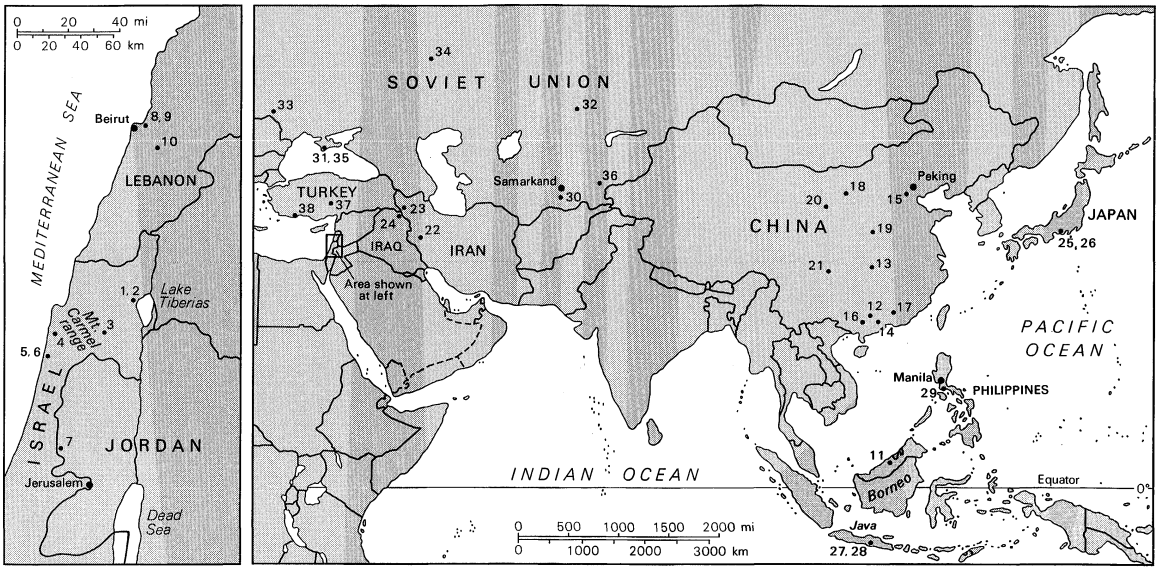
In eastern and Southeast Asia, hominid fossils resembling the western Asiatic Neanderthal hominids have been recovered from sites that date from the early part of the upper Pleistocene and even into the beginning of post-Pleistocene, or Recent, times (10,000 BP). The taxonomic status of these hominids is uncertain, but they have been given the colloquial appellation of Neanderthaloids, even though some populations of this hominid group were contemporaries of *Homo sapiens sapiens,* who emerged in western Asia, Europe, and Africa by the beginning of the Würm II Stadial of the last Ice Age, about 40,000 years BP.

Neanderthaloids from Southeast Asia and eastern Asia

Since the excavation in 1922 at Sjara-osso-gol River in Ordos, China has provided numerous loci with human remains. This site may have been inhabited contemporaneously with the occupation of the other Ordos site, Ti-shao-kou-wan, which was investigated in 1957. Both appear to be of upper Pleistocene date (*c.* 150,000–10,000 BP).

Earlier than these is the Ting-ts'un site, in Shansi, excavated in 1954 and dated to the Third Interglacial (100,000–70,000 BP). Still more ancient is the human skull from the cave at Ma-pa, in Kwangtung, which may be late middle Pleistocene; *i.e., c.* 125,000 years BP. Other series of Neanderthaloids appear in the fossil record of Java at the upper Pleistocene site of Ngandong, on the Solo River.

At the time of the excavation of 11 skulls from Ngandong in 1931–32, the specimens were named *Homo (Javanthropus) soloensis,* but modern systematists regard these fossils as similar to the Chinese Neanderthaloids and to the Neanderthals of western Asia. Two skulls from a limestone terrace overhanging an ancient lake near the village of Wadjak, in Java, were discovered in 1890, and these constitute the earliest discovered fossil hominids from Asia, although they were not reported until 1921. They have been dated to the very end of the Pleistocene or possibly to the beginning of post-Pleistocene times (*c.* 10,000 BP), but



ISRAEL
1. Amud
2. el-Zuttiyeh
3. Har Qedumim
4. Mt. Carmel (Maghārat at-Tabūn Maghārat as-Skhūl)
5. Mugharet el-Kabareh
6. Wādī Maghārah

JORDAN
7. Shuqbā

LEBANON
8. Abri Bergy
9. Grotte d'Antelias
10. Ksar 'Akil

BORNEO
11. Niah

CHINA
12. Chai-sha
13. Ch'ang-yang
14. Ch'i-lin Shan
15. Chou-k'ou-tien
16. Kai-t'o-tung
17. Ma-pa
18. Sjara-osso-gol
19. Ting-ts'un
20. Ti-shao-kou-wan
21. Tzu-yang

IRAN
22. Kermanshah
23. Tamtama

IRAQ
24. Shanidar

JAPAN
25. Aichi-ken
26. Ushikawa

JAVA
27. Ngandong
28. Wadjak

PHILIPPINES
29. Rio Pasig

SOVIET UNION
30. Aman-Kutan
31. Kiik-Koba
32. Malta
33. Molodova
34. Oundory
35. Staroselye
36. Teshik-Tash

TURKEY
37. Karain
38. Musa Dağı

Major sites of Upper Pleistocene hominid remains in Asia.

they resemble the Ngandong Neanderthaloid specimens in a number of striking ways. Fossil evidence of Neanderthal man has not been reported from southern Asia, although this region contains artifactual materials made by hominids of the period of time under consideration.

A few bone fragments from a late middle Pleistocene deposit (*c.* 125,000 BP) at Ushikawa, on the Japanese island of Honshū, were reported in 1957, but the bones are too incomplete to make possible a reliable identification of their taxonomic status.

Homo sapiens remains. In western Asia the first appearance of modern *Homo sapiens* coincides with the onset of the Würm II Stadial of the last Ice Age, about 40,000 years BP, approximately the same period of time that *sapiens* hominids replace Neanderthals in Europe and Africa. As noted above, Neanderthaloid hominids persisted in eastern and southeastern Asia until the close of the Pleistocene, but they appear to have constituted isolated populations in this part of the world, where modern-type *sapiens* had also appeared by late upper Pleistocene times (*c.* 30,000 BP).

The earliest date for *Homo sapiens* in Southeast Asia is 37,650 years BP, assigned to a skull specimen discovered in 1959 at Niah Cave, in Borneo. Late upper Pleistocene is also the date for the Chinese fossils found in 1930 at the Upper Cave of Chou-k'ou-tien, near Peking, in 1951 at Tzuyang (formerly Tzeyang), in Szechwan, in 1956 at Kai-t'o-tung in Laipin, Kwangsi, and at Ch'i-lin Shan, in the same province. Fossils were also found in 1958 at Chaisha (formerly Liukiang), also in Kwangsi; this site may have lower deposits dating to a period contemporary with the Würm I Stadial (70,000–50,000 BP), as is certainly the case at the site of Ch'ang-yang (Lungtung), in Hupeh Province, investigated in 1957.

Japan has yielded a *sapiens* specimen in the fossil record taken from Aichiken, in Honshū, in 1958, but the date can only be specified as upper Pleistocene. In 1921 a skull with Negrito or Pygmy features was reported from an alluvial deposit of the Río Pasig near Manila, but the dating of this specimen from the Philippines is uncertain.

Southeast Asia's record of *sapiens* fossils does not commence until the beginning of the geological Recent period (*i.e.*, after 10,000 BP), save for a single hominid lower molar found in northern Indochina in a deposit of uncertain antiquity. Similarly, southern Asia does not provide a *sapiens* fossil series for this period; the earliest skeletal remains, coming from Sai-Nahar-Rai near Allahabad in north-central India, date to 10,345 to 10,311 years BP.

Near Eastern *sapiens* fossils

The Late Stone Age (Mesolithic) site of Bellanbandi Palassa in Sri Lanka (Ceylon) has yielded a dozen skeletal specimens of *H. sapiens* from a deposit dated about 5000 years BC.

In the Near East, Lebanon offers two Pleistocene sites with *sapiens* fossils: a cave at Abri Bergy, near Antilas, which was excavated in 1948, and the previously mentioned site of Ksar 'Akil, which has a middle Würm deposit (*c.* 40,000 BP) superimposing the portion of the shelter from which the Neanderthal specimens had been removed in 1938. Similar hominids came from an upper level of the Har Qedumim cave, in Israel, as well as from the Mugharet el-Kebareh and Wādī Maghārah (Mugharet el-Wad) caves of the Mt. Carmel Range, where excavations in 1931 provided data on late upper Pleistocene occupations. At Malta, in Siberia, a late Würm (*c.* 15,000 BC) *sapiens* specimen was recovered in 1929, but other sites in the Soviet Union have provided few skeletal remains from this period.

MORPHOLOGY OF ASIAN FOSSIL REMAINS

Features of the skull

Neanderthal morphology in western Asia. The classification of upper Pleistocene Asian hominids as Neanderthal, Neanderthaloid, and sapient forms becomes meaningful when the physical anthropologist compares the osteological and dental anatomy of the fossils from the sites just discussed. Some Neanderthal specimens of western Asia closely resemble the Würm I Neanderthals of Europe with regard to the possession of a massive supraorbital torus (browridges), a low-lying forehead, large face with large eye sockets (orbits), large nose, dolichocranic or mesocranic cranial form, chinned or chinless mandible, bulging of the occipital bone (at the back of the skull) into a "chignon," taurodont molars, a cranial capacity with a mean of about 95 cubic inches (1,550 cubic centimetres), and a range from 78 to 104 cubic inches (1,270 to 1,700 cubic centi-

metres), inclusive, of both sexes, a modern-like dentition (apart from taurodonty), and a body conformation that was burly and muscularly well developed. The misconception that Neanderthals walked with slightly flexed knees and a hunched posture was based upon the examination of a Neanderthal skeleton from La Chapelle-aux-Saints, in France, an arthritic and aged male. The retroversion of the head of the tibia (large bone of the lower leg) was also considered to be indicative of a slumped posture and plodding gait.

Recent reinvestigation of the La Chapelle specimen and other Neanderthal skeletons does not indicate that this hominid's locomotor pattern was significantly different from that of later and modern hominids. Stature ranged from about 61 inches (154 centimetres) for females to 68 inches (173 centimetres) for males.

The Asiatic specimens that show the most striking physical resemblances to European specimens of Neanderthal man (such as those from La Chapelle-aux-Saints and La Ferrassie, in France, Spy, in Belgium, and the Neander Valley specimen, in West Germany) are from Amud and Shuqbā, in Israel, as well as some of the skeletons from the series of the Mt. Carmel caves, from Shanidar, in Iraq, and from Kiik-Koba, Aman-Kutan, and Teshik-Tash in the Soviet Union.

Other Neanderthal fossils of western Asia, however, bear closer phenotypic similarity to the Third Interglacial Neanderthals of Europe as represented in the sites of Ehringsdorf, East Germany, and Steinheim, in West Germany, and Fontéchevade, in France, where the specimens exhibit less morphological specialization than do the Würm I inhabitants of Europe. These Third Interglacial Neanderthals are more like modern hominids in their possession of lower mean cranial capacity, with a range of 67–89 cubic inches (1,100–1,450 cubic centimetres) for both sexes, the absence of an occipital "chignon," less muscular marking of the cranial vault, longer faces, and a tendency toward linearity of body build. The postcranial long bones are not bowed, as is the case with many Neanderthal specimens of Europe and western Asia of Würm I times. With these more modern-looking Neanderthals may be associated the western Asiatic fossils from Ksar 'Akil, in Lebanon, Har Qedumim, el-Zuttiyeh, and the majority of specimens from the Mt. Carmel series in Israel and those from the Soviet site of Staroselye. The specimens from Maghār-at at-Tabūn and Maghārat as-Skhūl have been studied in greatest detail and are recognized by most physical anthropologists as representative Neanderthal populations evolving into the sapient type of humanity that has dominated this part of the world for the past 40,000 years. For the other sites of western Asia noted earlier, either the fossil evidence is too fragmentary or the comparative anatomical studies are not sufficiently complete to permit classification beyond the statement that a Neanderthal phenotypic pattern seems to be represented.

Neanderthal morphology in eastern Asia. For the Neanderthaloids of eastern Asia, the fossil record is less complete. Incisor teeth from Sjara-osso-gol and Ting-ts'un are "shovelled," a feature found in high frequency among living peoples of modern Asia and not unknown among *Homo erectus* specimens at the middle Pleistocene site of Chou-k'ou-tien. Cranial bones from Ti-shao-kou-wan and Ma-pa are thick, the skull from the southern China locality having a capacity of about 75 cubic inches (1,225 cubic centimetres) as well as robust features similar to the skull vaults from Ngandong. Yet the Javanese skulls, which are 11 in number and the only case of a population series for this period of time, have a mean cranial capacity well under this value, the means for males and females being 71 cubic inches (1,158 cubic centimetres) and 64 cubic inches (1,042 cubic centimetres), respectively. The range of cranial capacity for the series is 63–77 cubic inches (1,035–1,255 cubic centimetres). In this feature the Chinese and Javanese Neanderthaloids do not resemble the Neanderthaloids of Broken Hill and Saldanha Bay, in Africa, but are more like the middle Pleistocene *Homo erectus* specimens from Asia, from which phylogenetic line they most likely evolved. Similarities with *Homo erectus* are also obvious in the pronounced angularity of the occipital bone (at the back of the skull), platyrrhinic nasal indices (indicating wide nasal openings in the skull), the thickness of the parietal bones (at the sides of the skull), and the location of the maximum width of the skull at points just below the parietal area. Yet the frontal torus (browridge) is of the divided pattern, as it is in Neanderthals, rather than forming a continuous bar. It is with the Wadjak specimens that cranial capacities increase to 94 cubic inches (1,550 cubic centimetres) for the female and 100 cubic inches (1,650 cubic centimetres) for the male. Specific features of large and chinned mandibles, broad and flat faces with broad noses and massive browridges, alveolar prognathism (forward projection of the tooth-bearing portion of the upper and lower jaws), and deep palates have led physical anthropologists to note the close similarity of the Wadjak specimens to modern aboriginal Australian and Melanesian populations rather than to modern Asians. Even less is known of the postcranial anatomy of these Neanderthaloids, but stature estimates based on two modern-looking tibial bones (from the lower leg) directly associated with the Ngandong crania suggest a height of about 70 inches (178 centimetres) for the male. The stature of the Ti-shao-kou-wan specimen is estimated for a thick-walled femoral (thigh bone) fragment as 66 inches (167 centimetres) if the specimen was male and 63 inches (160 centimetres) if female. The humerus (upper arm bone) from Ushikawa is thought to belong to a female less than 55 inches (140 centimetres) in height, and it differs from the humeri of modern Japanese by its narrowing of the proximal (shoulder) end and the pronounced thickening of the cortical area of the bone. The short stature and stocky body build of the greater number of Neanderthals and Neanderthaloids is considered by some anthropologists to have been an adaptive mechanism for surviving cold climatic conditions, where increased body bulk helps conserve body heat. Certain features of the Neanderthal-Neanderthaloid face have been explained as thermal adaptations to cold stress. The taller and more linear body conformation of the Solo specimens is especially interesting in view of the fact that this particular population lived in the tropical belt, where cold adaptation was not a factor of survival.

Skull morphology

Body build of Asian Neanderthals

The Asian sapiens fossils. The *Homo sapiens* fossil specimens are not unlike contemporary Asiatic populations, with the exception of the skull from Niah Cave, which resembles more closely the skulls of Australians than Borneans in structural features. Three of the seven specimens from the Upper Cave site at Chouk'outien have been compared to contemporary populations of Ainus, Melanesians, and Eskimos, and a broad range of phenotypic and polymorphic features may have been common in this part of Asia during later Pleistocene times (30,000–10,000 BP). Very similar to the skeletal anatomy of present occupants of eastern Asia are the specimens from Tzu-yang, Kai-t'o-tung, Ch'ang-yang, and Chai-sha, Kwangsi, save that the male postcranial skeleton from the latter site suggests a low height of 150 centimetres, which is just on the upper border of Pygmy stature. The skull fragments from Aichi-ken in Japan are sapient, as are the other fossil remains from upper Pleistocene sites in western Asia noted above.

LIFE-STYLES OF PREHISTORIC MAN IN ASIA

Stone-tool industries. The life of Asiatic hominids of this period can be reconstructed in large part from archaeological data of preserved stone tools and weapons. Two technological traditions are represented in deposits dated to the middle Pleistocene (*c.* 125,000 BP)—the bifacial-hand-ax and cleaver-tool industries of western and central Asia and the chopper–chopping-tool traditions of eastern and southeastern Asia. The transitional zone of these two traditions occurs in India, where it has been defined as the "Movius Line," after H.L. Movius, an American prehistorian who has conducted research in Asia. The bifacial-hand-ax and cleaver-tool tradition is called Acheulian in Europe, Africa, and western Asia, where it appears stratigraphically earlier than the flake-tool industry called Levalloisian, although flakes and bifaces occur together in some sites in this area.

Acheulian and Levalloisian tool traditions

Acheulian hand axes have been found in deposits in the caves of Mt. Carmel, in Israel, in Jordan and Syria, Turkey, Arabia, Iran, Iraq, Afghanistan, and eastward to Armenia, the Crimea, and the Caucasus Mountains. The Narmada Valley of India is particularly rich in hand axes, the first discovery of a paleolithic tool in South Asia having been made in the Madras area of India as early as 1863. Levallois flakes made by a prepared core technique occur in increasing frequency with the dawn of the upper Pleistocene in India, where they are a final phase (dating from *c.* 150,000 years ago) of another flake-tool tradition, called Soan. Apart from a few isolated cases of hand axes in China, the only place in Southeast Asia where these tools appear is in Java, where they are associated with the Patjitanian tool industry, which also has certain Levalloisian features. Chopper–chopping tools also appear in Java during this period, as do tool industries called the Sangiran flake industry and the Ngandongian industry. These are marked by the presence of antler picks, stingray barbs, and bola stones as well as small stone flake tools.

With the advent of colder climatic conditions and the onset of the Würm I Stadial (70,000 BP) in western Asia, the bifacial-hand-ax industries were replaced by flake-tool traditions; one of these was an evolved Levalloisian and another the Mousterian tool industry, in which stone flakes were retouched to make tools but not specifically manufactured from prepared stone cores. The Mousterian tradition is intimately associated with Neanderthal man in western Asia, as is seen at Ksar 'Akil, the caves of Mt. Carmel, a number of sites in Jordan and Syria, Kiik-Koba in the Crimea, Shanidar in Iraq, Teshik-Tash in central Asia, Molodova in the western Ukraine, and far eastward in the Ordos region near the Great Wall of China. In southern Asia, the later phase of the Soan tool tradition resembles the Mousterian. Elsewhere in eastern and southern Asia, the chopper–chopping-tool tradition persisted until the close of the Pleistocene (*c.* 10,000 BP). In Burma, in the vicinity of the Irrawaddy River, a silicified wood or tuff was manufactured into choppers and retouched flakes. These tools represent the Anyathian industry. The Tampanian industry of Malaysia is another local variation of the chopper–chopping-tool tradition. On the shores of the Sea of Azov in the southern U.S.S.R. this tradition merged with the limits of diffusion of the Mousterian tradition. At the upper Pleistocene localities of Chou-k'ou-tien, quartz flakes with working at both ends occurred with choppers, chopping tools, and small scrapers. A tool tradition of flaked stone balls and bifacial choppers but no bipolar flaking is called the Fenho industry, and its Third Interglacial date marks the transition to the upper Pleistocene traditions that have features of the Asiatic Mousterian in combination with older styles. In eastern and southeastern Asia many elements of this culture persist into terminal Pleistocene times, when Mesolithic cultures, characterized by small-blade microliths, replaced the older traditions.

Mousterian tool industries

Fire, shelter, and cultural data. Human occupation during the upper Pleistocene included the areas inhabited by *Homo erectus* of the middle Pleistocene as well as extensions into areas that earlier hominids had not entered. Biological and cultural adaptation to cold climatic conditions permitted settlement of higher altitude regions of Asia, and with the retreat of the ice sheets man followed game fauna into the newly opened country. Cold climates were made more tolerable by the use of caves and rock-shelters, some of which served as places of burial as well as hearths and industry sites. Open-air encampments continued to be used in some regions, however. Fire, which has been reported in Asia as early as 360,000 years ago from evidence of charcoal and charred bone at the *Homo erectus* deposits of Chou-k'ou-tien, in China, was controlled by the hominids of the upper Pleistocene, thus enabling them to move into wider areas of settlement and survive under a wide range of ecological settings. The control of fire by man, along with cave dwelling and the wearing of skins sewn together with bone awls, enabled prehistoric man to survive the cold conditions of the later upper Pleistocene. Traces of fire are found in most of the open-air, cave, and rock-shelter sites where skeletal and artifactual remains have been preserved, but the traces of charcoal and calcined bones do not appear in disturbed deposits, as at Ngandong. Among the undisturbed sites are the caves at Mt. Carmel, Aman-Kutan, Molodova, and Ch'i-lin Shan. Neanderthal man practiced burial of the dead at the Mt. Carmel caves, Kiik-Koba, Teshik-Tash, in the Upper Cave of Chou-k'ou-tien, Ksar 'Akil and Malta. At Molodova, the remains of a hut can be identified by the oval of mammoth bones and tusks enclosing some 15 separate hearths. This is evidence of open-air housing during a warm phase of the upper Pleistocene, although caves and rock-shelters continued to be the most common form of habitation in areas where they were naturally present. Evidence of cannibalism occurs at Solo, as it does in several Neanderthal sites of Europe. The presence of diseased and aged individuals in the Neanderthal community at Shanidar speaks of a more humane aspect of upper Pleistocene life. Pictorial or plastic art does not appear in the archaeological record for Neanderthal man and emerges only with the presence of *Homo sapiens.*

Fire and shelter

Throughout a broad geographical range of habitation in the Asiatic landmass, upper Pleistocene hominids were successful hunters of large game animals. For the western and central Asiatic Neanderthals these were Paleoarctic fauna, which can be identified as cold adapted or warm adapted according to their occurrence in periods of high glaciation or interglacial–interstadial recessions of the ice. In Israel, the warm-climate fauna are marked by remains of hippopotamus and rhinoceros, cold-climate fauna by deer and antelope. In central Asia the cold-adapted woolly rhinoceros, mammoth, and reindeer were abundant. South and Southeast Asian fauna were distributed also in southern China and merged with the Paleoarctic fauna of northern China. In India, some of the fauna of the middle Pleistocene survived until well into the late upper Pleistocene, as suggested by the presence of certain species of bovids. The persistence of tropical conditions throughout much of the Pleistocene in southern and Southeast Asia gives a very different faunal picture from that of western and Central Asia.

Game animals

PHYLOGENETIC AFFINITIES OF ASIAN FOSSILS TO MODERN MAN

Physical anthropologists are cautious in assuming phylogenetic affinities of living human populations to specific hominid fossil specimens of the Pleistocene, although some scholars have asserted that the living races of man can be recognized in the ancestral hominid record as far back as the middle Pleistocene. Beyond a few thousand years, however, physical resemblances of particular osteological or dental features between living and extinct populations become fewer in number and more tenuous as reliable data for establishing phylogenetic lines and classifications, and it is no longer a common practice in physical anthropology to attempt a "racial phylogeny" for a population beyond the limits of a few millennia. The traditional racial categories of Veddoid, Caucasoid, Mongoloid, Australoid, and the like are no longer regarded as particularly meaningful in relation to knowledge of prehistoric man.

In western Asia, the Natufian people of the Mesolithic of Israel do not resemble the present populations of the Near East; hence, it is not surprising that still earlier hominids, such as the Neanderthals of western Europe, fail to reflect obvious connections with contemporary Israelis, Lebanese, Syrians, etc. In Southeast Asia, the Neanderthaloid specimens from the sites of Solo and Wadjak in Java and Niah Cave in Borneo do not resemble the present inhabitants of mainland and island Southeast Asia, although some features of Neanderthaloid crania appear in the contemporary native populations of Australia and Melanesia. With the coming of Mesolithic (Middle Stone Age) post-Pleistocene cultures (after 12,000 BP) in Southeast Asia, the populations bearing these new traditions are represented in the osteological record by only a few specimens that are similar in anatomically significant ways to modern Southeast Asians. The earliest known fossil specimen of a Pygmy or Negrito population in Asia has come from Indochina, but it is only as ancient as the Neolithic Period of culture (*c.* 8000 BP). The antiquity of Pygmy populations in Malaysia, the Philippines, Anda-

Affinities of pygmy peoples

man Islands, and New Guinea remains unknown, but it no longer seems reasonable to conceive of a "Pygmy race" binding these dwarfed Asiatic groups to Pygmies of Africa. Pygmy populations of Asia bear many more resemblances of physical characters to the macropopulations surrounding them than to more distant Pygmy populations.

Skeletons that can be reasonably identified as Chinese first appear in the Far East around 3000 BC, although some phenotypic variables, such as shovel-shaped incisors, which appear in high frequency in populations of Asiatic descent, occur as well in *Homo erectus* fossils from Chouk'ou-tien. The high frequency of Chinese characters in the indigenous populations of Southeast Asia is explained by historic movements of Chinese into regions to the south of their ancient cultural area. In southern Asia, the hominid skeletal record has been recognized as useful in drawing some phylogenetic affinities between living Indians and their ancestors of a few thousand years ago, but the evidence for Pleistocene man in India was established upon the discovery of stone industries rather than upon a rich fossil record. It is with the food-producing Neolithic (New Stone Age) cultures of Asia that biological similarities between extinct and contemporary human populations become apparent, but this represents an antiquity of only four or five millennia. It was from Asia that man moved last to the uninhabited vastness of Oceania, Australia, and the Western Hemisphere. Movements of Asian people into Europe has continued to the present day in eastern Europe and the Aegean. (K.A.R.K.)

The ancient world

The history of Asia in the past 2,500 years is primarily the result of the interaction of five main influences: (1) Chinese, (2) Indian, (3) Islāmic, (4) European, including Russian, (5) Central Asian. Of these, the first four represent different kinds of civilization. The fifth has little originality but has been of significance in affecting the distribution of peoples and of political power.

China has molded the civilization of eastern Asia including Japan, Korea, and Annam and has been a primary influence on Mongolia, Tibet, Thailand, Kampuchea, and Burma. Wherever Chinese influence exerted itself, it introduced Confucianism, a distinctive style of art and, above all, the Chinese script.

Hindu and Buddhist influences

Indian influence has mainly expressed itself through Hinduism and Buddhism. These are not merely religious in nature but have carried with them Indian art and literature and often an Indian alphabet, as in Tibet, Java, and Kampuchea. Indian influences have affected the peoples of Central Asia and Southeast Asia, including the Malay Archipelago (the East Indies) and Indochina. Buddhism spread into China and Japan, but Indian culture on the whole has itself been little affected by Chinese art, literature, or ethics.

Islām spread widely in all directions from its original home in Arabia. It subjugated Southwest Asia, in which it is still the principal religion, and also eastern and northern Africa. It spread for a time into eastern Europe and Spain. In the other direction it got a firm hold in the northwest and northeast of the Indian subcontinent, and the states of Pakistan and Bangladesh now cover these two areas. Beyond India it reached the Malay Archipelago (the East Indies), where it submerged earlier Hindu influences. Through Central Asia it reached and affected China, but it gained no foothold in Japan and Indochina. Islām, like Buddhism, took with it everywhere a special style of art and culture. It was usually accompanied by the use of the Arabic alphabet, and the vocabulary of this language forms a large part of the languages of the Muslim peoples.

Central Asia has been the region into and through which these various powerful influences have been projected. Archaeological excavations have shown that early in the Christian Era there flourished in the Tarim basin small states, such as Khotan and Kucha, which possessed a mixed culture comprising Chinese, Indian, Iranian, and even Greek elements. Buddhist, Christian, and Manichaean edifices have been found as well as libraries in many languages, two of which were previously unknown. Through Central Asia, Greek influences and later Islām penetrated India, and Buddhism passed from India into the Far East and into parts of Southeast Asia.

The various tribes or peoples of the area have no common name. Linguistically they fall into several groups, such as Turks, Mongols, and Huns; and, in the face of difficult communications and terrain, only occasionally did several of the groups come together in one state. In history, Central Asia has acted rather like a sponge, absorbing pressures and sooner or later transmitting them. Time and again these pressures have produced invasions of the peripheral regions, and in the Christian Era the following may be mentioned: the early invasion of Europe by the Avars, Huns, and Bulgarians; the conquest of northern India by the White Huns; the conquest of Russia by the Mongols; the conquests of Timur; the conquest of Asia Minor and eastern Europe by the Turks; the invasion of India by the Mughals; the conquest of China first by the Mongols under Kublai Khan and later in the 17th century by the Manchus.

THE MIDDLE EAST

The ancient Middle East, or Near East, is a historical concept denoting the extent, in space and time, of the earliest civilized societies. Through archaeology, it is known that in this area the change from food gathering to food production first began and that the diffusion of agriculture, not only in the knowledge of grains but also in the technique of harvesting, took place. Toward the 3rd millennium BC the emergence of river-valley civilization in Mesopotamia and Egypt set this area apart from the peasant cultures of the rest of Asia and Europe. The cradle of Mesopotamian civilization was the southernmost part of the Tigris–Euphrates valley, and there the first cities arose. As is known from the remains of their pottery, the earliest settlers of this marshy plain had descended from the highlands of southwestern Persia. These people were probably Sumerians, speaking a very remarkable language, which has not been brought into relation with any known tongue. Physically they belonged to the Mediterranean group of peoples. The high civilization of the Sumerian south penetrated the Semitic-speaking peoples of the middle regions of the valleys, whence arose the Akkadian dynasty under the ruler Sargon. The Akkadians finally absorbed the Sumerians, and out of this mixture emerged the state of Babylon under Hammurabi, the lawgiver, in the 18th century BC. But the epics and books of wisdom of the Sumerians remained the classical texts of both Babylonians and, later, Assyrians. The Sumerian civilization had invented writing, at first as a practical requirement of the organization of their temples. They used pictograms, later supplementing them with phonetic signs. The form of the society that built the earliest cities has been called theocratic socialism. Its basis was a well-balanced mixed economy in which agriculture, stock breeding, and hunting existed side by side. Through the export of rugs and textiles, weapons and jewelry, the influence of Sumerian civilization permeated the whole of the ancient Middle East.

Tigris–Euphrates valley

The Babylonian state collapsed before invaders, the Kassites, from Elam, who controlled Babylon for five centuries. They adopted the civilization and Semitic language of their subjects. The Hittites, who first invaded Babylonia about 1595 BC, created a considerable empire covering northern Syria and the greater part of Asia Minor in the 14th century BC. In the archives found at their capital, eight languages are represented, including Sumerian and Akkadian. Subsequently the Assyrians, who seem to have been an offshoot of the Babylonians using almost the same language, asserted themselves and in the 11th century BC became the chief power. Their empire gradually broke up, finally succumbing before the Medo-Persian power at the close of the 7th century BC. Babylon itself was taken by the conqueror Cyrus the Great of Persia in 539 BC, but its culture and religion continued to exercise great influence long after the Persian conquest.

In Egypt the cultural continuity was even stronger than in Mesopotamia, and there was never any change corresponding with the displacement of Sumerian by Akkadian in Mesopotamia. Whereas that land was dotted with au-

Egyptian civilization

tonomous cities, Egypt began in the upper reaches of the Nile Valley as a royal domain, which extended to cover the whole valley and to found a 1st dynasty about 3100 BC. The administration and major activities of society were centralized to an extreme degree; and it was accepted that in the person of pharaoh, the living king, a god had taken charge of the people. With relatively minor breaks the established order continued for many centuries, Egyptian influence at times reaching eastward to the upper Euphrates. Up to about 1200 BC the history of the ancient Middle East had passed through two main phases, the emergence of the first great civilization in Mesopotamia and Egypt and subsequently the gradual spread of that civilization to the periphery. About 1200 BC new waves of invaders broke into Asia Minor and the Levant, destroying the Hittite Empire and disrupting Egypt, and after this the creative power of the Middle East waned. Its main achievement was in the consolidation of acquired knowledge. From this period the centres of power move both to the west and to the east of Egypt and Mesopotamia.

As has been seen, Persian power in the 6th century BC destroyed the Assyrian Empire and in 539 BC captured Babylon and created the Achaemenid Empire. The Persians, with whom the Medes are often coupled, appear to be Aryan in origin, their language and religion offering remarkable analogies to those of the early Hindus in India. These two peoples appear to have had a common origin in central Asia. The Achaemenid power at its greatest extended from the Oxus and Indus in the east to Thrace in the west and Egypt in the south, but it fell before Greece after lasting for more than 200 years. Darius and Xerxes were repulsed in their efforts to subjugate the Greek peninsula; and Alexander the Great of Macedonia conquered their successor, Darius III, in 331 BC. But the greater part of the empire continued to exist under new masters, the Seleucids, as a Hellenistic power that was of great importance for the dissemination of Greek culture in the east. About the same period (227 BC–AD 226) the Parthian Empire arose under the Arsacids in Khurasan. The Parthians were probably a Turanian tribe who had adopted Persian customs. At one time their power stretched from India to Syria. They withstood the Romans but succumbed to the Persian dynasty of Sassanids, who ruled for about four centuries, establishing the Zoroastrian faith as their state religion and maintaining an equal conflict with the eastern Roman Empire. But in the 7th century AD their power was overwhelmed in the first rush of the Muslim conquest that established Islām in Persia and in the neighbouring lands.

INDIA AND INDIANIZED ASIA

The subcontinent of India is divided from the rest of Asia by the Himalayan mountain ranges. This has by no means kept India in isolation, but it has resulted in the growth over a period of two to three millenniums of a Hindu civilization that in many of its aspects is unique. Archaeologists have unearthed the remains of a city civilization on the upper Indus valley that appears to have flourished in the 3rd millennium and to have had affinities with Sumerian civilization.

'he
.ryans

Hindu civilization came about in the 1st millennium BC as a result of the intermingling of Aryan and pre-Aryan cultures. Entering India sometime between 1800 and 1500 BC, the nomadic Aryans settled in the northwest, thence gradually during the 1st millennium BC encroaching eastward on the pre-Aryan peoples in the Ganges valley and producing a multiplicity of settled states and a fusion of cultures. Society came to be dominated by a hereditary priestly class of Brahmans. Political multiplicity is a characteristic of India's history, and the occasions on which most of the subcontinent was united under one rule were few and relatively brief. Such periods occurred during the Mauryan rule of Aśoka in the 3rd century BC, the Delhi sultanate in the 13th and 14th centuries AD, the Mughal Empire of the 17th century, and British rule in the 19th century.

India lacked a common political consciousness probably because its social consciousness had developed so strongly. The Indo-Aryan culture of the northern plains spread across the southern part of the peninsula during the second half of the 1st millennium BC, and the Dravidian peoples of the south (Tamils, Kanarese, etc.) were at one with the north in accepting Hinduism and the caste system, a division of the population into groups, based partly on race, partly on occupation. In Hinduism, India cradled the oldest surviving world religion. India's greatest achievements lie in the intellectual and cultural field, and its religious and philosophical systems and Sanskrit literature stand among the finest achievements of the human mind. From the 4th century BC two scripts were in use—Kharosthi in the northwest and the more important Brahmi elsewhere. From the latter, regional modifications developed not only for India but also for Central and Southeast Asia. Indian grammar, law, architecture, sculpture, painting, music, arts, and such crafts as metal casting, enamel work, jewelry, and ivory and wood carving were highly developed. In India the invention of calculation on a system of nine digits and zero took place. Indian art and science grew directly from its religions and philosophies.

In the main this was a Brahmanical achievement, but in the 6th and 5th centuries BC various reactions to Brahmanism began, the most important being the doctrines of Gautama Buddha, which in the form of Buddhism grew into one of the greatest religions in the world. For many centuries the intellectual development of the Hindus depended mainly on the interaction of Buddhism and Hinduism, but Buddhism was finally absorbed and disappeared in India. But it proved acceptable on the frontiers and spread far and wide. Ceylon was converted. In the northwest it crossed the passes into Afghanistan and moved along the trade routes through Turkistan into China, bearing with it in literature, sculpture, and painting various material forms of Indian culture. It passed into Korea and Japan, gradually adapting itself to its new environments. In the 7th century AD Buddhism was imposed on Tibet, and the country remains a stronghold of the faith. To the south, in the early centuries of the Christian Era, Buddhism followed the trade routes across the seas to Southeast Asia; and, as a result, mixed cultures sprang up in which Indian influences are discernible. In this direction, unlike the movements to the north, Hinduism also followed Buddhism; and petty Indianized kingdoms in Burma, Thailand, Malaya, Indonesia, and Indochina were set up, and a movement of traders, scholars, and travellers took place. Some of the greatest surviving architectural creations in the Indian world—for example, Borobuḍur and Lara Jonggrang in Java and Angkor Wat in Kampuchea—were conceived and built in Southeast Asia. The kingdom of Champa in Indochina marked the farthest reach of Indian culture. There it came directly into contact with Chinese civilization, which had molded the adjacent empire of Annam. Champa was overrun by Annamite armies in the 15th century AD.

Spread of Indian influences

In general, the contacts between India and the west took place on the material plane. Trade, for example, between the Roman Empire and Southeast Asia via southern India was considerable even in the early centuries of the Christian Era. On the whole, western Europe was little affected by India until the end of the 18th century.

SOUTHEAST ASIA

The peoples of this extensive region belong to many races, of whom the first with a determinable history were the speakers of the Mon-Khmer languages still used in Pegu Division of Burma and Kampuchea (historical Cambodia). Early in the Christian Era, Indianized kingdoms were established in Cambodia and in Champa.

Cambodia. The earliest Cambodian, or Khmer, kingdom, founded in the 1st century AD, is known only by its Chinese name, Fu-nan. Many of the features that distinguished later Khmer kingdoms made their first appearance in Fu-nan between the 1st and 6th centuries. Based upon the Mekong delta and a well-developed irrigation system, Fu-nan was prosperous and carried on a flourishing trade with China and India. As a kingdom it was organized in the Indian fashion, with a god-king reigning from a holy mountain in the capital at Vyadhapura (Ba Phnom) over surrounding vassal kingdom that extended from southern Vietnam in the east to the Malay Peninsula in the west

and northward over the Korat plateau. Continuous contact with India gave Fu-nan a cultured bureaucratic elite, accomplished in the arts and sciences and techniques of statecraft and artistic expression, while the social structure, manner of life, and beliefs of the Khmer peasant remained fundamentally unchanged.

Internal dynastic struggles brought about the fall of Funan in the latter half of the 6th century and of its successor state, Chen-la, founded on Phum Basac in the middle Mekong valley, in the 9th century. It was Jayavarman II (reigned *c.* 790–850) who established the capital in the region of Angkor, where he and his successors constructed grand monuments. The power of the Khmer Empire reached its height under Suryavarman II (reigned *c.* 113–*c.* 1150), who built the famous Angkor Wat, the ruins of which still stand. His armies ranged as far as North Siam in the west, the Bay of Ban Don in the south, and as far north as the fringe of the Red River delta of Tongking.

Zenith of Khmer empire

Domestic instability caused by the accession of weak rulers to the monarchy left the Khmers open to the attacks of their neighbours, and their difficulties were compounded when Buddhism began to undermine the Brahmanical hierarchy of the state; yet Jayavarman VII (reigned 1181–*c.* 1215/19) was able to extend the empire farther afield than any of his predecessors. The empire quickly crumbled in the 13th and 14th centuries; first as the Thai in the west seized power from their Khmer overlords and the Chams expelled an occupying army, and then as the Mongols extended their support to the Thai and Chams, invading Cambodia from Champa in 1283.

Champa. Champa was an ancient kingdom of Indochina extending over the southeastern coastal region of modern Vietnam from Tourane (Da Nang) in the north to Cape Varella in the south. It was occupied by the Chams, a people of Indonesian stock who had come in contact with the higher civilizations of the Chinese then ruling Tongking and of Indian merchants from overseas. This latter Indian influence prevailed, and Champa became a powerful state, second only to the Khmer Empire of Cambodia.

Appearing during the 2nd century AD under the name of Lin-yi in Chinese histories, Champa first comprised four small states named after Indian regions: Amaravati (Quang-nam), Vijaya (Binh-dinh), Kauthara (Nha-trang), and Panduranga (Phan-rang). Unified later under strong dynasties, these states disappeared one by one in the same order (from north to south) in the course of 14 centuries during which the Chams retreated under the pressure of the Vietnamese. Eventually the race was almost completely annihilated.

The Burmese

Other peoples. Meanwhile, the Burmese, who are linguistically allied to the Tibetans, entered Burma from the northwest. By the 16th century, Burma had become a united kingdom. The Thais, or Siamese, who speak a language of the Chinese type but use an Indian alphabet, infiltrated from southern China and took power in Thailand in the 13th century. The Annamites and Malays are discussed above. All these peoples have been closed about by the cultures of India and of China, and the higher elements of their civilization have been taken from these two primary sources.

EAST ASIA

China and Sinicized Asia. Chinese civilization appeared in northern China in the latter half of the 2nd millennium BC. The discovery at An-yang in Honan province in 1899 of thousands of bone fragments, many of them inscribed, has confirmed the existence of the Shang dynasty (18th–12th century BC), which previously had been thought to be legendary. Early Chinese civilization grew in the Huang Ho plain extending southward toward the Yangtze and westward and northward along the Wei and Fen valleys in Shensi and Shansi provinces. During the Chou dynasty (1111–255 BC), the great formative age of Chinese civilization, the intervening areas populated by groups of a lower culture were conquered and absorbed.

In Han times (206 BC–AD 220) the centre of Chinese culture was still in the north, but by the Sung dynasty (960–1279) the Yangtze valley began to outweigh the north in population and importance. Chinese expansion to the north, which reached the steppelands during the Chou dynasty, was much slower, and it came to a halt on the steppe, among the nomadic herdsmen, where the Chinese system of settled agriculture could not be applied. This conflict between two ways of life resulted in the building by the Chinese of a series of defensive walls, finally linked together by the Ch'in dynasty into the Great Wall.

Korea and Annam came under Chinese dominance in the Ch'in-Han period, but the former broke free again in the 4th century AD and the latter in the 10th century. Both absorbed Chinese culture. On the other hand, Japan, which was a united power by the 4th century AD, never came under Chinese rule. Japan, however, received the first elements of higher culture from China through Korea, and in later times Japan set itself with determination and success to absorb Chinese culture.

In early historical times in China, society was dominated by a hereditary ruling class whose religion, involving the cult of heaven and of the family and clan, was not shared by the masses. The rulers were the custodians of the written language and of the traditions, and the scholars among them gradually formed during a period of political troubles a system of ethics and political theory that the philosophers Confucius and Mencius preserved and transmitted to posterity. These thinkers, moreover, had evolved the rational, ethical ideal of the ruler, the "son of heaven," holding the mandate of heaven but not himself divine, and capable of being replaced if his conduct betrayed his position. Theirs, too, was the ideal of the supremacy of learning and of the scholar-ruler that became the accepted standard of the mandarin (higher civil-service) administrator of Imperial China.

Many centuries passed, however, before the ideal of government through bureaucracy, selected on the basis of learning, reached its fulfillment. Not until the T'ang dynasty (AD 618–907) was the examination system, through which the mandarins were selected, functioning fully. This administrative system undoubtedly provided the backbone of the remarkable political continuity of the Chinese Empire and helped to strengthen the ideal of political unity, "all under heaven," which was throughout a feature of Chinese political theory. The actual achievement of unity under the Ch'in dynasty (221–206 BC) set the standard for the following 2,000 years, a unity that persisted through about 20 successive dynasties. Different as was the empire of the 19th century from that of the Ch'in, it had in fact undergone no major political revolution in the interim. Rebellions might take place, provinces might break away, rulers might change or be changed, but the system persisted.

The Ch'in dynasty

China's artistic achievement, like its ethical and political system, was greatly influenced by the scholar bureaucracy. This is especially true of those arts that were based on the written character, whether in the form of literature or calligraphy or a great deal of Chinese painting. But the mainstream of Chinese culture was also affected from the outside. New ideas, especially in the early centuries of the Christian Era, entered freely from India and the Iranian world. Of these, Buddhism was much the most important, competing with Confucianism for the allegiance of the upper classes, deeply penetrating the later Taoist religion, and providing a pattern for the organization of the Taoist Church. With Buddhism, too, came a deep influence on all Chinese art.

An outstanding characteristic of Chinese civilization was its inventiveness, which produced, among other things, paper, printing, gunpowder, the mariner's compass, the sternpost rudder, and the wheelbarrow. Chinese silks, ceramics, jades, and bronzes early found a market in other parts of Asia and in Europe. The expansion of the Western (Former, or Earlier) Han dynasty (206 BC–AD 25) into Central Asia opened up a major caravan route through Turkistan that for centuries provided a link with the Roman world. China's relations with the nomad peoples affected the movements of peoples throughout Central Asia, from time to time creating repercussions in the Middle East, in Europe, in northern India, and in Iran. By the 2nd and 3rd centuries, Chinese influence was felt in Korea in

the north, and a southern sea route was opened from India and the west around Malaya to Annam and southern China; and by these land and sea routes trade and travellers passed freely. After the rise of Islām in Southwest Asia, trade by sea flourished, and Arab ships were to be seen at Canton and Chinese junks in the Persian Gulf.

In the 8th century the caliphate and the T'ang empires came into direct land contact and conflict (AD 751). Overland contact, interrupted from time to time by the pressures of the steppe peoples, was maintained from this time until the 13th century. Then a sudden major outburst of the Mongols created for a short period a single empire reaching from southern Russia to the Pacific, allowing Europeans for the first time to visit and write about China. The Mongol Empire soon dissolved, but the legend of Cathay that was born in European minds lived on and, in the 16th century, played its part in inducing the Portuguese to venture around the Cape of Good Hope and the Spaniards to cross the Pacific. Thus there came to the West a more detailed knowledge of Chinese civilization, which in the 18th century created in Europe a craze for things Chinese. But by this time the Chinese Manchu Empire was decaying, its administration corrupt and its inspiration dead.

The Mongol empire

Japan. Japan had early begun to assimilate Chinese culture introduced both directly and via Korea, and in AD 645 a deliberate and wholesale introduction of Chinese forms of government took place. But the Imperial administration remained subordinate to a number of great landed-warrior families who fought among themselves for control of the country. The Fujiwara, the Taira, and the Minamoto bore the brunt of the struggle, the Minamoto emerging victorious and in 1192 establishing a dual system under which an emperor ruled in name while the real power rested in the hands of a hereditary military chief called the shogun. This system was carried on by the Ashikaga family (1338–1573) and the Tokugawa family (1603–1868). In the 16th century the Portuguese reached Japan and attempted to introduce Christianity. In the resulting ferment, ideas of conquest developed among a remarkable group of leaders, one of whom, Hideyoshi, organized the invasion of Korea. Death interrupted his plans, and a reaction set in under his successor, Ieyasu, who decided on a policy of isolation. Christianity was forbidden, and Japan was closed to foreigners, and it so remained until after 1854. The early history of Japan was chiefly remarkable for the single-minded way in which its people were able to close ranks and follow a set line of policy such as deliberately setting out to copy Chinese culture, adapting Chinese forms of government, or shutting out the foreigner.

Islām

The term Islām covers the peoples and states that accepted the faith and law of Islām and professed to live by them. Islām begins with the life and teaching of the prophet Muhammad in Arabia in the early 7th century AD. The first Muslim state in Medina erupted in successive waves of conquest over Arabia into the Fertile Crescent and across Persia into Central Asia. Its influence reached into China and northern Asia; at the same time, it had also pushed westward across northern Africa to the Atlantic. Fresh impulses of conquest took Islām into southern Europe and through Spain into France. By sea first and then by land, Islām was also carried into India, thence by sea again southeastward to Malaya and the East Indies. The heart of the Islāmic world, however, was and is the Middle East. The caliphate under the Umayyads of Damascus (661–750) and then the 'Abbāsids of Baghdad (750–1258) became the principal power.

Early conquests

Wherever Islām was accepted it carried with it a sense of unity based on its strictly formulated faith and on its holy law, which, despite much variation of custom over so vast a zone of conquest, remains a common ideal and pattern of belief and conduct for the whole Muslim world. Wherever it has gone it has taken its language, Arabic, which is the holy tongue of Islām, the language of the Qur'ān and of the traditions of the Prophet. Almost all the languages of the Muslim world have borrowed heavily from Arabic and are written in its script. The art and architectural forms of Islām, too, are distinctive, proclaiming the unity of the Islām cultural pattern. Through their military and political power, reinforced by their culture, the Muslims brought together two formerly conflicting worlds, the diversified Mediterranean tradition of the ancient Middle East, Greece, and Rome and the rich civilization of Persia, a fusion which produced great scientific and philosophic developments. Islāmic scholars preserved something of the heritage of Greek antiquity, which was later handed on to Christian Europe. Through the Arabs the Chinese art of papermaking and the Indian system of numerals reached Europe.

Arab power was explosive and quickly burned itself out, and by the 11th century fissiparous tendencies produced a singularly complete collapse of the empire of which the European crusaders and traders in the 12th and 13th centuries took advantage by invading the Middle East. But the peoples on the periphery carried Islām with them. The khanate of the Golden Horde, which between 1241 and 1395 ruled from the Danube to the Urals, was a Muslim state, as also were its successors in the Crimea, the Caucasus, and the Volga. In the 14th century the Islāmized Ottoman Turks brought large areas of Europe under Muslim rule and created an Ottoman Empire that lasted until 1922. Muslim dynasties ruled in Persia, and Islām was carried through Central Asia in two main waves into northern India by the Turks (1000–1526) and the Mughals (1526–1707). The Muslims never fully subjugated southern India, but their rule in the north under such rulers as Akbar and Shāh Jahān was brilliant, both politically and culturally. But Hindu powers, especially in western India, struck back at the Muslims in the 17th century, creating disorder and disrupting the empire. In these disturbed conditions, European trading companies already established on the Indian coasts, especially the English and the French, were drawn in to compete for political supremacy.

Through trade Islāmic influence had reached out in the 16th century from India to Malaya and Indonesia, peacefully overlaying Hindu culture in the main centres, though not in Bali.

The influence of Asia on Africa was until the 19th century greater than that of Europe. The ancient Middle East drew on the resources both in men and material of the regions beyond the Upper Nile Valley. The ancient Abyssinian kingdom was founded by Semites from southern Arabia, and Islām penetrated both north and south of the Sahara across to western Africa. There was a continuous Arab migration to eastern Africa, resulting in the founding of a series of cities on the coast. There was also an ancient connection between India and eastern Africa, which was reinforced in the 19th century by Indian immigration.

Asian influences in Africa

The great civilizations of Asia spread out over immense areas. Although they were separated one from the other by distance and slow communication, their contacts were numerous and between them the cross-fertilization of ideas and material culture was extensive. They were slow growing and slow changing, based in the main on subsistence economies. Their upper classes, however, created an active trade in ideas and luxury goods and in each civilization developed a way of life, whether Indian, Chinese, or Islāmic, that was chiefly remarkable for its inner harmony. Each religion or ethical code formed a way of life for its peoples, and all aspects of life served to express this. Nevertheless, that their ways of life were very different one from the other is to be seen, for example, in their respective attitudes toward history. The Chinese regarded the writing of their own history as an important state function and a significant branch of literature. Official historians were regularly at work producing official histories of each dynasty. In this way the doctrines of Confucianism were perpetuated, and the official record of the times was made to conform to the accepted doctrines. What man did in life was taken to be supremely important. The civilizations of the ancient Middle East and of Islām produced histories and chronicles justifying the ways of God to man and, in the case of Islām, seeking to show that a good Muslim would be a good ruler or a good subject. What

a Muslim did was necessarily important. In contrast, the Brahmans of India disregarded their own history. They did not regard the recording of man's practical activities as of any significance compared with spiritual issues. These differences illustrated the fact that Oriental civilizations did not find uniform solutions to their problems.

The modern era

ASIA AND WESTERN DOMINANCE

At the close of the 15th century—that is, about a quarter of a century before the Mughals began to conquer northern India and a half-century before the last great dynasty of China (the Manchu) established itself—a Portuguese fleet under Vasco da Gama found its way around the Cape of Good Hope into Indian waters. While the Mughals went on to conquer the Indian mainland, the Portuguese made themselves masters not only of Indian waters but of all the eastern seas. Where they led, the Spanish, Dutch, French, and English followed. Above all, trade was sought, especially in silks and spices; but various attempts were also made to spread Christianity, especially by the Portuguese and Spanish, who had themselves felt the proselytizing impact of Islām in Europe and North Africa and were not reluctant to carry the fight to the East.

European contact

During the 15th to the 19th centuries Europe itself was changing rapidly. Nation-states were growing up. The foundations of a scientific, industrial, and technological revolution were laid. Europe learned how to navigate the oceans, apply sea power, wield artillery, organize representative government, cultivate religious toleration, and use money and credit in trade. Thus, it was a supremely confident Europe that began to feel its way along the coasts of the traditional, slow-changing societies of Asia.

The Portuguese quickly established themselves in Goa (India) and Macau (Macao) off the Chinese mainland; the Spanish took the Philippines in 1565. The rivalries of Europe were extended to Asian waters, and the Dutch evicted the Portuguese and consolidated a spice empire in Java and Sumatra, centring on Batavia (Jakarta). Between 1740 and 1805 the French and British on the coasts of India struggled with each other for supremacy, which the British won, in the process being drawn into the vacuum that the collapse of the Mughal Empire had created. During the 19th century British power easily spread across India to the Himalayas, and the whole subcontinent and its people became a major field of British investment. To protect this, new British bastions of power were erected in Aden, Persia, Arabia, Burma, and Singapore.

Meanwhile the French, pushed out of India and forestalled in Burma, established control over Cochin China, Annam, Tongking, Laos, and Cambodia. The independence of Siam (Thailand) was preserved because of its buffer position between the British in Burma and the French in Indochina.

In the Middle East, the Ottoman Empire was crumbling under internal stresses and external pressures. In 1882 the British took control of Egypt in order to safeguard the new route to India through the Suez Canal (1869), and the French, who had already crossed the Mediterranean into northern Africa, later found a foothold in Syria. The Far East was to some extent protected by distance and the rivalries of the European states elsewhere, but in 1842 and thereafter China was forced to open some of its ports to foreign trade and residence and to accept treaty provisions giving a special position to foreigners, an "Open Door" status that was substantially maintained until the 1930s. Beginning in 1854, the United States opened Japan, under arrangements similar to those forced on China, and, in turn, Japan opened Korea (1876).

The opening of China and Japan

Meanwhile, across central and northern Asia, Russia was advancing by military force, political manipulation, and colonization. By 1900 Russia was established on the Pacific coast, and both China and Korea were being threatened from land and sea.

THE RECOVERY OF ASIA

By the late 19th century the peoples and governments of Asia had begun to react to the impact of Europe. Japan, a tightly organized, compact island state, moved first by again making a deliberate decision to acquire the material power and organization of the apparently superior power, this time the European. By 1894 the legal and judicial order had been modernized; a new school system had been created; the foundations for a modern economy had been laid; the military system had been modernized; and a constitutional system had been established (1889). These changes enabled Japan to negotiate revision of its "unequal" treaties with the Western states. Its new power also enabled Japan to defeat China in 1894–95 and Russia in 1904–05. These two wars enlarged Japan to include Formosa (1895) and Korea (1910) and gave it a position in Manchuria from which to move in empire building on the Asian mainland.

European domination in Asia, together with the introduction of "Western learning" into Asian countries, gave rise to nationalist movements. Capitalist enterprise and higher-education policies in the Islāmic world, China, Southeast Asia, and India had brought into existence new intellectual and middle classes, which sought to introduce into their several countries the political and economic institutions of the West but, at the same time, to reject Western political and economic supremacy. Everywhere in Asia nationalism and anticolonialism formed the coordinates of these movements. Although nationalism itself tended to isolate the Asian countries into self-contained units, the individual nationalist movements interacted on one another. The victory of Japan over Russia and the early successes of the Chinese reform and revolutionary movements stimulated action elsewhere in Asia. The United States, after displacing Spain in the Philippines (1898), committed itself to Philippine autonomy. The rise of the freedom movement in India encouraged similar developments in Ceylon, Burma, and Indonesia. The revival of Islām in the Middle East after World War I stimulated Muslim consciousness in India and Southeast Asia.

In China the first reaction to the Western impact was anti-Manchu as well as antiforeign. The failure of the Boxer Rebellion (1900) inaugurated a period of Manchu reform designed to avoid revolutionary change. Revolutionaries were, however, successful in replacing the dynasty with a parliamentary republic in 1911–12. The traditional provincialism, however, soon made ineffective the national parliamentary republic, although China as a state retained its identity. National political unity began to be reestablished when the Kuomintang, or Nationalist Party, took over the government of China (1928). At the same time, however, Marxist–Leninist ideas were introduced into China with the organization of the Chinese Communist Party, affiliated with the Communist International. In the ensuing struggle the Nationalist government maintained its ascendancy until 1949. The establishment of the People's Republic by the Chinese Communist Party not only inaugurated the Communist phase of the Chinese revolution but also enabled Chinese and Russian influence to replace Western in a number of peripheral areas. Japan's defeat in World War II resulted in the loss of its empire. Korea became independent but was divided. But following the occupation entailed by defeat, Japan again quickly reestablished itself economically and resumed its position among the important states.

Chinese civil struggle

In the Indo-British empire an effort was made by the British to establish forms of representative and responsible government. But the attempt to hand over power roused the self-consciousness of the Muslim and Hindu communities so that the subcontinent, when freed in 1947, was also partitioned into two independent countries—Pakistan and India. Nationalism had been sufficiently strengthened in Burma and Ceylon during World War II so that Britain was obliged to concede independence by 1947. A new relationsip with Malaya was similarly established a decade later. With British imperial responsibilities being thus largely liquidated, other European powers began to yield their positions. The United States granted independence to the Philippines in 1946. The Dutch gave way to the Indonesian nationalists (1949); and the French, by 1954, had been compelled to withdraw from Indochina. The U.S. occupation of Japan was officially terminated

by the peace treaty that became effective in 1952. Thus the tide of the West in Asia was reversed after World War II. Internal political and economic stability, however, had not always been fully established following independence. In the background the U.S.S.R., on the landward side, strengthened its grip on central and northern Asia, and the United States, on the seaward side, sought to fill the power vacuum left by the Japanese defeat in the Pacific, with the U.S.S.R. supporting Communist Vietnam against the United States and its anti-Communist allies. In the Middle East the Arab and Muslim states had regained or reasserted their independence, but they were small, weak, and divided. Continuing hostility to the new Jewish state, Israel, after its proclamation (1948), as well as fear of European domination, served at least as a symbolic bond of union for the Arabs. Turkey, which had become a republic in 1923, deliberately, abandoned Islām as the state religion and under Kemal Atatürk began to turn itself into a modern Westernized nation-state.

Economic and social developments in Asia had not kept pace with the rate of political change. An enormous increase in population, especially in India, China, and Japan, had occurred; and, except in Japan, the masses were probably poorer at the close of the period of European dominance than they had been at the beginning. The disparity between the masses and their new, mainly middle-class rulers had become more glaring than ever; and the new nations of Asia, though free, were neither secure nor stable. In short, the influence of the West had created revolution in Asia and had complicated, not simplified or clarified, the problems facing Asians.

(C.H.Ps./Hd.M.V./Ed.)

ASIAN GEOGRAPHICAL FEATURES OF SPECIAL INTEREST

Mountain ranges

ALTAI MOUNTAINS

A great topographic spine of Inner Asia, the Altai straddle parts of China, Mongolia, and the Soviet Union, stretching more than 1,000 miles (1,600 kilometres) southeast to northwest, from the Gobi (desert) to the West Siberian Plain. The jagged mountain ridges, whose name derives from the Turkish–Mongolian *altan,* meaning "golden," separate the waters of such great rivers as the Ob (flowing north to the Arctic Ocean) and its major tributary, the Irtysh, from the rivers draining into the vast Central Asian basin. There are three main prongs of the system—the Soviet, Mongol, and Gobi Altai; and the Soviet peak, Belukha, at 14,783 feet (4,506 metres) is the highest point. Although historically these mountains have been something of a remote barrier, the Altai are now famous for their mining and hydroelectric potential, while the ancient ways of life of their peoples are being rapidly transformed.

Physical features. *Physiography.* The Soviet Altai lies in eastern Kazakhstan and the Gorno-Altai sector of the Russian Soviet Federated Socialist Republic. A belt of northern foothills separates it from the rolling West Siberian Plain, while in the northeast it borders the Zapadny (Western) Sayan uplands and the mountains of the Lebed River Basin. The Mongol Altai (rising to Hüyten Orgil, over 14,000 feet) thrusts away to the southeast, then eastward. The Gobi Altai begins some 300 miles southwest of Ulaanbaatar, the Mongolian capital, and occupies the country's southern portions, towering over the Gobi expanses.

Geology. The Altai were originally formed in the great geological upthrusts of the Hercynian orogeny, about 300,000,000 years ago, and were worn down, over the long periods of geological time, into a peneplain, a plateau with generally accordant summit heights. At the beginning of the Quaternary Period, about 2,500,000 years ago, new upheavals thrust up magnificent peaks of considerable size along a zone of weakness in the Earth's surface. Quaternary glaciation scoured the mountains, carving them into rugged shapes, and deepened valleys from a V- to a U-shaped cross section; river action was also intensive and left its marks on the landscape. As a result of these differential geological forces, the highest ridges in the contemporary Altai—notably the Katun, Severo- (North) Chu, and the Yuzhno- (South) Chu—tower more than 13,000 feet, running latitudinally in the central and eastern portions of the Soviet sector of the system. The Tabyn-Bogdo-Ola (Tavan Bogd Uul in Mongolian), the Mönh Hayrhan Uul, and other western ridges of the Mongol Altai are somewhat lower. The highest peaks are much steeper and rockier than their Alpine equivalents, but the ranges and massifs of the middle Altai, to the north and west, have ridges of about 8,200 feet, whose softer outlines betray their origins in presenting ancient, smoothed surfaces. Valleys are nevertheless jagged and gorgelike. The ridges are separated by steppelike structural hollows (notably the Chu, Kuray, Uymen, and Kansk), which are filled with crumbly deposits up to 50,000,000 years old. Heights range from 1,600 to 6,600 feet above sea level.

The extreme dislocations suffered by the Altai over the course of geological time have occasioned a variety of rock types, many of them altered by volcanic activity when great former geosynclines, sediment-filled structural downwarps, were thrust upward in mountain-building epochs. There are considerable accumulations in the various structural depressions. The layers of nonferrous and rare metal-bearing rocks contain exploitable deposits of iron, mercury, gold, manganese, and tungsten that have a very high commercial value.

Climate. The regional climate is severely continental: because of the influence of the great Asiatic anticyclone, or high-pressure complex, the winter is long and bitterly cold. January temperatures range from 7° F (−14° C) in the foothills to −26° F (−32° C) in the sheltered hollows of the east, while in the Chu steppes temperatures can plunge to a bitter −76° F (−60° C). There are occasional tracts of the permanently frozen soil (or permafrost) that coats great stretches of the Soviet Far North. July temperatures are warm (often to 75° F [24° C], sometimes up to 104° F [40° C]) on the lower slopes, but summers are short and cool in the higher portions. In the exposed west, particularly at heights of between 5,000 and 6,500 feet, precipitation is high: as much as about 80 inches (2,030 millimetres) may fall quite evenly throughout the year. The amount decreases by a third farther east, and some areas have no snow at all. Glaciers coat the flanks of the highest peaks: some 1,200 in number, they cover an area approaching 600 square miles.

Temperature ranges

Drainage. The Soviet and Mongol portions of the Altai are criss-crossed by a network of turbulent, youthful rivers, fed mainly by melted snow and summer rains, occasioning spring and summer floods. The Katun, Bukhtarma, and Biya are among the biggest. Rivers of the Gobi Altai are shorter, shallower, and often frozen in winter and dry in summer. There are more than 3,500 lakes, most of structural or glacial origin. Those of the Gobi Altai are often bitterly salty.

Plant life. There are four fairly distinct vegetation zones: mountain subdesert, mountain steppe, mountain forest, and the Alpine regions. The first, found on lower slopes and in hollows of the Mongol and Gobi Altai, reflects the high summer temperatures and low rainfall: the sparse life includes wormwood and halophytic plants (those growing in salty soils). The mountain steppe zone rises to about 2,000 feet in the north and to 6,600 feet in the south and east. Meadows and mixed grass steppes characterize the former, and grassy herbs of the wormwood-fescue type the latter: only in the higher areas are mountain vines and bush steppes found. The mountain forest zone is most characteristic of the Soviet Altai, where it covers about 70 percent of the territory, mostly in the low and medium

mountain regions. Forests reach up to 6,500 feet but climb to 8,500 feet on less exposed slopes of the central and eastern Altai. Trees include coniferous growths, birches, aspens, and larches, and there are many huge clearings. A forest belt is practically nonexistent in the Mongol and Gobi Altai, but isolated clumps of Siberian trees grow in sheltered spots. Alpine vegetation—sub-Alpine shrubs giving way to meadows widely used for summer pasture and then mosses and bare rock and ice—is found only on the highest ridges.

Animal life. Animal life follows vegetation patterns. Various small rodents populate the mountainous semideserts and steppes, while bird life includes eagles, hawks, and kestrels. Most birds and animals are of Mongolian origin: marmot, jerboa (a jumping rodent), and antelope. Siberian animals (bear, lynx, musk deer) and birds (grouse, woodpecker, and crossbill) frequent the moist coniferous forests. Alpine animal life includes the mountain goat, snow leopard, and mountain ram.

The people. The Soviet Altai is peopled by groups of Altaic, Russian, and Kazakh origin. Only in the 1960s did the formerly tribal Altaic peoples—distinguished on the basis of dialect—cohere into a unified group. Their principal occupation is the breeding of cattle or sheep. The Russians are mainly agriculturalists or, significantly, miners. The Kazakhs mostly raise sheep, cattle, and wild horses, utilizing the summer alpine pastures (and erecting temporary dwellings), while keeping their stock under shelter in winter. The Mongol Altai is the home of groups of Kazakhs, Khalkhas, and the nomadic Oyrat. Sheep, goats, and camels are raised in the drier south, with large horned cattle becoming prominent in the north. Horse breeding is widespread. The shortages of water in the south necessitate migratory movements of up to 120 miles on the part of the nomads, who erect temporary villages. The development of agriculture, however, is producing more settled ways of life.

Peoples and their ways of life

Study and exploration. Scientific research into the Altai dates only from the 18th century but was stimulated when gold was discovered in 1828. Russian geologists and engineers pioneered the collection of data. Soviet expeditions from the U.S.S.R. Ministry of Geology and Tomsk V.V. Kuibyshev State University and expeditions from the Academy of Sciences of the Mongolian People's Republic have uncovered a wealth of geological and topographical data, and many peaks have been scaled. (N.I.M.)

HIMALAYAS

The Himalayas of Asia include the highest mountains in the world, with more than 30 peaks rising to heights of 24,000 feet (7,300 metres) above sea level. One of these peaks is Mt. Everest, the world's highest, which reaches a height of 29,028 feet (8,848 metres). The great heights of the mountains rise above the line of perpetual snow. The vast permanent snowfields attracted the attention of the pilgrim mountaineers of ancient India, who coined the Sanskrit name Himalaya—from *hima,* "snow," and *ālaya,* "abode"—for this great mountain system. In modern times, the Himalayas have constituted the greatest attraction and the greatest challenge to mountaineers throughout the world.

Forming the northern border of the Indian subcontinent and an almost impassable barrier between it and the lands to the north, the ranges form part of the great mountain belt stretching halfway around the world from northern Africa to the east coast of Asia. The Himalayas themselves stretch uninterruptedly for about 1,550 miles (2,500 kilometres) from west to east between Nānga Parbat (26,660 feet), in the disputed state of Jammu and Kashmir, and Namcha Barwa (25,445 feet), in Tibet. Between these eastern and western extremities lie the three Himalayan kingdoms of Nepal, Sikkim, and Bhutan. The Himalayas are bordered to the northwest by the mountain ranges of the Hindu Kush and Karakoram and to the north by the high Plateau of Tibet. The width of the Himalayas from south to north varies between 125 and 250 miles. Their total area amounts to about 229,500 square miles (594,400 square kilometres).

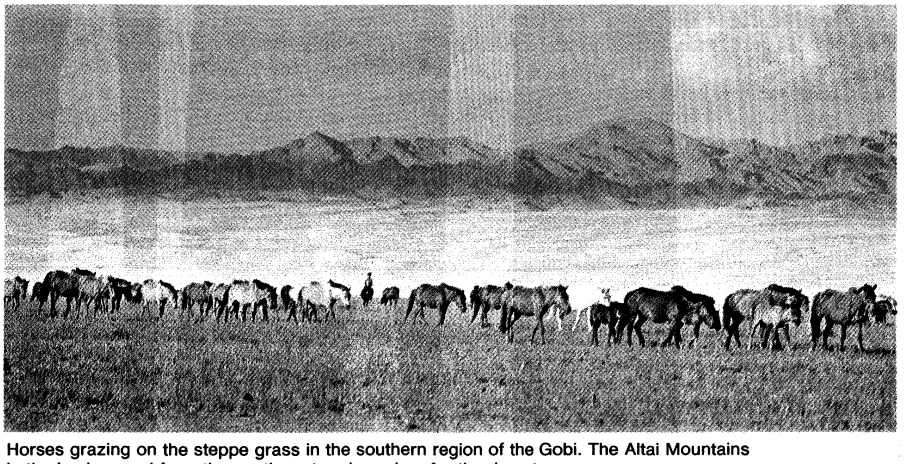
Though India has sovereignty over most of the Himalayas, Pakistan and China occupy parts of them. In the state of Jammu and Kashmir, Pakistan has administrative control of 32,362 square miles "line of control" of the range lying north and west of a cease-fire line established between India and Pakistan in 1972. China's occupation of 14,000 square miles in the Ladākh district of Kashmir, as well as Chinese incursions to the south of the McMahon Line (a 1914 boundary line limiting Tibetan sovereignty in the Assam Himalayas of northeast India) in the North East Frontier Agency (now Arunachal Pradesh) in 1962, have accentuated further the boundary problems faced by India in the Himalayan region.

Physical features. The Himalayas' most characteristic features are their soaring heights, snowcapped and steep-sided jagged peaks, valley glaciers often of stupendous size, topography deeply cut by erosion, seemingly unfathomable river gorges, complex geological structure, and a rich temperate and alpine vegetation. Viewed from the south, the Himalayas appear as a gigantic crescent, with its main axis rising above the snow line, where snowfields feed the valley glaciers and constitute the sources of most of the Himalayan rivers. The greater part of the Himalayas, however, lies below the snow line. The mountain-building process that created the range is still active and is accompanied by erosion by rivers and landslides of great dimension.

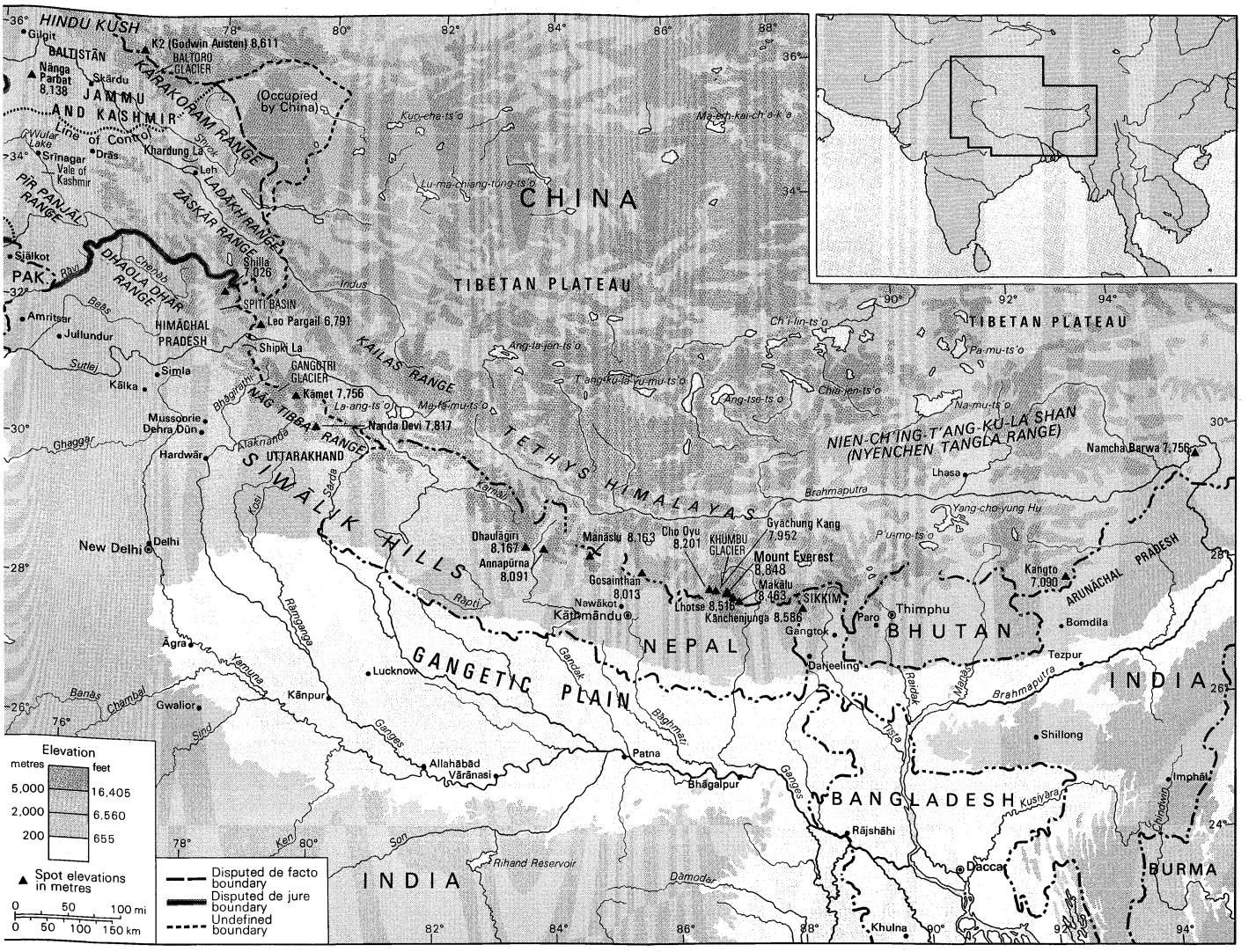
From south to north, the Himalayan ranges can be grouped into four parallel, longitudinal mountain belts of varying width, each having distinct physiographic features and its own geological history. They are designated as the Outer, or Sub-Himalayas; the Lesser, or Lower Himalayas; the Great, or Higher, Himalayas; and the Tethys, or Tibetan Himalayas. Farther north lie the Trans-Himalayas in Tibet proper, eastward continuations of some of the most northerly Himalayan ranges. From west to east the

Principal divisions

Paolo Koch—Rapho/Photo Researchers



Horses grazing on the steppe grass in the southern region of the Gobi. The Altai Mountains in the background form the southwestern boundary for the desert.



The Himalayan region of the Indian subcontinent.

Himalayas are divided broadly into three mountainous regions: Western, Central, and Eastern.

Physiography. The Outer Himalayas comprise flat-floored structural valleys and the Siwalik Hills, which border the Himalayan mountain system to the south. Except for small gaps in the east, the Siwalik run for the entire length of the Himalayas with a maximum width of 62 miles in the Indian state of Himachal Pradesh. In general, the 900-foot contour line marks their southern boundary; they rise to another 2,500 feet to the north. The main Siwalik range has steeper southern slopes facing the Indian plains and descends gently northward to flat-floored basins, called *dūns*. The best known of these is the Dehra Dūn, in Uttarakhand, which is in the mountainous parts of Uttar Pradesh.

Northward, the Siwalik Range abuts against a 50-mile-wide massive mountainous tract, the Lesser Himalayas, where mountains rising to 15,000 feet and valleys with altitudes of 3,000 feet run in different directions. There is a general conformity of altitude among neighbouring summits, which creates the appearance of a highly dissected plateau. The three principal ranges of the Lesser Himalayas, the Nāg Tibba, the Dhaola Dhār, and the Pīr Panjāl, have branched off from the Great Himalayan Range lying farther north. The Nāg Tibba, the most easterly of the three ranges, is 26,795 feet high near its eastern end, in Nepal, and forms the watershed between the Ganges and the Yamuna, in the Uttarakhand.

To the west, the picturesque Vale of Kashmir, a structural basin (*i.e.*, an elliptical basin in which the rock strata are inclined toward a central point), forms an important section of the Lesser Himalayas. It extends from southeast to northwest for 100 miles, with an average elevation of 5,100 feet, having a width of 50 miles; it is traversed by the meandering Jhelum River, which runs through the Wular Lake, the largest freshwater lake in India.

The backbone of the system is the Great Himalayas, a range rising above the line of perpetual snow.

Highest peaks of Himalayas

The Great Himalayan Range rises to its maximum height in Nepal, having in that section nine of the 14 highest peaks of the world. From west to east they are: Dhaulāgiri 1 (26,810 feet; 8,172 metres), Annapūrna 1 (26,504 feet; 8,078 metres), Manāslu 1 (26,760 feet; 8,156 metres), Cho Oyu (26,750 feet; 8,153 metres), Gyāchung Kang 1 (25,991 feet; 7,922 metres), Mt. Everest (29,028 feet; 8,848 metres), Lhotse 1 (27,923 feet; 8,511 metres), Makālu 1 (27,824 feet; 8,481 metres), and Kānchenjunga 1 (28,208 feet; 8,598 metres).

Farther east the range changes from a southeasterly to an easterly direction as it enters Sikkim. After this, it runs eastward for another 260 miles through Bhutan and the eastern part of Arunachal Pradesh as far as the peak of Kangto (23,260 feet) and finally turns northeast, terminating in Namcha Barwa.

There is no sharp boundary between the Great Himalayas and the ranges, plateaus, and basins lying to the north of the Great Himalayan Range, generally grouped together under the name of the Tethys Himalayas and extending far northward into Tibet. In Kashmir the Tethys are at their widest, forming the Spiti Basin and the Zāskār Mountains, the highest peaks of which, to the southeast, are Leo Pargail (22,280 feet; 6,791 metres), rising north of the Sutlej River opposite Shipki La (pass), and Shilla (23,050 feet; 7,026 metres).

Geology. A study of the geological history of the Himalayas reveals that marine sediments of Paleozoic and Mesozoic eras (between about 65,000,000 and 570,000,000 years ago) deposited on the floor of the ancient Tethys Sea; the frontal part of the crystalline massif (mountain mass) of peninsular India; estuarine deposits along the flanks of the embryonic mountains; and finally products of surface erosion of the rising mountains—all contributed to the formation of the present-day range. The uplift of the Himalayas took place in at least three distinct and widely separated phases. The first phase of the major mountain-building movement took place at the close of the Eocene Epoch (about 38,000,000 years ago), although the beginning of the main Himalayan uplift started in Middle and Upper Cretaceous times (from about 100,000,000 to 65,000,000 years ago), with the advancement of the crystalline massif of peninsular India toward the Plateau of Tibet. This movement caused the rise of the Tethys Himalayas, along with the greater part of the Great Himalayas. In the second phase of upheaval, which occurred in the Miocene Epoch (7,000,000 to 26,000,000 years ago), the estuarine deposits and the Indian Massif formed the ranges of the Lesser Himalayas. The final mountain-building phase started at the end of the Tertiary Period (about 7,000,000 years ago), lifting the detrital deposits accumulating at the base of the Himalayas to form the Siwalik Range, the foothills of the Outer Himalayas. Since the middle Pleistocene Epoch (about 1,500,000 years ago), the Himalayas have risen at least 4,500 feet, an occurrence witnessed by early man. That the Himalayas still continue to rise is evidenced by the upheaval of the younger river terraces.

Formation of the ranges

Precambrian metamorphic rocks (rocks formed by heat and pressure from 570,000,000 to 4,600,000,000 years ago) form the bulk of the Himalayas. These rocks represent the frontal part of the Indian Shield, which, according to the theory of continental drift, pushed northward, uplifting the Himalayas as it pressed against the Asian landmass. Only in the Spiti Basin and in a few other localities can large outcrops of marine sediments from Paleozoic and Mesozoic times be seen.

Plutonic rocks (formed deep down from a molten state), such as granites and granodiorites of pre-Miocene age (*i.e.,* more than 26,000,000 years old), outcrop in extensive areas in north Kashmir, forming the whole of the Ladākh Range and the greater part of the area of Baltistān to the west of the Drās River and to the south of the Karakoram Range. The intrusion during post-Miocene times (*i.e.,* within the last 7,000,000 years) of tourmaline granite into an older series of gneisses (rocks formed by heat and pressure and made of bands that differ in colour and composition) and schists (crystalline rocks, the constituent minerals of which are usually arranged in a foliated or parallel pattern), aided by upthrusts in many areas, has given rise to many of the high peaks of the Himalayas, such as Makālu, Manāslu, and Nānga Parbat, which are typical examples. Mt. Everest and its two associated peaks, Lhotse and Cho Oyu, are, however, formed of limestones and pelitic (clay and mudstone) rocks, the latter dipping toward the north. It is possible that the entire formation of Mt. Everest is thrust up over a foundation of gneiss, as nappes or overturned folds.

The nappe structure of the Himalayas can also be seen elsewhere in the ranges. The Krol Nappes of the Simla region, in Himachal Pradesh, and the Garhwal Nappes, of Uttarakhand, are typical. These nappes are also evident in Nepal in the Nawākot and Kāthmāndu areas, which were formed along the line of the main central thrusts. This thrust zone borders the Great Himalayas to the south, rising abruptly in height and showing changes in the sequence of geological beds. The Pīr Panjāl Range, for example, owes its origin to thrust faulting; in the Simla region Krol limestones of Carboniferous Period (*i.e.,* from 280,000,000 to 345,000,000 years old) are overthrust onto much younger deposits of Pliocene Epoch (from 7,000,000 to 2,500,000 years old).

The geological structure is much simpler in the Outer Himalayas, where the foothills are mainly composed of Tertiary formations (from 2,500,000 to 65,000,000 years old), grouped under the Lower, Middle, and Upper Siwaliks. These consist mostly of freshwater deposits, such as sandstones, shales, and conglomerates. The Lower Siwaliks are from 1,800 to 6,000 feet thick, the Middle Siwaliks from 3,000 to 4,500 feet thick, and the Upper Siwaliks from 4,500 to 6,000 feet thick. At the same time that the Siwalik deposits were occurring, lacustrine (lake) deposits known as Karewas (flat-topped terraces) were being formed in the Vale of Kashmir. Both the Karewas and Siwaliks show evidence of glaciation during Pleistocene times (from 10,000 to 2,500,000 years ago).

Drainage. The Himalayas are drained by 19 major rivers, of which the Indus and the Brahmaputra are the largest, each having catchment basins about 100,000 square miles in extent in the mountains. Of the other 17 rivers, five belong to the Indus system—the Jhelum, Chenāb, Rāvi, Beās, and Sutlej, with a total catchment area of 50,958 square miles; nine belong to the Ganges system—the Ganges, Yamuna, Rāmganga, Kāli (Sarda), Karnāli, Rāpti, Gandak, Bāghmati, and Kosi, draining another 84,098 square miles; and three belong to the Brahmaputra system—the Tista, Raidak, and Manās, draining another 70,769 square miles.

Major Tibetan rivers

Most of the Himalayan rivers flow in troughs, the trends of which are generally determined by the branching ranges of the Great Himalayas. The rivers of the Indus system as a rule follow northwesterly courses, whereas most of the rivers of the Ganges–Brahmaputra systems take easterly courses while in the mountain region.

To the north of India, the Karakoram Range, with the Hindu Kush (Mountains) on the right and the Ladākh Range on the left, forms the great water divide, shutting off the Indus system from the rivers of Central Asia. The counterpart of this divide on the east is formed by the Kailas Range and its eastward continuation, the Nienchʼing-tʼang-ku-la Shan (Nyenchen Tangla Range), which prevent the Brahmaputra from flowing northward. South of this divide, the Brahmaputra flows eastward for about 900 miles before cutting across the Great Himalayan Range in a transverse gorge, although many of its Tibetan tributaries flow in an opposite direction, as the Brahmaputra may once have done.

The Great Himalayan Range, which normally would form the main water divide throughout its entire length, functions as such only in limited areas. This situation exists because the major Himalayan rivers, such as the Indus, Brahmaputra, Sutlej, and at least two headwaters of the Ganges—the Alaknanda and Bhāgirathi—are older than the mountains they traverse. It is believed that the Himalayas were uplifted so slowly that the old rivers had no difficulty in continuing to flow through their channels and, with the rise of the Himalayas, even acquired a greater momentum, which enabled them to deepen their valleys more rapidly. The elevation of the Himalayas and the deepening of the valleys thus proceeded simultaneously, with the result that the mountain ranges emerged with a completely developed river system cut into deep transverse gorges, ranging in depth from 5,000 feet to 16,000 feet and in width from six to 30 miles. The earlier origin of the drainage system explains the peculiarity that the major rivers drain not only the southern slopes of the Great Himalayan Range but to a large extent its northern slopes as well, the water divide being north of the crest line.

The role of the Great Himalayan Range as a watershed can, nevertheless, be seen between the Sutlej and Indus valleys for 360 miles; the drainage of the northern slopes is carried by the north-flowing Zāskār and Drās rivers, which drain into the Indus. Glaciers also play an important role in draining the higher altitudes and in feeding the Himalayan rivers. Several glaciers occur in Uttarakhand, of which the largest, Gangotri, is 20 miles long. The Mahalangur *himāl* ("snowfield"), with its Khumbu Glacier, drains the Everest region in Nepal. The rate of movement of the Himalayan region glaciers varies considerably; in the neighbouring Karakoram ranges, for example, the Baltoro Glacier moves about six feet a day, while others, such as the Khumbu, move only about a foot daily. Most of the Himalayan glaciers are in retreat.

Glaciers

Soils. Not much is known about the Himalayan soils. The north-facing slopes generally have a fairly thick soil

cover, supporting dense forests at lower altitudes and grasses higher up. The forest soils are dark brown in colour and silt loam in texture and occur mainly in Uttarakhand; they are ideally suited for growing fruit trees. The mountain-meadow soils are well-developed but vary in thickness and in their chemical properties. Some of the wet, deep, upland soils of this type in the Eastern Himalayas—for example in the Darjeeling Hills and in the Assam Valley—have a high humus content that is good for growing tea. Podzolic soils (infertile, acidic forest soils) occur in a 400-mile-long belt along the valley of the Indus and of its tributary the Shyok, to the north of the Great Himalayan Range, and are also found in patches in Himachal Pradesh. Farther east, saline soils occur in the dry Ladākh Plateau. Of the soils that are not restricted to any particular area, alluvial soils (deposited by running water) are the most productive, though they occur in limited areas, such as the Vale of Kashmir, the Dehra Dūn and on the high terraces flanking the Himalayan valleys. Lithosols consisting of imperfectly weathered rock fragments deficient in humus content cover many large areas at high altitudes and are the least productive soils.

Climate. The Himalayas, as a great climatic divide affecting air- and water-circulation systems, exercise a dominating influence upon meteorological conditions in the Indian subcontinent, to the south, and in the Central Asian highland, to the north. By its situation and stupendous height, the Great Himalayan Range obstructs the passage of cold continental air from the north into India in winter and also forces the southwest monsoonal (rain-bearing) winds to give up most of their moisture before crossing the range northward, thus causing a heavy precipitation of rain and snow on the Indian side but arid conditions in Tibet. The average annual rainfall on the south varies between 60 inches (1,525 millimetres) at Simla and Mussoorie in the Western Himalayas and 120 inches at Darjeeling in the Eastern Himalayas. At places such as Skārdu, Gilgit, and Leh, in the Indus Valley, to the north of the Great Himalayan Range, only three to six inches of rainfall occur.

The precipitation pattern

Local relief and situation determine the meteorological variations experienced not only in different parts of the Himalayas but even on different slopes of the same range. Because of its favourable location on top of the Mussoorie Range facing the Dehra Dūn, the town of Mussoorie, at a height of about 6,100 feet, receives 92 inches of rainfall annually, as against 62 inches recorded in the town of Simla, which lies behind a series of ridges at a height of 6,600 feet. The Eastern Himalayas, being at a lower latitude than the Western Himalayas, are relatively warmer; the lowest minimum temperature so far recorded was at Simla, in the Western Himalaya, −13° F (−25° C). The average minimum temperature for the month of May, recorded in Darjeeling at 6,380 feet elevation, is 52° F (11° C). In the same month, at an altitude of 16,500 feet in the neighbourhood of Mt. Everest, the minimum temperature is about 17° F (−8° C); at 19,500 feet it falls to −8° F (−22° C), the lowest minimum being −21° F (−29° C). At this time during the day, in areas sheltered from strong winds that blow at more than 100 miles an hour, the sun is often pleasantly warm, even at that altitude.

There are two periods of wet weather—the winter rains and the rains brought by the southwest monsoon winds. Winter precipitation is due to the depressions advancing into India from the west, causing heavy falls of snow. Within the regions where western disturbances are felt, condensation takes place in upper air levels at a height of 10,000 feet from the surface; as a result, precipitation is much greater over the high mountains. It is at this season that snow accumulates around the Himalayan high peaks and that the Western Himalayas receive more rain than the Eastern Himalayas. In January, for example, Mussoorie in the west receives almost three inches, while Darjeeling to the east receives less than an inch. By the end of May, meteorological conditions are reversed. Southwest monsoon currents passing over the Eastern Himalayas reach heights of 18,000 feet; in June, therefore, Darjeeling receives about 24 inches, and Mussoorie less than eight inches. The rains cease in September, after which the finest weather in the Himalayas prevails until the beginning of winter in December.

Plant life. Himalayan vegetation can be broadly classified into four groups: tropical, subtropical, temperate, and alpine. This division is based mainly on altitude and rainfall. Local variations in relief and climate, as well as exposure to sun and winds, cause considerable variation in the composition of the vegetation within each group. Tropical evergreen rain forest is confined to the humid foothills of the Eastern and Central Himalayas. The evergreen dipterocarps—a group of timber- and resin-producing trees—are common; their different species grow on different soils and on hill slopes of varying steepness. *Mesua ferrea* (rose chestnut) occurs on porous soils at altitudes between 600 and 2,400 feet (185–730 metres); bamboos grow on very steep slopes; oaks and chestnuts grow on the lithosol, covering sandstones from Arunachal Pradesh westward to central Nepal at altitudes of from 3,600 to 5,700 feet. Alder trees grow along the watercourses on the steeper slopes. At higher elevations they are succeeded by mountain forests in which the typical evergreen is *Pandanus furcatus,* a type of screw pine. Besides these trees, some 4,000 species of flowering plants, of which 20 are palm, are estimated to occur in the Eastern Himalayas.

Types of vegetation

With the decrease of rainfall and the increase of altitude westward, the rain forests give place to tropical deciduous forests, where the timber tree, sal, is the dominant species, thriving best on high plateaus at 3,000 feet (wet sal), as well as higher up, at 4,500 feet (dry sal). Westward, steppe forest (*i.e.,* forest on an extensive plain), steppe, subtropical thorn steppe, and subtropical, semidesert vegetation occur successively. Temperate forests extend from about 4,500 to about 11,000 feet and contain conifers and broad-leaved temperate trees. Evergreen forests of oaks and conifers have their westernmost outpost on the hills above Murree, some 30 miles northwest of Rāwalpindi, in Pakistan; these are typical of the Lesser Himalayas, being conspicuous on the outer slopes of the Pīr Panjāl, in Kashmir, India. *Pinus roxburghi* (chir pine) is the dominant species at altitudes of from 2,700 to 5,400 feet. In the inner valleys this species may occur even at an altitude of 6,300 feet. Deodar cedar, a highly valued timber tree, is another species particular to the Himalayas, occurring mainly in the western part of the range. Stands of this species occur between 6,300 and 9,000 feet and also tend to grow at still higher altitudes in the upper valleys of the Sutlej and the Ganges. Of the other conifers, blue pine and spruce make their appearance between about 7,300 and 10,000 feet.

The alpine zone begins above the tree line between 10,500 and 11,700 feet and extends as far as 13,700 feet in the Western Himalayas and 14,600 feet in the Eastern Himalayas. In this zone all the wet and moist alpine vegetation is to be found. Juniper is widely distributed, preferring sunny sites, steep and rocky slopes, and drier areas; on Nānga Parbat they are found even at an altitude of 12,750 feet. Rhododendron occurs everywhere but more abundantly in the wetter parts of the Eastern Himalayas, where it grows in all sizes from trees to low scrub. Mosses and lichens grow in shaded areas at lower levels where the humidity is high; flowering plants occur at high altitudes, especially on Nānga Parbat and Mt. Everest.

Alpine plants

Animal life. The animal life of the Eastern Himalayas is derived mainly from that of the South Chinese and Indo-Chinese region. It is primarily the type of animal life found in the tropical forest and is only secondarily adapted to the subtropical, mountain, and temperate conditions prevailing at higher altitudes and in the drier western areas. The animal life of the Western Himalayas, however, has more affinities with that of the Mediterranean, Ethiopian, and Turkmenian regions. The past presence in the region of some African animals, such as the giraffe and the hippopotamus, can be inferred from fossil remains in the Siwalik deposits of the Outer Himalayas. The animal life at higher altitudes above the tree line consists almost exclusively of species, adapted to the cold, that have originated in the area, having evolved from the wild life of the steppes (extensive plains) after the Himalayan uplift. Elephants, bison, and rhinoceroses are restricted to certain

areas of the forested *terai* (moist or marshy lands, now largely drained) at the base of the low hills in the Outer Himalayas. The Indian rhinoceros was once abundant all over the foothill zone of the Himalayas but is now becoming extinct; the musk deer and the Kashmir stag, or hangul, are also on the point of extinction. The Himalayan black bear, the clouded leopard, the langur monkey (a long-tailed Asian monkey), and the cat are some of the other denizens of the Himalayan forests. Himalayan goat antelopes are also found.

Exotic game animals

In higher altitudes above the tree line, the snow leopard, the brown bear, the red panda, and the Tibetan yak can occasionally be seen. The yak has been domesticated and is used as a beast of burden in Ladākh. The typical inhabitants of higher altitudes above the tree line are, however, diverse types of insects, spiders, and mites, which form the only animal life that can live as high up as 20,700 feet.

Catfish of the genus *Glyptothorax* live in most of the Himalayan streams, on the banks of which is found the Himalayan water shrew. Lizards of the genus *Japalura* are widely distributed. *Typhlops,* a genus of blind snake, is common in the Eastern Himalayas. The butterflies of the Himalayas are extremely varied and beautiful, especially the genus *Troides.*

The bird life is equally rich but is more in evidence in the east than in the west. Among some of the common Himalayan birds are different species of magpie—the black-rumped, the blue, and the racket-tailed; titmouse; chough (related to the jackdaw); whistling thrush; and redstart. A few strong fliers, such as the lammergeier (bearded vulture), the black-eared kite, and the Himalayan griffon (an Old World vulture), can also be seen in Sikkim. The snow partridge and the Cornish chough are found at elevations of 18,600 feet.

The people. Of the three principal ethnic types in the Indian subcontinent—the Indo-Aryan, the Mongolian, and the Dravidian—the first two are well represented in the Himalayas, although mixed in varying proportions in different areas. Waves of immigration into the mountains have occurred from all directions in the past and have caused intermingling of peoples. Generally speaking, the Great Himalayas and the Tethys Himalayas are inhabited by Tibetan and other Mongoloid people; the Lesser Himalayas are the home of the tall, fair Indo-Aryans. In the Outer Himalayan region of Jammu and Kashmir the Indo-Aryans are called Dogras and are divided into two main castes, Brahmins and Rājputs. In the Kashmir Valley the same type is represented by the Kashmīrī people. The Gaddis and Gūjars, who live in the hilly areas of the Lesser Himalayas, also belong to the Aryan type. The Gaddis are essentially a hill people; they possess large flocks of sheep and herds of goats and come down with them from their snowy abode in the Outer Himalayas only in winter, returning again to the highest pastures in June. The Gūjars are a migrating, pastoral people living from their herds of sheep, goats, and a few cattle, for which they seek pasture at various altitudes.

The Champa, Ladākhī, Balti, and Dard peoples live to the north of the Great Himalayan Range in the Kashmir Himalayas; the Dard are Aryan, the others Mongoloid. The Champa lead a nomadic pastoral life in the Upper Indus Valley. The Ladākhī have settled on terraces and alluvial fans flanking the Indus in Kashmir. The Balti have spread farther down the Indus Valley and have adopted Islām.

The Aryan racial type is represented by the Kanets in Himachal Pradesh and the Khāsas in Utturakhand. In Himachal Pradesh the majority of the inhabitants of Kinnaur and Lāhul-Spiti districts are Mongoloid, having immigrated from Tibet.

In Nepal, the Tibeto-Nepalese and Indo-Nepalese form the two main ethnic divisions, which are further subdivided into a large number of ethnic groups, including the Newārs, the Tamangs, the Gurungs, the Magars, the Sherpas, and the Kirāts. The Kirāts were the earliest inhabitants of the Nepal Valley. The Newārs are also one of the earliest Nepalese groups. The Tamangs inhabit the high valleys of Ganesh Himal (Nepal, southwest of Himachal Pradesh). The Gurungs live on the southern slopes of the Annapūrna Massif (mountain mass), pasturing their cattle as high as 12,000 feet. The Magars inhabit western Nepal but migrate seasonally to other parts of the country. The Sherpas, who live to the south of Mt. Everest, are famed mountaineers.

The people of Sikkim belong to three distinct ethnic groups—the Lepchās, the Bhutias, and the Nepalese. Generally speaking, Nepalese and Lepchās live in western Bhutan and Bhutias of Tibetan origin in eastern Bhutan. Arunachal Pradesh is the homeland of several groups—the Abors or Adis, Akas, Apa Tanis, Daflas, Khamptis, Khowas, Mishmis, Mombas, Miris, and Singpho. Ethnically, all these groups are Indo-Mongoloid; linguistically, they are Tibeto-Burman. Each group lives in a distinct river valley, practicing shifting cultivation (*i.e.,* constantly changing the land on which they raise crops).

The economy. *Resources.* The Himalayas abound in economic resources. These include rich arable land, extensive grassland and forest, workable mineral deposits, and easily harnessable waterpower. The most productive arable lands in the Western Himalayas are in the Vale of Kashmir, the Kāngra Valley, the Sutlej Basin, and the terraces flanking the Ganges and Yamuna in Utterakhand; these produce rice, corn (maize), wheat, and millet. In the Central Himalayas in Nepal most of the arable land is in the foothills and on the adjacent plains; this land produces four-fifths of the total rice production of the country, amounting to 1 percent of world production. The region also produces large crops of corn and wheat. In addition to cereals, most of the cash crops of the country—jute, sugarcane, and oilseeds—are grown in this region.

Arable lands, minerals, and water resources

Most of the fruit orchards of the Himalayas lie in the Vale of Kashmir and in the Kulu Valley of Himachal Pradesh. Such fruit as apples, peaches, pears, and cherries, for which there is a great demand in the cities of India, are grown extensively. There are rich vineyards on the shores of Dal Lake in Kashmir, which produce grapes of good quality from which wine and brandy are made. On the hills surrounding the Vale of Kashmir grow walnut and almond trees, the nuts of which are exported to India where oil is extracted from them. Bhutan also has fruit orchards and exports oranges to India.

Of the plantation crops, tea is grown mainly on the hills and on the plain at the foot of the mountains in the Darjeeling district. Tea in limited quantity is also grown in the Kāngra Valley. Plantations of cardamom, a spice used in curry, are to be found in Sikkim, Bhutan, and the Darjeeling Hills. Medicinal herbs are grown in plantations in the Uttarkāshi and Pithorāgarh districts of Uttarakhand.

Transhumance (the seasonal migration of livestock) is widely practiced during the summer months in the Himalayas pastures, called *margs,* in Kashmir. Sheep, goats, and yak are raised on the rough grazing lands available.

Woodlands occupy at least one-third of the Himalayas, covering more than two-thirds of the total area of Bhutan and Sikkim. They constitute the greatest asset of the mountains, although their fuller use is hampered because of inaccessibility. Logs of timber are floated down the Himalayan streams to sawmills located at the foot of the mountains. Forest-based industries, to manufacture matches, rayon, and paper pulp, are being established in Bhutan.

The Himalayas are rich in minerals, although their exploitation is restricted to the more accessible areas. Jammu and Kashmir state is the most mineralized region. Sapphires are found in the Zāskār Mountains, and alluvial gold in the nearby bed of the Indus. There are deposits of copper ore in Baltistān, and iron ores are found in the Vale of Kashmir. Ladākh contains borax and sulfur deposits.

Mineral resources

Coal deposits exist in the Jammu Hills. Bauxite occurs in Jammu and Kashmir. Nepal, Bhutan, and Sikkim have extensive deposits of coal, mica, gypsum, and graphite and ores of iron, copper, lead, and zinc.

The Himalayan rivers have a tremendous hydroelectric potential, which has been harnessed more intensively since the five-year plans were introduced in the 1950s in India. A giant multipurpose river-valley project is located at Bhākra-Nāngal on the Sutlej River in the Outer Himalayas; completed in 1963, it has a storage capacity of

348,218,000,000 cubic feet (9,860,000,000 cubic metres) of water and a total installed capacity of 1,050 megawatts. Three other Himalayan rivers, the Kosi, Gandak, and Jaldhāka, have been harnessed by India, which then supplies the power to Nepal and Bhutan.

Transportation. The difficulty of transport in the Himalayas has always constituted a barrier to economic growth. Only in recent years have new highways been built, making the Himalayan region accessible from both the north and the south. Of these, the 75-mile, all-weather Tribhuvan Rajpath road connecting Kāthmāndu, capital of Nepal, with India and the 65-mile road connecting Kāthmāndu with Kodari, on the Tibetan border, deserve special mention. The Hindustān–Tibet road—Indian National Highway No. 22—which passes through Himachal Pradesh has also been considerably improved by the government of India; this 300-mile highway, which runs through Simla, once the summer capital of India, connects the Punjab plains with the Indo-Tibetan border near the Shipki La (pass). There are only two main railroads, both of narrow gauge, penetrating into the Lesser Himalayas from the plains of India—one in the Western Himalayas, between Kālka and Simla, and the other in the Eastern Himalayas, between Siliguri and Darjeeling. There is another narrow-gauge line in Nepal, running 29 miles from Raxaul to Amlekhganj and connected with Kāthmāndu by an electrically operated aerial cableway, which transports cargo in baskets. Two other short railroads run to the Outer Himalayas—one, the railroad of the Kulu Valley, from Pathānkot to Jogindarnagar; the other from Hardwār to Dehra Dūn. A short railway, formerly running between Wazīrābād and Jammu through Siālkot, is now permanently closed.

There are two major airstrips in the Himalayas—one at Gaucher, Kāthmāndu, capital of Nepal, and the other in Srīnagar, capital of Kashmir—which are served by national and (except Srīnagar) international airways. Besides these, there are also many airstrips of local importance in the hills and in the *terai* district of Nepal.

From the Punjab plains the only direct approach to the Vale of Kashmir is by the National Highway No. 1A from Jullundur to Uri through Jammu, Banihāl, Srīnagar, and Bāramūla. It crosses the Pīr Panjāl Mountains through a tunnel at Banihāl. The old road from Rāwalpindi to Srīnagar through Pakistan has lost much of its former importance. Within the Kashmir Himalayas, the Srīnagar to Leh road connects Leh in Ladākh with the Nubra Valley, passing over the 17,730-foot-high Khardung La—the first of the high passes on the historic caravan trail to Central Asia from India. Many other new roads have been built in recent years.

Sikkim commands the historic Kālimpong to Lhasa caravan trade route, which passes through Gangtok. Before 1956, there was only one (30-mile) motorable highway running between Gangtok and Rongphu, on the Tista River near the Sikkim–West Bengal border, which then continued southward to Silīguri for another 70 miles. Since then, several roads passable by jeep have been built in the southern part of Sikkim, and a highway in northern Sikkim connects Gangtok with Lachen (Lachung).

Arunachal Pradesh is connected with the Brahmaputra Valley by roads running from Namsai to Chowkham, Sadiya to Roing, Pāsighāt to Dibrugarh, Along to Sonarighat, North Lakhimpur to Hapoli, and Tezpur to Bomdila.

Study and exploration. The first Himalayan sketch map of some accuracy was drawn up by Father Antonio Monserrate, a Spanish missionary to Akbar's court in 1590. In 1733 a French geographer, Jean-Baptiste Bourguignon d'Arville, compiled the first map of Tibet and the Himalayan range based on systematic exploration. In the middle of the 19th century, the Survey of India organized a systematic program to measure correctly the heights of the Himalayan peaks. The Nepal and Uttarakhand peaks were observed and mapped between 1849 and 1855. Nānga Parbat, as well as the peaks of the Karakoram to the north, were surveyed between 1855 and 1859. The surveyors did not allot individual names to the innumerable peaks observed but designated them by figures and roman numerals. Thus, at first Mt. Everest was simply labelled as "h", this was later changed to Peak XV in 1849–50. In 1865, Peak XV was renamed for Sir George Everest, surveyor-general of India from 1823–1843. Not until 1856 were the computations sufficiently advanced for it to be realized that Peak XV was higher than any other peak in the world. By 1862 more than 40 peaks of 18,288 feet and above had been climbed for surveying purposes.

Early explorations

The Survey of India has prepared some large-scale maps of the Himalayas from aerial photographs. Parts of the Himalayas have also been mapped by German geographers and cartographers with the help of ground photogrammetry. (S.P.C.)

HINDU KUSH

The Hindu Kush is one of the great watersheds of Central Asia, forming part of the vast alpine zone that stretches across the continent from east to west. Broadly defined, it is a mountain system nearly 1,000 miles (1,600 kilometres) long and possibly 200 miles (320 kilometres) wide, running northeast to southwest, and dividing the valley of the Amu Darya (the ancient Oxus River) to the north from the Indus River Valley to the south. There is no agreement among geographers as to its precise boundaries. To the east, the Hindu Kush buttresses the Pamir Plateau near the Chinese border, after which it runs southwest through Pakistan and into Afghanistan, finally merging into minor ranges in western Afghanistan. The highest peak is Tirich Mīr, which rises on the Pakistan–Afghanistan border to 25,263 feet (7,700 metres).

Historically, the passes across the Hindu Kush have been of great military significance, providing access to the northern plains of India. Alexander the Great, king of Macedonia, who passed over the Hindu Kush in the 4th century BC, was among those who invaded India by this route. During the period of British rule in India, the Indian government was keenly concerned with the security of these passes, and more especially with their own control of an associated physical feature to the south, the Khyber Pass. The Hindu Kush range has rarely constituted the frontier between major powers, but has usually formed part of an intermediate buffer zone.

The name Hindu Kush first appears in 1333 in the writings of Ibn Baṭṭūṭah, the medieval Berber traveller, who said that the name meant "Hindu killer," a meaning still given by Afghan mountain dwellers, who are traditional enemies of Indian plainsmen. More likely the name is a corruption of the classical term Hindu-Caucasus, or else Hindu-Koh, meaning "Indian Mountains."

Physical features. The eastern limit of the Hindu Kush is difficult to determine because of a locally complex topography, although the Karambar Pass (14,225 feet, or 4,345 metres) between the valleys of Chitrāl, Pakistan, and Gilgit, Jammu and Kashmir, may be tentatively accepted as marking the boundary. The western limit is still more uncertain, as the mountains lose height and fan out into minor ranges. The Kermū Pass (10,879 feet, or 3,316 metres) to the west of Kābul may, nevertheless, be regarded as indicating the approximate boundary of the range. Geologists, however, consider the Hindu Kush range to extend much further west not only into Afghanistan but also into Iran.

Physiography. Three main sections of the Hindu Kush may be defined. These are the eastern Hindu Kush, which runs from the Karambar Pass in the east to the Dorāh Pass (14,940 feet, or 4,554 metres) not far from Tirich Mīr; the central Hindu Kush, which then continues to the Khāvāk Pass (11,640 feet, or 3,548 metres) to the north of Kābul; and the western Hindu Kush, also known as the Kūh-e-Bābā, which gradually descends to the Kermū Pass.

Three main regions

In its extreme eastern section, between the passes of Karambar and Baroghil (12,480 feet, or 3,804 metres), the eastern Hindu Kush region is not very high and has mountains that often take the form of rounded domes. Further to the west the main ridge rises rapidly to Baba Tangi (21,368 feet, or 6,513 metres) and becomes rugged, after which, within the space of about 100 miles, are concentrated the highest mountains of the entire region—about two dozen summits of more than 23,000 feet, or 7,000 metres, in height. A first cluster of high peaks around

Ūrgand, Afghanistan (23,094 feet, or 7,039 metres) is followed further south by the massif (principal mountain mass) of Saraghrar (24,111 feet, or 7,349 metres). Another line of imposing mountains, which includes Koh-i-Langar (23,162 feet, or 7,060 metres), Shachaur (23,346 feet, or 7,116 metres), Udrem Zom (23,376 feet, or 7,125 metres), and Nāder Shāh (23,376 feet, or 7,125 metres), leads to the three giant mountains of the Hindu Kush, which are Noshaq (24,580 feet), Istro Nal (24,242 feet), and Tirich Mīr (25,263 feet). Most major glaciers of the Hindu Kush—among them Kotgāz, Ushko, Niroghi, Atrak, and Tirich—are in the valleys of this section.

Watershed of the Central Hindu Kush

The central region lies almost entirely within Afghanistan. According to the report of a British expedition in 1967, this region "has no nice, easily definable east–west ridge but rather a tortuous twisting watershed with massive off-shoots running north towards the Soviet Union and south into Nuristan (a region of Afghanistan)." Maximum heights, which are lower than those in the eastern section, include Koh-i-Bandakor (22,451 feet, or 6,843 metres), Koh-i-Mondi (20,498 feet, or 6,248 metres), and Mīr Samīr (19,878 feet, or 6,059 metres). These peaks are surrounded by a host of lesser mountains. Glaciers are poorly developed, but the mountain passes—which include Putsigram (13,450 feet, or 4,100 metres), Verān (15,400 feet, or 4,700 metres), Ram Gol (15,400 feet, or 4,700 metres), and Anjoman (13,850 feet, or 4,225 metres)—are high, thus making transmontane communications difficult.

The mountains of the western region fan out gradually toward the Afghan town of Herāt, near the Iranian border, declining into hills of lesser importance. Communication is easier in this region, as the passes, such as the Shebar Pass (9,800 feet, or 2,987 metres), have long since been crossed by roads.

A wider definition of the limits of the Hindu Kush would lead to the inclusion of a fourth region known as Hindu Rāj in Pakistan. This is formed by a long, winding chain of mountains—with some lofty peaks, such as Darkot (22,447 feet, or 6,842 metres), and Būni Zom (21,499 feet, or 6,553 metres)—which strikes southward from the Lupsuk Peak (18,853 feet, or 5,746 metres) in the eastern region, then continues to the Lawarai Pass (12,100 feet, or 3,700 metres) and beyond to the Kābul River. If this chain were to be considered as part of the Hindu Kush, then the outlying mountains of the Swāt Kohistān region of Pakistan to the south would also form part of the complex. For most purposes of this article, however, the Hindu Rāj and its associated ranges are excluded from consideration.

International boundaries running through the Hindu Kush are primarily those of Pakistan and Afghanistan. The Karambar Pass lies about 40 miles west of the Chinese borders, while to the west the Hindu Kush, strictly considered, approaches the border between Afghanistan and Iran without extending into Iranian territory. Between these extremes the Pakistan–Afghanistan border follows the main watershed of the Hindu Kush throughout its eastern region, from Lupsuk Peak just north of the Karambar Pass to the Dorah Pass just south of Tirich Mīr. Not far from the Dorah Pass the boundary leaves the main watershed and follows minor spurs until it crosses the Kābul River, continuing along the crest of the Spīn Ghar Range toward the south. The Khyber Pass constitutes an important strategic gateway because it cuts through the Spīn Ghar instead of through the Hindu Kush thus offering a comparatively easy route between the valley of the Kābul and the plains of Punjab.

The Khyber Pass

The erratic boundary line is the result of a series of compromises reached at the end of the 19th century between the British and the ruler of Afghanistan; called the Durand Line, after the British negotiator, it has been inherited by the modern states of Pakistan and Afghanistan. Another curious configuration established about the same time and as yet unchanged is the Vākhān region (Wakhan Corridor), a panhandle of Afghan territory designed to act as a buffer between British India and tsarist Russia.

Geology. In many of its features the Hindu Kush resembles its eastern neighbour, the Karakoram Range that extends from Tibet into Pakistan. The Hindu Kush, which some authorities consider to be a continuation of the Karakoram, has the same core of igneous metamorphic rock (*i.e.,* rock formed by heat and pressure that has solidified from a molten state) flanked, especially toward the north, by sedimentary material. In the Hindu Kush, however, there appears to be a greater prevalence of sedimentary rocks—a fact that may explain the softer and more rounded forms of many of the mountains. In the Afghan section the core of the chain is formed by a complex sequence of metamorphic rocks containing marble, together with intrusions of granodiorites (rocks formed deep down by heat and pressure, and containing a certain admixture of both dark-coloured and light-coloured minerals). In Nūrestān, Afghanistan, the hills consist mainly of schists (medium- or coarse-grained metamorphic rocks), gneiss (a coarse-grained rock in which bands containing granular materials alternate with bands of schistose materials), and intrusions of granite with zones of migmatite (rock consisting of alternate layers of granite and schist). The Hindu Kush differs markedly from the Karakoram, however, in the winding direction of its strike (*i.e.,* its course, or bearing).

Drainage. The Eastern Hindu Kush appears to be formed of two parallel chains, consisting of a lower one to the north, which acts as a watershed, and a higher southern one that carries the main peaks. Drainage is comparatively simple on the northern side but highly complex on the southern one, where valleys follow two contrasting directions—northeast to southwest and roughly east to west. Most of the rivers, such as the Panjsher, the Alīngar, the Konar, and the Pānjkora, follow the northeast to southwest direction and are then suddenly deflected toward the east–west axis by the Kābul River, into which they flow. The Karambar and Ghizar valleys also take the same east to west direction. The Indus River, however, swerves in its course from one direction to another as it makes its roundabout descent toward the lower plains.

Climate. Since the range separates one important zone of Asia from another, the climate shows great variations. The mountains of Swāt Kohistān are virtually within the area of the rain-bearing monsoon winds, and most of the Eastern Hindu Kush, as well as the Hindu Rāj, rises up at the extreme limit of monsoonal Asia. The Central and Western Hindu Kush, however, border the Mediterranean climatic zone. Thus, moving from the southeast to the northwest and west, one moves from a region of rainy or snowy summer (from July to September) and of dry winters into a region of hot dry summer and cold and rainy or snowy winter (from December to early March). Climatic variations between these opposites also occur, producing often striking local contrasts.

Climatic contrasts

A graphic image of climatic conditions is presented by the glaciers. The mantle of snow and ice is heaviest at the extreme eastern end of the Hindu Kush in Pakistan, where the Chiāntar Glacier is situated, and is also heavy in the higher section around Tirich Mīr and Saraghara and in parts of the Hindu Rāj. Toward the west, however, glaciation is more sporadic. In the Central Hindu Kush, mountains 12,000 feet high are often bare almost to the summit. Most of the glaciers of the Hindu Kush appear to be retreating. A striking feature of some glacial regions are the so-called nieves penitentes, which are protruding spikes of frozen snow forming the illusion of kneeling human figures, sometimes two or three feet high, which are especially noticeable in the early morning; they are caused by the alternation of fierce sun and rapid evaporation during the day and of severe cold at night.

Plant life. Differences in latitude and the variety of climates make it difficult to generalize about vegetation. Compared to the Himalayas of Nepal, and still more to those of Sikkim and Bhutan, the mountains of the Hindu Kush appear bare, stony, and poor in vegetation, although there are local exceptions. Some rich forests and pastures are found in the extreme southeast of the extended Hindu Kush region, as well as in the hills of Swāt Kohistān and in the Pānjkora Valley of the Dīr District, Pakistan. Parts of the valleys of the Gupis and Yāsīn rivers in the Gilgit area enjoy sufficient summer precipitation to be partially covered with vegetation. In the valleys of the Swāt and Dīr districts, as well as in some parts of the Chitrāl area, rice is cultivated.

The highlands of the eastern extremity of the Hindu Kush, with their rolling pastures, bear some resemblance to the plateau landscape of Tibet, but further south the valleys become arid and stony. A typical view in parts of Chitrāl, for example, would include snowy peaks in the distance, dry, barren, brick-red or ochre-coloured mountains all around, and bright emerald-green islands of vegetation near the villages where springs and irrigation furnish abundant water. In such oases, poplar trees are a distinct feature, often being accompanied by old and gigantic plane trees.

Much moisture carried by the monsoon winds penetrates the lower part of Chitrāl across the Lawarai Pass, so that stands of coniferous trees occur in the surrounding districts. The valleys of Nūrestān receive sufficient precipitation to have some pastureland and forests.

To the west of Afghanistan, and to the north of the Hindu Kush, extremely dry summers prevail; much of the country is stony, or sparsely covered with thorny and spiny plants, or with poor grass.

Animal life. The meagre vegetation of most parts of the Hindu Kush does not favour an abundance of wild animal life. The snow leopard manages barely to survive in the most remote valleys. Bears formerly roamed Nūrestān, but few remain today. The markhor (a kind of wild goat) was once abundant, but hunters have now thinned the stock. Birdlife is rich, however, and eagles are occasionally to be seen.

The people. A long and tormented history, together with fragmented topography, has produced a veritable mosaic of ethnic units in the region. Kirgiz nomads, about 30,000 in number, graze their herds on the Vākhān uplands. The lower parts of the Vākhān and the higher parts of the Sang Lēch and Anjoman valleys, all on the northwest slopes of the Hindu Kush, are sparsely inhabited by the so-called Pamir Tadzhiks, most of whom are Shī'ah Muslims. Other Tadzhiks (who are Sunnī Muslims), Uzbeks, and some Ḥazāra (Persian-speaking Mongols) live in the valleys of the central and western parts of the Hindu Kush. Afghans are found in the major towns, in Kābul, and in many districts to the south of the Hindu Kush, with the exception of Nūrestān.

On the southeast (Pakistan) side of the Hindu Kush, most people are Chitrāli, a racially mixed ethnic group that shows a marked cultural unity.

The Kafirs of Nūrestān

The Kafirs of Nūrestān and of Chitrāl are an exceptionally interesting people. Their name means "infidel" or "non-Muslim" and seems to have been used since the 11th century. Traditionally, they are divided into two groups—the *kalash* ("black") Kafirs of Chitrāl, and the *kati* ("red") Kafirs of Nūrestān. In the remote past, the Kafirs possibly inhabited a much larger area. The Kafirs of Nūrestān were forcibly converted to Islām in 1896.

Physically, the Kafirs do not seem to differ much from their neighbours; they speak a language classed by some as Dardic. It is in their religion and culture that their ethnic individuality is most strikingly expressed. In religion they practise a form of polytheism; worship consists mainly in the sacrifice of animals. Dancing is important, and divination through shamans is practiced. The dead are disposed of, unburied, in heavy wooden coffins. Large wooden statues of ancestors, often on horseback, stand near graveyards; many are works of vigorous and elemental beauty. Kafir homes are strong rectangular wooden buildings. Their economy is based on agriculture, hunting, and the raising of goats and oxen.

The economy. *Resources.* The economic resources of the region remain virtually undeveloped. It is thought that the northern slopes may contain coal and perhaps petroleum or natural gas; the presence of gold, silver, copper, zinc, and lead also is suspected. Quantities of salt are known to exist, and lapis lazuli (a deep blue mineral, used as a gem), rubies, and beryl (a mineral, usually green, one variety of which is emerald) have been reported. There is some antimony (a white metallic element used in alloys of medicine) in Chitrāl. There is a hydroelectric potential, especially in the Daryā-ye Qondūz and Kowkcheh river basins on the northern side. Forests are still partially standing, but much merciless lumbering is rapidly reducing this resource. While many local peoples, especially the Chitrāli, make ingenious use of springs and rivers by building small aqueducts, agriculture would benefit greatly from more modern methods of irrigation. The local economy is at the bare subsistence level. Some dried fruit, a little lumber, mineral salt, charcoal, fodder grass, mats, and ropes are exported, mainly from Chitrāl. Sheep and goats are also raised.

Transportation. The Hindu Kush offers a formidable barrier to communication. There are, however, some important passes. The Baroghil Pass (12,480 feet, or 3,804 metres high), at the head of the Chitrāl Valley, is one of the lowest and easiest openings in the 1,500 miles of forbidding mountains that border the north of India and Pakistan. Further west the highest section of the Hindu Kush offers, for about 100 miles, a wall quite unpassable by normal means of communication. In the Central Hindu Kush the passes are also high; only in the western section do more accessible passes occur. In 1964 a tunnel was completed under the Sālang Pass (12,008 feet, or 3,660 metres) north of Kābul; consequently, for the first time in history, the great mountain wall has been pierced, making the north of Afghanistan accessible to the south at all seasons.

The Kakal Tunnel

The Hindu Kush can now be approached by motor transport from many directions. Chitrāl in the south is accessible, via the Lawarai Pass from Peshāwar, while the Kakal Tunnel has made Khānābād in the north accessible from Kābul. Lesser roads lead on to Feyzābād and, from there, to the Vākhān corridor, or to Zībāk, both situated in the heart of the mountain ranges. Many major peaks are now barely three or four days march from the last village accessible to motor transport.

Study and exploration. The West took note of the Hindu Kush when Alexander and his armies crossed the mountains, which according to some authorities, he did twice. The Hindu Kush was well known to Arab geographers, as well as to the Chinese, who occupied Chitrāl in the 8th century AD. Marco Polo, the Venetian traveller, and his group passed along the Hindu Kush through the Vākhān region. The mountains were also traversed by Timur (Tamerlane), the Mongol conqueror, in the 14th century and by Bābur, the Turkish conqueror who founded the Mughal empire, in the 15th century, in their expeditions against India. A number of explorers visited the region in the 19th century, and much knowledge was gained by the British during the two wars that they waged against Afghanistan from 1838 to 1842 and from 1878 to 1879. Further detailed knowledge of the area has been gained in the 20th century. In the second half of the 20th century hundreds of mountaineering and exploring expeditions have visited the Hindu Kush. (F.M.)

KARAKORAM RANGE

A mountain system of Central Asia, the Karakoram Range extends some 300 miles from the easternmost part of Afghanistan to the southeast. The borders of the Soviet Union, China, Pakistan, Afghanistan, and India all converge within the system.

Physical features. *Physiography.* The Karakorams consist of a group of parallel ranges with several spurs. Only the central part is a monolithic range. The width of the system is about 150 miles; the length is increased from 300 to 500 miles, if the easternmost extension—the Ch'iang-ch'en-mo Shan (called Chāng Chenmo in Jammu and Kashmir) and Pantong ranges of the Tibetan Highlands—is included. The Karakoram Range is one of the highest mountain systems in the world; its average height is around 20,000 feet (6,100 metres), and four peaks exceed 26,000 feet, the highest being K2 (Godwin Austen, Chogori, Dapsang) at 28,250 feet (8,611 metres). K2, the second highest peak in the world, was first climbed by an Italian expedition in 1954; the remaining high peaks were climbed in 1956–58.

Glaciation

The topography is characterized by craggy peaks and steep slopes. The southern slope is long and very steep, the northern slope steep and short. Cliffs and taluses (great accumulations of large, fallen rocks) occupy a vast area. In the intermontane valleys, rocky inclines occur widely. Transverse valleys usually have the appearance of

narrow, deep, steep ravines. Because of their great height, the Karakorams are characterized by heavy glaciation, the total glaciated area amounting to 6,900 square miles (17,800 square kilometres). Glaciers occur on both slopes, but glaciation is more developed on the southern, more humid slope. The snow line on the southern slope of the Karakorams begins at an altitude of 15,400 feet; glaciers begin at 9,440 feet. On the northern slope the figures are 19,400 feet and 11,580 feet, respectively. Often, glaciers combine to form complex glacial systems occupying not only valleys but also the watersheds. Seasonal thawing of the glaciers give rise to serious floods on the southern slopes. Traces of ancient glaciation are evident at altitudes of 8,500 to 9,500 feet.

The Karakorams serve as a watershed for the basins of the Indus and Tarim rivers. The formation of river channels, for the most part, occurs in the high-altitude zone, the melted waters of seasonal and perpetual snows and glaciers being principal feeders of the rivers. Groundwaters accumulate in the rocky taluses and contribute to a more even flow throughout the year. During winter, huge layers of ice are formed.

Geology. Structurally, the Karakorams originated from folding in the Cenozoic Era (up to 65,000,000 years ago). Granites, gneisses, crystallized slates, and phyllites dominate the geological composition. To the south and north, the crystal centre of the Karakorams is edged by a region of limestones and micaceous slates of the Paleozoic and partly of the Mesozoic eras (from 190,000,000 to 570,000,000 years old). To the south the sedimentary rock is sometimes cut by intrusions of granite. Certain areas expose slate at the surface, which yields more rapidly to weathering.

At the end of the Mesozoic Era (65,000,000 years ago), the region of Karakorams was characterized by great structural changes, and the Karakorams emerged as the result of intensive, geologically recent upheavals. Today there is still frequent seismic activity in the region, some events being of great violence. Hot springs are found in several areas.

There is also evidence of stannic tungsten on the northern slopes of the Karakorams and of alluvial gold on the southern slopes.

Climate. The climate of the Karakoram Range is for the most part semi-arid and sharply continental. The southern slopes are exposed to the humidifying influence of the monsoons coming in from the Indian Ocean, but the northern slopes are extremely dry. In the lower and central part of the slopes, rain and snow is precipitated in small quantities; average annual precipitation does not exceed four inches. At altitudes of more than 16,000 feet precipitation always takes a solid form, but, even lower down, in June, snow is not infrequent. At altitudes of around 18,700 feet, the average temperature during the warmest month is lower than 32° F (0° C), and, at altitudes between 12,800 and 18,700 feet, the temperature is lower than 50° F (10° C). Rarified air, intensive solar radiation, strong winds, and great ranges of temperature are characteristic climatic features of the region.

High-altitude vegetation patterns

Plant and animal life. The high-altitude vegetation of the northern and southern slopes of the Karakorams is varied. On the northern slope, at altitudes of 7,900 to 9,200 feet on the rocky desert soil, a complex of plant combinations of such species as flowering plants of the genus *Kalidium,* and horsetail (genus *Equisetum*) has developed. In this area it is not uncommon to encounter vast expanses completely devoid of vegetation. Only at the source of the Yarkand River (Tarim Basin) and its tributaries up to altitudes of 10,000 feet is there enough moisture to support individual thickets of brushwood (mainly barberry) and poplars. In the central part of the northern slope, at altitudes of 8,500–10,200 feet, a desert-steppe landscape is developed, with vegetation consisting of sparse thickets of coarse grasses and wintergreen. At altitudes of 10,500–11,500 feet, a mountain steppe predominates, and, in places that are most humid and well sheltered from winds, there are prairie steppes. Still higher up are found high-altitude expanses of wintergreen, wormwood, and meadows of desert-like plants. Sparse combinations of wintergreen and prickly herbs of the genus *Acanthus* are located on the coarse soils near the arable-land zone. On the moister southern slopes, more extensive and varied vegetation is found. Valleys up to 10,000–11,500 feet support forests of pine, Himalayan cedar, and, near streams, willows and poplars. Higher up, high-altitude steppes rather like typical alpine meadows predominate.

Notable animals of the region include the snow leopard, wild yak, and Tibetan antelope; in the southern foothills wild asses are also found. There are a great number of pikas and marmots. Among birds, the Pallas sand grouse, Tibetan capercaillie, partridge, ibis white dove, and red brambling are characteristic.

The people and economy. Being centred in an area of very high ranges covered with immense glaciers, the Karakorams are extremely inaccessible. Mountain passes are situated at altitudes of about 16,000 feet and are open only five or six months of the year. Sections along the banks of rivers and lakes are utilized as pasture, and in places, on the southern slopes, agriculture is developed, with two harvests in years when water is plentiful. Apricot orchards are extensive.

Exceptionally severe natural conditions in the Karakorams make life hard for man. An area of almost 80,000 square miles supports a population of only a few tens of thousands of people, mainly Tibetans, who live in villages at altitudes of up to 14,800 feet. Most of the Tibetans are farmers who grow barley, oats, and millet and who breed cattle. Individual groups of Tibetans lead a nomadic or seminomadic way of life and are occupied with the breeding of yaks, *dso* (a hybrid of the yak and common cow), sheep, and goats; they also do some hunting. Other peoples include the Baltis, who are Muslim, and the Ladākhis, who, like the Tibetans, are Tibetan Buddhist. Also there are the Burishki people, who speak a language that occupies a singular place in linguistic classification. The Burishki are Muslims and are settled on farms. (G.D.B.)

KUNLUN MOUNTAINS

The Kunlun Mountains (K'un-lun Shan), which extend approximately 1,675 miles (2,680 kilometres) from the Pamirs in the Soviet Union on the west to the Sino-Tibetan ranges on the east, constitute the longest mountain system in Asia, uniting dozens of ranges that are among the highest on Earth. Located in the People's Republic of China, within the autonomous regions of Sinkiang Uighur and Tibet and the province of Tsinghai, the Kunlun Mountains form the northern wing of the geologically uplifted region known as High Asia—the highest such region in the world.

The position of the Kunlun Mountains, between Tibet and the northern plains of Central Asia, determines the sharply asymmetrical structure of the system. Although the average elevation of the watershed ridges in the southern ranges is about 21,325 feet (6,500 metres), when viewed from the south the ranges rise only from about 3,300 to 4,900 feet above the Plateau of Tibet, which itself has an average height of between about 14,500 and 16,500 feet. As seen from the Tarim and Ala Shan plains to the north, however, which have an average altitude of only about 2,600 to 3,900 feet, the watershed ridges of the northern ranges (the average elevation of which is about 19,650 feet) create the impression of gigantic mountains that tower up to 14,765 feet over the surrounding plains.

General alignment of the mountains

Physical features. *Physiography.* General alignment is nearly latitudinal (from east to west), but individual segments change direction significantly, following the outlines of the Tarim Basin, the Ala Shan Desert, and the Tibetan massifs (mountainous masses). The most significant such deviations occur in the Khotan–Keriya (Ho-t'ien–Yü-t'ien) sector (about 81° E), which faces the Lop Nor plain (about 90° E). Here, the strike of the mountains shifts abruptly from northwest–southeast to southwest–northeast, before again resuming its northwest–southeast alignment.

Almost the full length of the Kunlun Mountains consists of parallel chains of ranges, the tallest of which are those closest to Tibet, separated by vast depressions and narrow

valleys. Because they are higher in relation to the surrounding plains, the northern slopes (the Tarim and Ala Shan) are steeper and more complexly dissected, while the southern slopes (those facing Tibet) are shorter, sometimes taking the form of ledges that are only weakly dissected. Whereas the heights of the major Kunlun peaks are several thousand feet shorter than the famous Himalayan peaks, the average height of the Kunlun and Himalayan ridges are almost equal. The elevations of the Kunlun mountain passes (from 18,700 to 20,340 feet), however, surpass those of the deeply eroded Himalayas.

The Kunlun system is not uniform in structure, being subdivided into two unequal sections—the smaller western and the principal eastern part. The western Kunluns (between the Pamirs and the alignment change in the Khotan–Keriya region) form three parallel chains of ranges crowded closely together. The Sha-li-k'o-erh, T'a-shih-k'u-erh-kan (Tashkurgantag), Agyl, and Sugettag ranges form the southern chain, which adjoins the Karakoram Range; the inner and highest chain is formed by the Mu-ssu-t'a-ko-a-t'e, Tokhtakoram, and Karangutag ranges; the Tiznaf and Sandzhutag ranges make up the northern chain. Because the mountain chains of the western Kunluns are divided only by narrow intermontane depressions, the width of this part of the system usually does not exceed about 60 miles.

The highest groups of peaks in the western Kunluns are found on its flanks—in the Gissar sector, the 24,865-foot (7,579-metre) Kung-ko-erh and the 24,757-foot (7,546-metre) Mu-ssu-t'a-ko-a-t'e massifs; in the Khotan sector, the 23,008-foot (7,013-metre) Karangutag and the 23,891-foot (7,282-metre) Mu-ssu massifs. In the intervening So-ch'e–Yarkand sector, however, all the ranges are lower; even the main peaks rarely reach 19,650 feet. Where the alignment changes in the Khotan–Kariya region, the northern chain of the Kunluns is interrupted, and ranges of the high inner chain border the Tarim Basin, on the side of which they are abruptly bounded by a gigantic ledge.

The eastern Kunluns are characterized by a complex branching of mountain chains that pass around broad intermontane valleys. The unitary direction characteristic of the western Kunlun ranges is lost here; individual ridges are often situated at an angle to one another. Moreover, the width of the mountain system increases sharply, reaching about 375 miles in places.

The Russian, the A-erh-chin Shan-mo (Astin Tagh), and the group of Nan Shan ranges form the northern chain of the eastern Kunluns, which many consider an independent mountain system. Near the Lop Nor plain these ranges describe an arclike curve in which their alignment changes from southwest–northeast to northwest–southeast. The western branch of the arc (with northeastern alignment) is made up of two ranges with a width of from 19 to 25 miles. The Russian Range, whose crest rises to 21,738 feet (6,626 metres), is the only one of noteworthy height within this branch; A-erh-chin Shan-mo (Astin Tagh) is as much as 11,500 to 13,000 feet lower, especially in the segment contiguous with the Tsaidam Basin.

The Nan Shan ranges

The eastern branch of the arc, which is formed by the system of Nan Shan ranges,is separated from the principal mountain chains of the eastern Kunluns by the vast Tsaidam depression. The five to seven ranges that constitute the Nan Shan system, which has an overall width of up to 186 miles, include Ch'i-lien Shan-mo (Richthofen), T'o-lai Shan (Khrebet Tkholo-shan'), Su-lo Shan (Suess), Wu-lan-ta-pan Shan (Humboldt), Ta-k'en-ta-fan Shan (Ritter), and southern Koko Nor. While the height of their crests varies from about 16,700 to 20,700 feet, the elevation of the longitudinal (north–south) intermontane valley floors between them varies from 10,500 to 11,800 feet. The Su-lo Shan, which rises to 20,820 feet (6,346 metres), is the highest of these ranges; but the most developed, as a mountain range, is the Ch'i-lien Shan-mo.

The principal chains of the eastern Kunluns are located between the Tsaidam Basin and the Tibetan uplands. The northernmost of these, the Ch'i-man–Burkhanbudda chain, is the lowest; the level of its series of summits is from about 17,400 to 18,000 feet. It rises massively above the Tsaidam Basin but stands out in only mild relief when seen from the Tibetan side. Forming an extension of the inner chain of the western Kunlun ranges is the Przhevalsky Range (Arkatag), the highest and longest in the eastern Kunluns and the principal structural pivot of this part of the mountain system. It contains the tallest peaks of the entire Kunlun Mountains, the 25,341-foot (7,724-metre) Wu-lu-k'o-mu-shih Ling and the 25,328-foot (7,720-metre) T'ieh-k'o-li-k'o Shan (also called the Chong-Karlyk-Tag [Great Snow Range] or Shapka Monomakha [Hat of Monomakh]). The crest of the range is snow covered and steep walled, rising vertically almost 3,300 feet; its mountain passes (K'a-la-mu-lun Shan-k'ou, Rekviem, and others) have elevations of close to 18,000 feet. The southern chain, formed by the K'u-k'u-shih-li (K'o-k'o-hsi-li) and Pa-yen-k'a-la Shan (Bayan' Karashan') ranges, is about 3,300 feet lower than the Arkatag.

The high plains separating the inner mountain chains of the eastern Kunluns from those to the north and south lie at the same elevation as the plateaus in the foothills of the Tibetan uplands and have a similar landscape. The most extensive of these is the Kul'tala Plain, which lies between the Przhevalsky Range and the Ch'i-man chain and which is up to 59 miles wide. Covering its surface next to the mountains are fields of broken stone; in areas of lower elevation there are salt marshes, as well as the A-ya-ko-k'u-mu Hu (Aiag Kum Kul') and A-tz'u-k'o Hu (Achchik Köl) lakes.

Geology. The principal folded structures and granitic rocks of the Kunluns are of Hercynian age, a period that lasted from about 230,000,000 to 250,000,000 years ago, during which there was much mountain building in the Eastern Hemisphere. The inner depressions of the Kunluns, however, are relatively recent structures in their entirety, being formed by deposits that are no more than 26,000,000 years old; only the largest of them, the Tsaidam depression, contains a thick sedimentary cover of which Jurassic deposits (136,000,000 to 190,000,000 years of age) represent the oldest strata. The Kunluns also represent a region of very recent movements of the Earth's crust, and great seismic (earthquake) activity.

Drainage. The Kunluns form a part of that region in Central Asia in which there is only internal drainage, associated mainly with the Tarim, Ala Shan, Tsaidam, and Kul'talin basins. Only the most easterly spurs of the mountain system, where the sources of the Huang (Yellow) and Yangtze rivers are located, have drainage systems that empty into the ocean.

Two groups of rivers compose the river network of the Kunluns: the large streams that rise in the Karakoram Range and in northern Tibet, cutting through the entire chain of Kunlun ranges by way of gorges, and the small streams that drain the slopes of the peripheral ranges.

Major rivers

The major rivers—the Gez, Yarkand, K'a-la-k'a-shih (Kara Kash), Yü-lung-k'a-shih (Yurang Kash), K'o-li-ya (Keriya), Ha-la-mu-lan (Kara Muran), and Ch'e-erh-ch'en (Cherchen)—form lengthy, zigzag valleys. In the broad and open longitudinal valleys between mountain ranges, the rivers flow quietly and calmly for long distances; in the ravines that bisect the ranges, however, short sections of the rivers flow through narrow gorges in violent torrents.

Although they receive some rainwaters, the Kunlun rivers are fed mainly by snows and glaciers. Therefore, the volume of flow varies with the seasons; 60 to 80 percent of it occurs in the summer months, when intensive thawing of snow and ice in the mountains is combined with maximum precipitation. In winter, the discharge of the rivers is extremely insignificant; in spring and autumn it is somewhat greater.

Glaciation. In spite of the great elevation, there is little glaciation in the Kunluns because of the extreme dryness of the climate; external snows persist only along the deep crevices of the highest peaks.

The main centres of glaciation, in which are found dozens of glaciers that are usually not more than six miles long, are the Kung-ko-erh, Mu-ssu-t'a-ko-a-t'e, Wu-lu-k'o-mu-shih, and Tyumenlik massifs, where elevations approximate about 23,000 feet. All the glaciers are notable for their unusual steepness, dropping 20 to 30 feet in every 100 feet of length.

Soils. Because the surface layer of the Kunluns does not receive moisture during a large part of the year, soil formation proceeds at a very slow rate. In general, the soils have little productive capacity, contain many coarse skeletal soil elements, and are almost devoid of humus. Brown soils appear only in the vicinity of the Pamirs and in the eastern Kunluns, where the moisture levels are better and vegetation richer.

Aridity and extremes of temperature

Climate. Located within the arid region of Central Asia, the Kunluns are almost totally isolated from the climatic influence of the Indian and Pacific Ocean monsoons. Instead, they are under the constant influence of the continental air mass, which causes great annual and daily temperature fluctuations. Desert yields to mountain steppes only near the Pamirs (King Ata Tagh) and the Tibetan mountains (Pa-yen-k'a-la Shan [Bayan' Karashan'] and the eastern Nan Shans), where the amount of annual precipitation increases to about 18 inches (455 millimetres). In these areas, up to 80 percent of the precipitation falls in summer—an indication of its relationship with the monsoons.

In the high-altitude zone, with its extremely sharp daily fluctuations of temperature, weathering from heat and frost reaches great intensity, accounting for the presence of an enormous quantity of loose material. Wind erosion and accumulations of loose debris are very apparent on the Gobi slope of the peripheral ranges. Maximum aridity occurs in the middle segment of the mountain system, between 78° and 93° E; to the west and east, however, the climate is somewhat moderated. Characteristic of the Kunluns as well as of the entire arid region of Central Asia are the winds; the strongest of which occur in autumn, when they often reach gale intensity.

Scanty cloud cover and prolonged sunshine are also characteristic of the Kunluns. Naturally, the amount of warmth in the Kunluns varies with the altitude: in the lower tier of mountains, (those bordering the northern plains), the average temperature in June is 77° to 82° F (25° to 28° C) and not lower than 16° F (−9° C) in January; in the upper tier of mountains and on the border of Tibet, however, the average temperature in July is less than 50° F (10° C) and often falls to −31° F (−35° C) in winter. Summers in the lower belt are long and hot; in the upper belt, they are short and cold.

In the most arid part of the Kunluns, precipitation is less than two inches annually in the foothills and about four to five inches in the high altitudes. The climate of the lower tier is that of the hot, Gobi-type desert; of the upper belt, it is that of the cold, Tibetan-type desert. Associated with these, in turn, are accumulations of loess—fine, wind-blown material—that forms a cover on the dissected mountain relief and is found at altitudes up to about 13,000 feet.

Soil characteristics

Plant life. Because of the primitiveness of the soil cover, the extreme deficiency of moisture, and, in the high altitudes, the insufficient warmth, the natural conditions of the Kunluns are not very favourable for plant growth. Plants have a stunted appearance and possess a number of distinctive physiological peculiarities. Most often they are perennial, dwarfish semi-shrubs with stiff leaves and deeply penetrating root systems. The number of species of plants is very limited, and the plant cover itself extremely thin.

In the most arid part of the Kunluns, desert-like conditions prevail at high altitudes. Only two types of landscape are to be seen—the hot desert of the lower part of the mountains and the cold desert of the Alpine region. In both zones, completely barren spaces predominate, alternating with small areas that support a thin plant cover.

In the western and eastern extremities of the mountain system, however, the higher moisture levels and richer vegetation are more conducive to the formation of more fertile soil. In these areas, the vertical zonality of the plant cover is more complex and includes zones that do not occur in those parts of the Kunluns that consist entirely of desert. In the middle-altitude region there is a zone of desert-like steppe (a grassy and almost treeless plain) in which there is a somewhat thicker plant cover as one proceeds upwards. Mountain steppes then emerge, as well as scattered forests, small at first but then larger. A well-developed forest zone, interspersed with mountain steppes and meadow sections, appears in the extreme west and the extreme east of the Kunluns.

Animal life. In the totally arid part of the Kunluns, animal life is meagre and has little variety; it becomes somewhat richer and more varied in the approaches to the Pamirs and in the eastern spurs of the system, where mountain steppe and forest vegetation appears. The predominant animal life in these areas consists of hoofed animals and rodents. Especially characteristic of hoofed animals are the mountain sheep (in the west the *arkhar* and in the east the *kukuyaman* type); the mountain goat; the wild ass; and the wild yak. The rodents are represented by mouse hare, field vole, and, on the meadowland slopes, marmot. Beasts of prey commonly found include wolf, fox, bear, and, in the west, snow leopard.

The people and economy. Most of the Kunluns is unpopulated; only the large river valleys, up to an altitude of about 10,000 feet, contain any inhabitants. Settlements become fewer deep in the mountains and nonexistent in the high-altitude zone and on the Tibetan slope. The Uighurs, the most numerous population group, are concentrated in large settlements in the foothills bordering the Tarim Basin; Tadzhiks live in the western and Mongols in the eastern mountains. Agriculture and small-scale, nomadic animal husbandry are the basic occupations of the mountain population. The main agricultural crops are wheat and barley; domestic animals that are bred include sheep, goats, and yaks.

Throughout their extent the Kunluns are almost impassable, because of their deep and narrow ravines, steep slopes, mountain passes of great height, and torrential streams. In addition there is an absence of fuel facilities for vehicles and of forage for pack animals. Even the caravan trails are rare and difficult, especially in the highly eroded middle belt of the mountains.

Of the three automobile roads through the Kunluns, one proceeds along the southern edge of the Tarim Basin, partially using the ancient Silk Road that until the 16th century connected China with Central and western Asia. The other two roads, both of which lead into Tibet, cut across the Kunluns: the western road passes along the T'i-shih-yuëh-fu (Tiznaf), Yarkand (Raskem), and K'a-la-k'a-shih (Kara Kash) river valleys and the eastern road reaches Tibet by a route that extends from Ko-erh-mu (Golmo; in the Tsaidam Basin) to the upper reaches of the Yangtze River and then runs through the Ch'i-man Shan, the eastern spur of the Przhevalsky (Marco Polo Range) and the K'o-k'o-hsi-li (Kukushili).

Because the interior of the Kunluns has not been studied extensively, information about the mineral resources of this vast and geologically heterogeneous territory is extremely scant. Evidences of gold-bearing deposits have been found in various parts of the mountain system, and gold was extracted until the beginning of the 20th century, principally along the upper course of the Yu-lung-k'a-shih Ho (Yurung Kash). Nephrite (a type of jade) is widely distributed and has been extracted from rock deposits and river alluvium. There are well-known rock deposits of this mineral in the Karangutag Range, while alluvial deposits are found in the river valleys of the K'a-la-k'a-shih and Yulung-k'a-shih Ho (Yurung Kash). Small deposits of lead and zinc have been discovered at a number of points in the northern chain of the western Kunluns, while in the K'a-la-k'a-shih Ho Valley, above Sai-t'un-la (Shakhidulla), indications of tin-bearing deposits have been found. Deposits of iron and chromitic ores and evidence of copper mineralization have been located in the Nan Shan and A-erh-chin Shan-mo (Astin Tagh) ranges. Oil is recovered in the Tsaidam Basin, and there are extensive coal-bearing deposits in the intermontane depressions of the western and eastern Kunluns. Although coal is mined in many of these areas, it is on a small scale and solely for local needs.

The exploitation of the mineral resources of the Kunluns is hampered by the harsh natural environment, the great height of the mountains, difficulties of access, lack of water supplies, and scarcity of population.

(V.M.S.)

PAMIRS

The Pamir mountain area belongs mainly to the Gorno-Badakhshan *avtonomnaya oblast* (autonomous region) of the Tadzhik Soviet Socialist Republic of the U.S.S.R. It is surrounded by the Central Asian mountain systems of the Hindu Kush, Kunlun, Gissar-Alai, and Tien Shan.

The territory of the Pamir mountain area is bounded on the north by the Trans Alai Range (Zaalaysky Khrebet); on the east by the Sarykol Range, which forms the border between China and the U.S.S.R.; on the south by Lake Zorkul, the Pamir River, and the source of the Pyandzh River bordering Afghanistan; and on the west by the north–south segment of the Pyandzh Valley. The eastern parts of the Peter I and Darvaz ranges on the northwest are included within the Pamirs, since their glaciers comprise a single system with the glaciers of the northwestern Pamirs. The derivation of the word Pamir has not been definitely established.

Physical features. *Physiography.* The Pamirs are a combination of east–west and north–south ranges, with the former predominating. The east–west Trans Alai Range, which forms the northern frame of the Pamirs, falls steeply to the intermontane Alai Valley. The high central part of the range, between the Tersagar Pass on the west and Kyzylart on the east, averages between 19,000 and 20,000 feet (5,790–6,095 metres), reaching its highest point at Lenin Peak, 23,400 feet (7,134 metres). South from the Trans Alai extend three north–south ranges. Of these, the western, the Academy of Sciences Range, and the central, Zulumart, are relatively short; and the eastern, Sarykol Range, forms the eastern border of the Pamirs. The area east of the Sarykol Range is sometimes called the Chinese Pamirs.

Principal mountain ranges

The north–south Academy of Sciences Range enters into the northwestern Pamir system, where it rises into a huge barrier, reaching 24,584 feet (7,495 metres) in Communism Peak (the highest point in the U.S.S.R.). The eastern slope of the Academy of Sciences Range is covered on the south face by the Fedchenko Glacier. The western slope intersects other ranges that lie still farther to the west: the Peter I Range, with Moscow Peak (22,300 feet); the Darvaz Range, with Arnavad Peak (20,000 feet); and the Vanchsky and Yazgulemsk ranges, with Revolution Peak (22,900 feet). The ranges are separated by deep ravines. To the east of the Yazgulemsk Range, in the central portion of the Pamirs, is the east–west Muzkol Range, reaching 20,400 feet (6,980 metres) in the Soviet Officers Peak. South of it stretches one of the largest ranges of the Pamirs, called Rushan on the west and Bazar-Dara or Northern Alichur on the east. Still farther south are the Southern Alichur Range and, to the west of the latter, the Shugnan Range. The extreme southwestern Pamirs are occupied by the Shakhdarin Range, composed of north–south (Ishkashim Range) and east-west elements, rising to the Mayakovsky Peak (20,000 feet) and Karl Marx Peak (22,100 feet). In the extreme southeast, to the south of Lake Zorkul, lies the east–west Vakhan Range.

Eastern and western Pamirs

It is customary to divide the Pamirs into a western area and an eastern area, distinguished by their forms of relief. In the eastern Pamirs a medium-mountain relief predominates on a high raised foundation. While the heights above sea level average 20,000 feet or more, the relative heights of the peaks above their foundation do not in most cases exceed 3,300–5,900 feet. The ranges and massifs have mainly rounded contours, and the wide and flat-bottomed valleys and troughs between them, situated at heights of 12,100–13,800 feet, are occupied either by quietly running, meandering rivers or by dry channels. The valleys and slopes of the ranges are covered by layers of loose material.

In the western Pamirs the relief is high mountain and sharply disjointed, alternating between low ranges and alpine ridges capped by snows and glaciers; and there are deep, narrow ravines with high, rapid rivers. The valleys and depressions are filled with debris, so that almost the only suitable places for human settlement are the alluvial fans in the valleys of tributaries of the Pyandzh River. The transition from the eastern-Pamirs type of relief to the western-Pamirs type occurs gradually. The conventional boundary is a line joining the ridge of the Zulumart Range with Kara-Bulak Pass in the Muzkol Range; from Pshart Pass, it follows the ridge of the Northern Alichur Range to Lake Zortashkol, where it turns south to the valley of the Pamir River.

Geology. Geologists divide the Pamirs into three zones according to the characteristics of their rock formations: the northern, central, and southern Pamirs. The southern zone consists of metamorphic rocks (gneiss, quartzite, marble, and others) to which a majority of researchers attribute a Precambrian age (more than 570,000,000 years ago). The zone on the whole represents a huge anticlinorium, or series of stratified arches. The central zone of the Pamirs contains limestone, sandstone, and shale rocks of the Jurassic, Triassic, and Permian periods (136,000,000 to 280,000,000 years ago) and also red-coloured terrestrial rocks of the Lower Cretaceous Period (136,000,000 years ago). There are some marine rocks of the Lower and Middle Paleozoic Era (345,000,000 to 570,000,000 years ago) and lava and tufaceous rocks of the Paleocene (54,000,000 to 65,000,000 years ago). The structure of the central Pamirs is that of a huge synclinorium (an inverted arch caused by fracturing); it is separated from the northern Pamirs by a deep fracture.

In the structure of the northern Pamirs, two subzones can be discerned: a Paleozoic zone, which stretches to the ridge of the Trans Alai Range, and a zone beyond, which is composed of more recent deposits. The Paleozoic subzone of the northern Pamirs is a huge anticlinorium with a complex internal structure. It is separated from the Trans Alai subzone by the Karakul fracture. The Trans Alai subzone is very complex. Its western part is a fan-shaped anticlinorium, in the centre of which emerge Jurassic deposits; radiating outward are more recent, dislocated rocks of the Lower Cretaceous Period. The eastern part has Cretaceous and Paleocene deposits in a system of conflicting folds. Because of the numerous overthrusts, or horizontal faults, in some places the layers overlap each other. On the north the Trans Alai subzone is bounded by the deep Gissaro-Trans Alai fracture, separating the Pamirs from the Gissaro-Alai system.

Severity of climate

Climate. The climate of the Pamirs is arid and continental. These features are more pronounced in the eastern part, where there are broad, closed basins in which cold air is retained and a barrier of high ranges intercepting moist air currents. There is a mixed circulation of air—western (cyclonic) and southern (monsoonal). In the valleys of the western Pamirs the amount of annual precipitation is four to 10 inches (100 to 254 millimetres), and in the eastern Pamirs two to five inches. In the high altitudes and on slopes of mountains, the amount of precipitation increases, reaching 32 to 40 inches on the Fedchenko Glacier. The thickness of the snow blanket in the western Pamirs reaches 20 to 28 inches, and in the eastern it reaches 1.5 to four inches. The average January temperature in the eastern Pamirs at heights around 11,500 feet is 0° F (−17.8° C). Winter here is long (October through April) and severe, and extremes of temperature have been recorded at below −58° F (−50° C)—in Bulunkul in 1953, −64.9° F (−53.3° C). In the short summer, temperatures do not rise above 68° F (20° C). The climate of the valleys of the western Pamirs is more moderate. The average temperature in January at heights of about 6,900 feet is 18.7° F (−7.4° C), and in July it is 72.5° F (22.5° C), but temperatures vary greatly. The period of vegetation (with temperatures of 41° F [5° C]) is 223 days in Khorog and 140 days in Murgab.

Glaciation. The Pamirs have 1,085 glaciers, covering an area of 3,105 square miles (8,040 square kilometres). The largest centres of glaciation occur in the Academy of Sciences, Trans Alai, Rushan Northern Alichur, Yazgulems, Darvaz, Peter I, and Zulumart ranges. The largest valley glacier, Fedchenko (length 44.3 miles; 72 kilometres), starts in the Academy of Sciences Range; other glaciers in this range include the Grumm-Grzhemaylo (23 miles), Garmo (17 miles), Sagran (15 miles), and Geographic Society (13 miles). The main glaciers of the Trans Alai Range are the Sauk-Dara (16 miles), Korzhenevsky (around 13 miles), October (10.9 miles), and Lenin (5.9 miles). In

the eastern Pamirs shallow bog glaciers and snow fields predominate.

Drainage. The rivers belong mainly to the basin of the Amu Darya (the ancient Oxus), which forms the frontier between the U.S.S.R. and Afghanistan and flows into the Aral Sea. Its upstream extensions are the Pyandzh and the Pamir. The area also contributes to the basin of the Tarim, which flows eastward into China. Lakes include Karakul (a salt lake), Rangkul, Shorkul, Zorkul, Yashilkul, and Sarez.

Plant life. Vegetation in the Pamirs, especially in the eastern Pamirs, is poorly developed. Bare cliffs or a cover of surface rubble predominate. At high altitudes the vegetation changes with increasing altitude from mountain desert flora to mountain drought-resisting plants, then to mountain steppe vegetation, and finally to cold-resistant plants. Native forms predominate, but some forms from Central and Southwest Asia are found. The eastern Pamirs are a cold, high-mountain desert, with woody vegetation completely absent and with low-growing plants that are adapted to the severe conditions. On the dry mountain slopes there are low shrubs of winter fat, the only form of vegetable fuel in the area, and plant cover, such as the Pamir tansy, oxytropis, astragalus, local species of wormwood, the bulbous iris, and meadow grass. On the bottoms of the moist valleys, sedge and cobresia meadows are abundant.

Mountain flora

The vegetation of the western Pamirs is richer, although on the mountain slopes and bottoms of valleys there is a prevalence of wormwood and haloxylon. At heights over 8,500 feet, spiny pillow-form plants (*e.g.,* acanthus, spiny astragalus) are widespread. At heights over 10,500 feet, yugan, kamol, fescue, and feather grass are found. At 12,500 to 14,100 feet are alpine cobresia meadows; above 14,500 feet vegetation is sparse. Along the rivers of the western Pamirs are dense growths of willow, sea buckthorn, birch, poplar, and hawthorn. A thinner wood–shrub vegetation reaches heights of 12,500 feet, including Juneberry, almond, birch, and juniper. On irrigated lands there are cultivated plantings of grape, apricot, apple, pear, walnut, and mulberry trees.

Animal life. The fauna of the Pamirs are not numerous. In the eastern Pamirs are found the *arkhar* (a mountain sheep), the long-tailed marmot, and the large-eared Tibetan wolf. Birds include the Tibetan mountain turkey, the Tibetan raven, the horned lark, and the snow vulture. The western Pamirs have the mountain goat (*kiik*), brown bear, wolf, fox, snow panther, lynx, weasel, marten, *zayatstolay* (a hare), dormouse, and flying mouse. Birds include the Indian oriole, bluebird, dark-breasted pheasant, stone partridge, and paradise flycatcher. The Pamirs have few fishes; only the carp and the Tibetan loach are known.

The people and economy. More than 90 percent of the population are Tadzhiks living in the western Pamirs. Their languages belong to the Iranian group, and those who are religious are Shī'ite Muslims. The inhabitants of the eastern Pamirs are Kirgiz, who speak a Turkic language and are Sunnite Muslims.

Nearly all are peasants whose chief occupation is farming and livestock breeding. The small amount of arable land is planted with grain, beans, gourds, and potatoes, and there are orchards of apples, pears, apricots, and mulberry trees. Sheep and goats are the predominant livestock.

Agriculture

Industry in the Pamirs includes several small hydroelectric power stations and a few mines. In the eastern Pamirs, brown coal and common salt are mined. There is evidence of industrial deposits of gold, gems, jasper, lazulite, mica, asbestos, and talc. Thermal and mineral springs are common. There are auto routes across the Pamirs from Dushanbe in the west to Khorog in the south and from there to Osh in the east. Another route runs from Khorog south through the once impassable ravine of the Pyandzh–Pamir rivers. Buses go from Khorog to several regional centres. Many other routes are accessible only to pedestrians and pack animals.

Study and exploration. Modern exploration began with the Russian A.P. Fedchenko, who in 1871 succeeded in reaching the northern foot of the Pamirs from the Alai Valley. In 1877 the Russian geologist I.V. Mushketov visited the valley of the Muksu River and the vicinity of Lake Karakul. In 1877–78 the Russian zoogeographer N.A. Severtsov, penetrating into the depth of the mountain country, made a chart of its structure. In 1878 an expedition under the Russian naturalist V.F. Oshanin discovered the large valley glacier subsequently named A.P. Fedchenko. From 1884 through 1887, the Pamirs were explored by the zoologist G.Y. Grum-Grzhimaylo, who provided valuable data on glaciers of the northwestern and northeastern Pamirs.

Under the Soviets, explorations in the Pamirs have become systematic. In 1928 an expedition of the Academy of Sciences of the U.S.S.R. explored the region of the Fedchenko Glacier, making possible the first accurate topographical maps of the northwestern Pamirs. In 1933 the first high-mountain glaciological observatory in the world was constructed on the Fedchenko Glacier, at a height of about 14,800 feet. The Tadzhik–Pamir Expedition of 1932 resulted in important monographs on the geology, geomorphology, and hydrogeology of the Pamirs. Soviet alpinists have contributed much to the investigation of the Pamir region—a part of the world not easily accessible.

(T.K.Z.)

TIEN SHAN

The Tien Shan (Pinyin Dian Shan), Chinese for "Celestial Mountains," forms one of the great mountain systems of Central Asia. Straddling the border between the U.S.S.R. and China, it stretches for about 1,800 miles (3,000 kilometres) from west-southwest to east-northeast. It is about 300 miles wide in places at its eastern and western extremities but narrows to about 220 miles in width at the centre.

The Tien Shan are bounded to the north by the Dzhungarian and southern Kazakhstan plains and to the southeast by the Tarim Basin; to the southwest, the Gissar-Alai Mountains form part of the Tien Shan, making the Alai (Alayskaya), Surkhandarya (Surkhobskaya), and Gissar (Gissarskaya) valleys the boundaries of the system with the Pamir mountain ranges. The Tien Shan also include the Chu-Ili Mountains (Chu-Iliyskiye Gory) and the Karatau Range (Khrebet), which extend far to the northwest into the Kazakhstan lowlands. Within these limits the total area of the Tien Shan is about 386,000 square miles (1,000,000 square kilometres).

The highest peaks are a central cluster of mountains forming a knot, from which ridges extend along the boundary between China and the Soviet Union; these peaks are the Pobeda Peak (Pik Pobedy, or Victory Peak), which is 24,406 feet (7,439 metres) high, and the Khan-Tengri Peak, which is 22,949 feet (6,995 metres) high.

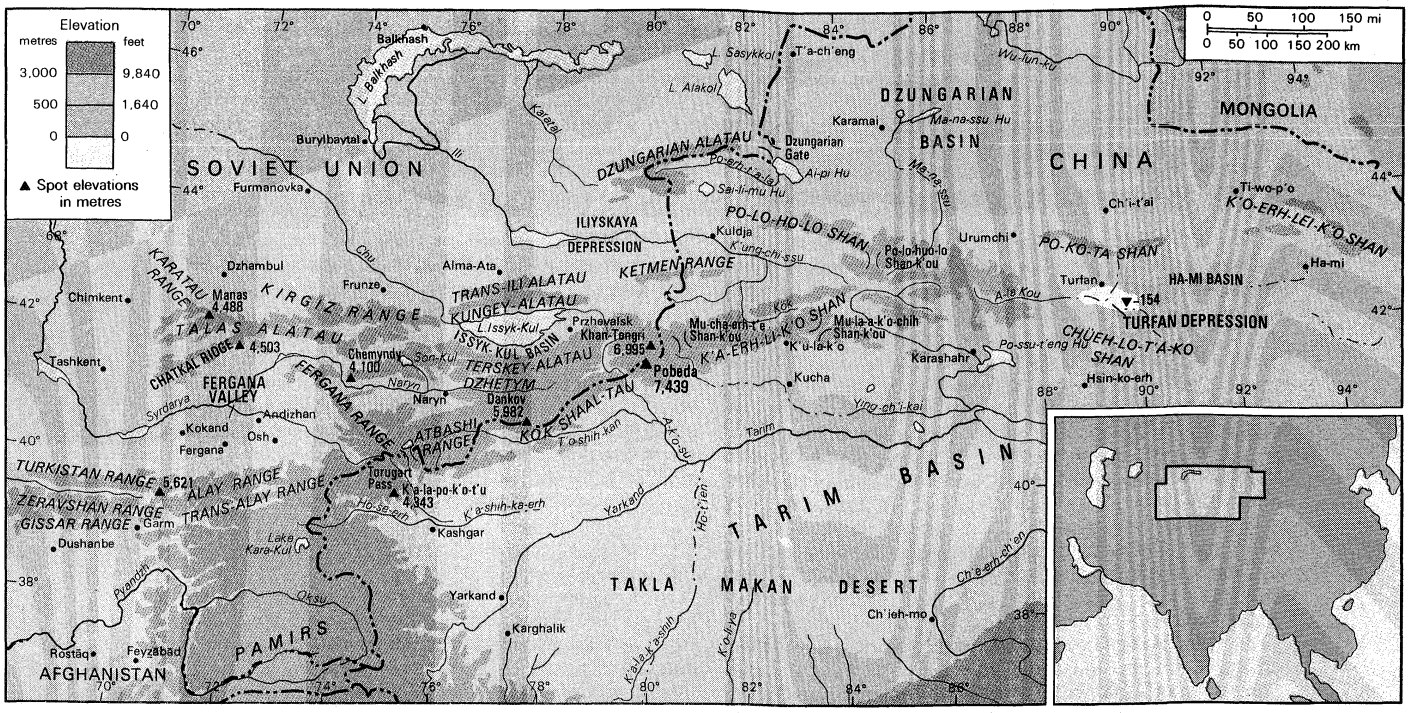
The highest peaks

Physical features. *Physiography.* The relief is characterized by a combination of mountain ranges and intervening valleys and basins, trending generally from east to west. The deepest depression in the eastern Tien Shan is the Turfan Depression, within which is the lowest point in Central Asia—505 feet (154 metres) below sea level. Thus, the differences in elevation in the Tien Shan are extreme, exceeding four and a half miles. The eastern extension of the Turfan Depression is the Ha-mi Basin; both basins are bounded on the north by the Poko-to Shan, with elevations of up to 17,864 feet (5,445 metres), and by the eastern extremity of the Tien Shan, the K'o-erh-lei-k'o Shan, which reaches heights up to 16,158 feet (4,925 metres).

The ranges are of the alpine type, with steep slopes; glaciers occur along their crests. The basins are bounded on the south by the low rising Chiao-lo Shan. West of the Turfan Depression is one of the greatest mountain knots of the eastern Tien Shan: the O-ha-pu-t'e Shan, which reaches. elevations of up to 18,208 feet (5,550 metres). The ridge has considerable glacial development, as well as numerous forms of relief that indicate the area was the site of ancient glaciation.

West of 84° east, the eastern Tien Shan ridges fork, trending in southwestern and northwestern directions, and enclose the vast Ili Depression (Iliyskaya Vpadina), which gradually widens and loses height as it proceeds westward. It is bounded on the north by the Po-lo-ho-lo Shan, which

The Ili Depression



The Tien Shan.

has glaciers in the eastern part and is characterized by steeply sloping ridges. This range also gradually descends westward, where, at a height of 6,801 feet (2,073 metres), lies the great undrained lake of Sai-li-mu Hu. The Ili Depression is bounded on the south by the highest mountains in the eastern Tien Shan—the K'a-erh-li-k'o Shan, reaching heights up to 22,346 feet (6,811 metres), and the isolated Ketmen Range (Khrebet Ketmen), which rises to an elevation of 11,936 feet (3,638 metres) in the central part of the depression.

The northern extremity of the Soviet part of the Tien Shan forms the Dzhungarian Alatau Range (14,645 feet [4,464 metres]), which is subject to considerable glacial action. To the south, the Trans-Ili Alatau Range rises abruptly above the Ili Depression to a height of 16,315 feet (4,973 metres). The successive transition of climatic zones, determined by altitude, from arid and dry steppe at lower levels to glacial at the summit is evident on the northern slopes of this range. The Kirgiz (Kirgizsky) and Talas (Talassky) Alatau ranges, rising above 13,000 feet and located farther west, also belong to the outer chain of the northern Tien Shan. There is a great difference in elevation between these outer mountain ridges and the plains at their base. Streams, therefore, usually plunge down the mountainsides through deep gorges and, as they flow out onto the plains, form vast fan-shaped deposits of silt and mud. On the fertile land formed by this process are located many oases and population centres, including the cities of Alma-Ata and Frunze in the Kazakh and Kirgiz Soviet Socialist republics. The Kungey-Alatau and Terskey-Alatau ranges also belong to the northern Tien Shan. They rise to a height of 17,100 feet (5,212 metres) and border the vast Issyk-Kul Basin, the centre of which is filled by the lake Issyk-Kul.

The Aksay River (T'o-shih-kan Ho in Chinese) Basin and most of the Naryn River Basin are situated within the inner Tien Shan. This region is characterized by the alternation of comparatively short mountain ranges and valleys, both extending east and west. The predominant elevations of the mountains vary from about 10,000 to 15,000 feet, while the elevations of the depressions that separate them vary from between about 6,000 to 10,500 feet. The most important ranges are Borkoldoy (16,565 feet [5,049 metres]), Dzhetym (16,178 feet [4,931 metres]), Atbashi (15,702 feet [4,786 metres]), and the Kok Shaal-Tau Range, in which Dankov Peak reaches a height of 19,626 feet (5,982 metres).

The elevation of the mountains increases in the Saryd-zhaz River (called K'un-a-li-k'o Ho in Chinese) Basin area in the central Tien Shan, which lies to the east of the Akshiyrak Range. The separate ranges gradually converge, forming the high-altitude mountain knot already mentioned, which includes the Khan-Tengri crests and Pobeda (Victory) Peak.

The Fergana Range

In contrast to most of the Tien Shan ranges, which run approximately east–west, the Fergana Range (Fergansky Khrebet), separating the inner region from the western and southern Tien Shan, extends from southeast to northwest. Its maximum elevation is 15,354 feet (4,680 metres). The southwestern slopes display a variety of climatic zones in the course of their gradual descent.

The western Tien Shan ranges lie north of the Fergana Valley (Ferganskaya Dolina). Several short but high and steep ranges running southwest–northeast here meet the southern sides of ranges running westward and northwestward. The highest peak is the Chatkal Ridge (14,773 feet, or 4,503 metres), and the predominant elevations vary from about 7,500 to 10,500 feet.

The southern Tien Shan ranges (including Turkistan, Zeravshan, and Alai, among others) border the Fergana Valley on the south and extend chiefly east and west. The maximum elevation is 18,441 feet (5,621 metres), with several peaks above 15,000 feet. To the south, the Tien Shan meets the Pamirs. Foothills approach the northern slopes of the ranges; there are oases on the plains below the mountains.

Geology. The mountains of Tien Shan are composed in the main of crystalline and sedimentary rocks of the Paleozoic Era (from 570,000,000 to 225,000,000 years ago). The basins that lie between the mountains are filled with sediments from the Mesozoic (225,000,000 to 65,000,000 years ago) and Cenozoic (65,000,000 years ago to the present) eras. These sediments were chiefly formed by the erosive action of the area's rivers. Granite-like rocks crop out over much of the area in the north and east of the Tien Shan.

The north and east portions of the region underwent folding during the mountain-building period that occurred during the Early Paleozoic Era; it has been uplifted dry land since, and its original sedimentary cover has been almost completely obliterated by erosion. The southern and western parts of the Tien Shan, however, consist principally of sedimentary metamorphosed (structurally changed by heat and pressure) rock and, to a lesser degree,

of intrusive and volcanic rock. These regions experienced folding during the Late Paleozoic Era.

A new stage of development began in the middle of the Tertiary Period (about 26,000,000 years ago) and has continued to the present time. It has been characterized by sudden movements of the Earth's crust. Loose fragments of rock have slid into the valleys and formed accumulations; those in the Fergana Valley are almost five miles thick. Shallow lakes were formed in many valleys and later evaporated, leaving behind salty deposits.

Subsequently, glaciers deposited boulder moraines (accumulations of earth and stones) in the mountains, while gravel and loess (wind-borne deposits) strata accumulated in the valleys. Zones of deep faulting occur, usually along the boundaries between the ridges and the valleys. Large-scale horizontal movements have occurred along the great Talas Fergana fault, which traverses nearly the entire Tien Shan system along the northeastern slopes of the Fergana Range and its northwestern extension. The deep faults are associated with catastrophic earthquakes that occurred at Verny (1887), at Kashgar (1902), in the northern Tien Shan chains (1911), and at Chatkal (1946), and Khait (1948).

Largest glacial areas

Glaciation. The total area of the Tien Shan glaciers exceeds 3,800 square miles, of which more than four-fifths is in the Soviet Union. Largest among the several glacier areas are the Khan-Tengri-Pobeda region and the O-ha-put'e Shan. There are also many glaciers in the Kok Shaal-Tau Range, the Akshiyrak Range, the Trans-Ili Alatau Range, and the southern Tien Shan. The largest glacier in the Tien Shan is Inylchek Glacier (Lednik), which is approximately 37 miles long; it descends from the western slopes of the Khan-Tengri massif and branches into numerous tributaries. Other large glaciers in this area include North (Severny) Inylchek (24 miles) and Mu-cha-erh-t'e Shan-k'ou (21 miles). The length of the largest Tien Shan glaciers elsewhere is usually between six and 12 miles; the most usual size is that of the relatively small valley glaciers, from about one and a half to three miles long.

The glaciers are usually fed by snowfall upon the glaciers themselves or by snow avalanches from the surrounding slopes. Glacial action in the Tien Shan is apparently decreasing; most glaciers are either receding or standing still. During recent decades, however, large glaciers in the inner Tien Shan region have made short-term advances. The glaciers of Tien Shan feed many large rivers, including the Naryn, Sarydzhaz, Ili, and Zeravshan.

Drainage. The rivers of the Tien Shan flow into major inland depressions, such as the Azalskaya and Tarim. The largest rivers are the Ili and Chu in the northern Tien Shan, Naryn in inner Tien Shan, Sarydzhaz in central Tien Shan, and Zeravshan in southern Tien Shan. Their maximum flows occur at the end of spring and in summer. Freshets sometimes cause catastrophic flows of mud and stone. Much water is diverted for irrigation. Hydroelectric power plants have been constructed on the Naryn, the largest river of the Tien Shan; the largest is Toktogul, completed in 1979.

The largest lake is the undrained Issyk-Kul, situated at 5,279 feet (1,609 metres). The lake, which has an area of 2,425 square miles, is saline and does not freeze in the winter; it is used for navigation and is a popular resort and tourist attraction. Po-ssu-t'eng Hu (533 square miles in area) is situated in the eastern Tien Shan.

Climate. The position of Tien Shan in the centre of Eurasia governs its sharply continental climate, characterized by great extremes of temperature in summer and winter. The characteristic aridity of the region is manifest in the surrounding deserts and dry regions. The area absorbs much solar heat, and there are about 2,500 hours of sun each year. The climate becomes progressively cooler and more humid as the elevation of the mountains increases. Permafrost (perennially frozen subsoil) is extensive above 9,000 feet. The prevalent air masses are transported over the Tien Shan by moisture-bearing westerly winds from the Atlantic Ocean. Most of the precipitation falls on the windward western and northwestern slopes at altitudes of between about 7,500 and 9,000 feet; it varies from between about 28 and 31 inches at one extreme and 59 and 79 inches at the other. To the east and in the interior regions of the Tien Shan, the total precipitation decreases to between eight and 12 inches, and it amounts to less than four inches in places. Maximum precipitation falls on the southern Tien Shan in March and April, and the summer is dry. In western and northern Tien Shan most of the rain falls during the warm period of the year, with a maximum in April or May. Most of the rain in the inner and eastern Tien Shan regions falls during the summer months. Many mountain valleys here are used as winter pastures because of the small amount of snow that falls in wintertime.

Precipitation

Temperatures vary in the Tien Shan, mostly depending on height. Summer is hot in the foothills: the mean temperature in July in Fergana Valley may reach 81° F (27° C); in the Ili Depression it may reach 73° F (23° C); and up to 93° F(34° C) to the east, in the Turfan Depression, where the climate is even more continental. The temperature in July at a height of about 10,500 feet in inner Tien Shan drops to 41° F (5° C), and frost is possible throughout summer. The mean temperature in January in Fergana Valley is 25° F (−4° C); in the Ili Depression it is 14° F (−10° C); and it drops to −9° F (−23° C) in the Alpine regions of inner Tien Shan, while in places (in particular, Aksay Valley) temperatures of −58° F (−50° C) have been recorded.

Plant life. The characteristics of the living world of the Tien Shan are largely determined by the region's distinct zones of elevation, which provide a diverse distribution of soils and vegetation. In the foothills and plains at the base of the mountains semidesert and desert areas have usually developed; these zones continue to heights of between 5,250 and 5,800 feet. In the Tien Shan they are characterized by ephemeral vegetation growths that die out at the beginning of summer; xerophyte (adapted to a scant supply of water) grasses, wormwood, and the desert shrub ephedra are generally distributed. The most common landscape in the Tien Shan is steppe, which occurs at elevations of between about 3,500 and 11,000 feet.

The forests of the Tien Shan alternate with steppes and meadows. They are principally on the northern slopes and extend to an elevation of 9,000 to 9,800 feet. On the lower slopes of the outer ranges the forests are principally deciduous, consisting of maple and aspen, with extensive admixtures of wild fruit trees (apples and apricots). Vast areas of the southwestern slopes of the Fergana Range are occupied by very ancient nut-bearing forests. Stands of pistachio, walnut, and juniper are found up to 7,500 feet on the shaded slopes of several western and southern Tien Shan ranges. North and east of the Fergana Valley, coniferous forests predominate. At the upper boundary they are often replaced by sparse juniper forests. The water meadow forests in the river valley bottoms, in which aspen, birch, poplar, and various brushwoods ordinarily grow, lie far outside the forest zone. The forest glades and areas adjacent to the upper tree line are usually covered with meadow vegetation. Sub-Alpine meadows of mixed grasses and cereals extend up to almost 10,000 feet on the moist northern slopes but on southern slopes are usually replaced by mountain steppes. There are short-grass Alpine meadows up to 11,500 feet. In the inner and eastern Tien Shan regions, at elevations between 11,500 and 12,000 feet and sometimes higher, the level areas and gentle slopes are "cold deserts," with sparse and short vegetation. Mosses and lichens are found in the areas of the glacial zone that are free of snow and ice.

Animal life. Animals in the Tien Shan include the wolf, fox, and ermine. There are also many typical Central Asian species, inhabiting chiefly the high mountains; these include snow leopard, mountain goat, Manchurian roe, and mountain sheep. The forest–meadow–steppe zone is inhabited by bear, wild boar, badger, field vole, members of the jerboa family (nocturnal jumping rodents), and members of the Ochotonidae family (short-eared mammals related to the rabbits). The many birds include the mountain partridge, pigeon, Alpine chough, crow, mountain wagtail, redstart, Himalayan snow cock, and other species. The lower zones—desert and semiarid regions—are visited by animals from the neighbouring plains, such

as antelope, gazelles, Tolai hares, and gray hamsters. Lizards and snakes are also found. (Y.Y.R.)

The people and economy. Several million people live in the Tien Shan. The Fergana Valley is the most densely populated, with more than 500 persons per square mile (195 per square kilometre) in places. Most of the Tien Shan is occupied by the Kirgiz and Uighur ethnic groups. Tadzhiks, Uzbeks, Kazakhs, and Mongolians reside along the periphery of the region. Substantial Russian and Ukrainian populations have been established in the Soviet part of the Tien Shan in recent decades; Chinese populations live in the eastern Tien Shan. Agriculture has developed in the valleys and on the mountain slopes with the aid of irrigation, and livestock herding is practiced in the mountains.

After the Russian Revolution of 1917, the nomadic Kirgiz and Kazakhs adopted a settled way of life. Their principal occupation is the herding of livestock; in the summer herds of horses, sheep, and cattle are driven to the mountain pastures. Where conditions permit, agriculture is developed. The Uighurs live principally by irrigated agriculture, supplemented by handicraft production. Except for the Mongolians, the other peoples of the Tien Shan also engage in agriculture. (S.I.B.)

Study and exploration. The Russian Geographical Society played a major role in the scientific exploration of the Tien Shan. From 1856 to 1857 the Russian geographer P.P. Semyonov-Tyan-Shansky gave the first scientific description of many regions of northern and inner Tien Shan, while the expeditions of another Russian geographer, G.Y. Grum-Grzhimaylo, in the 1880s, contributed greatly to the exploration of the eastern Tien Shan. The peak of Khan-Tengri was ascended for the first time in 1931 by a Soviet expedition led by M.T. Pogrebetsky. Pobeda Peak, the highest point, was conquered in 1956 by another Soviet expedition led by V.M. Abalakov. (Y.Y.R.)

Deserts

ARABIAN DESERT

One of the great desert regions of the world, the Arabian Desert occupies almost the entire Arabian Peninsula. Covering an area of about 900,000 square miles (2,331,000 square kilometres), it is bordered on the west by the Red Sea, on the north by the Syrian Desert, on the south by the Arabian Sea and the Gulf of Aden, and on the east and northeast by the Persian Gulf, the Gulf of Oman, and the Arabian Sea.

A large part of the Arabian Desert lies within the modern kingdom of Saudi Arabia. Yemen (Ṣan'ā'), on the coast of the Red Sea, and Yemen (Aden), on the coast of the Gulf of Aden, border the desert to the southwest. Oman, bulging out into the Gulf of Oman, borders the desert at its eastern extremity. The sheikhdoms of the United Arab Emirates (former Trucial States) and Qatar rim the region to the north, stretching along the southern coast of the Persian Gulf. The sheikhdom of Kuwait abuts the northern Persian Gulf between Saudi Arabia and Iraq; a Neutral Zone—diamond shaped—lying to the west of Kuwait is shared by Saudi Arabia and Iraq. In the northwest the desert extends into Jordan.

Seen from the air the Arabian Desert appears as a vast expanse of light sand-coloured terrain, with an occasional indistinct line of escarpments or mountain ranges, black lava flows, or reddish systems of desert dunes stretching to the horizon. Camel trails crisscross the surface between watering places. On the ground, features become distinctly individual and the relief seems more prominent. Vegetation at first seems nonexistent, but to the discerning eye can be seen as a minor fuzz on the surface, or as bits of green where shrubs strive to survive. There is almost always a breeze, which changes seasonally to winds of gale force. Cold or hot, these air currents chill the body or roast it. The sun and moon are bright in clear skies, although dust and humidity may cut visibility with little warning. Contrary to ideas commonly held in more northern climes, the desert is often a lovely place.

Physical features. Western Arabia formed part of the African landmass before a rift occurred in the earth's crust, as a result of which the Red Sea was formed and Africa and the Arabian Peninsula became separated. The southern half of the peninsula consequently has a greater affinity with the Somalia and Ethiopia regions of Africa than with northern Arabia or the rest of Asia. The northern Arabian Desert merges imperceptibly into Arab Asia through the Syrian steppe (treeless plain). The bulge of Oman contains a mountain range similar to ranges in Iran.

African affinities

The greatest length of the peninsula from northwest to southeast is about 1,600 miles (2,600 kilometres), from north to south not quite 1,400 miles (2,300 kilometres), while from east to west the maximum width is about 1,300 miles (2,100 kilometres). It is narrowest from Rābigh on the Red Sea to Manama on the Persian Gulf—a distance of about 700 miles.

Three corners are high: the southwest corner in Yemen, where Ḥaḍūr Shu'ayb reaches a height of 12,336 feet (3,760 metres); the northwest corner in Hejaz (a part of Saudi Arabia), where Jabal al-Lawz reaches a height of 8,562 feet (2,580 metres); and the southeast corner in Oman, where al-Jabal al-Akhḍar attains an altitude of 9,774 feet (2,980 metres). Much of the Yemen Plateau is over 9,840 feet (2,999 metres) high. To the north and east elevations decrease. Steep cliffs and steep canyons descend from highlands into adjacent seas south and west. Northeast slopes in Oman are short and steep, but on the southwest flanks the slopes grade gently to the Rub' al-Khali desert basin. The southern plateau is cut by great steep-walled canyons into rugged limestone masses that have held the peoples of that region in isolation for hundreds of years. On lower surfaces eastward in Dhofar grow shrubs that yield the fragrant frankincense and myrrh.

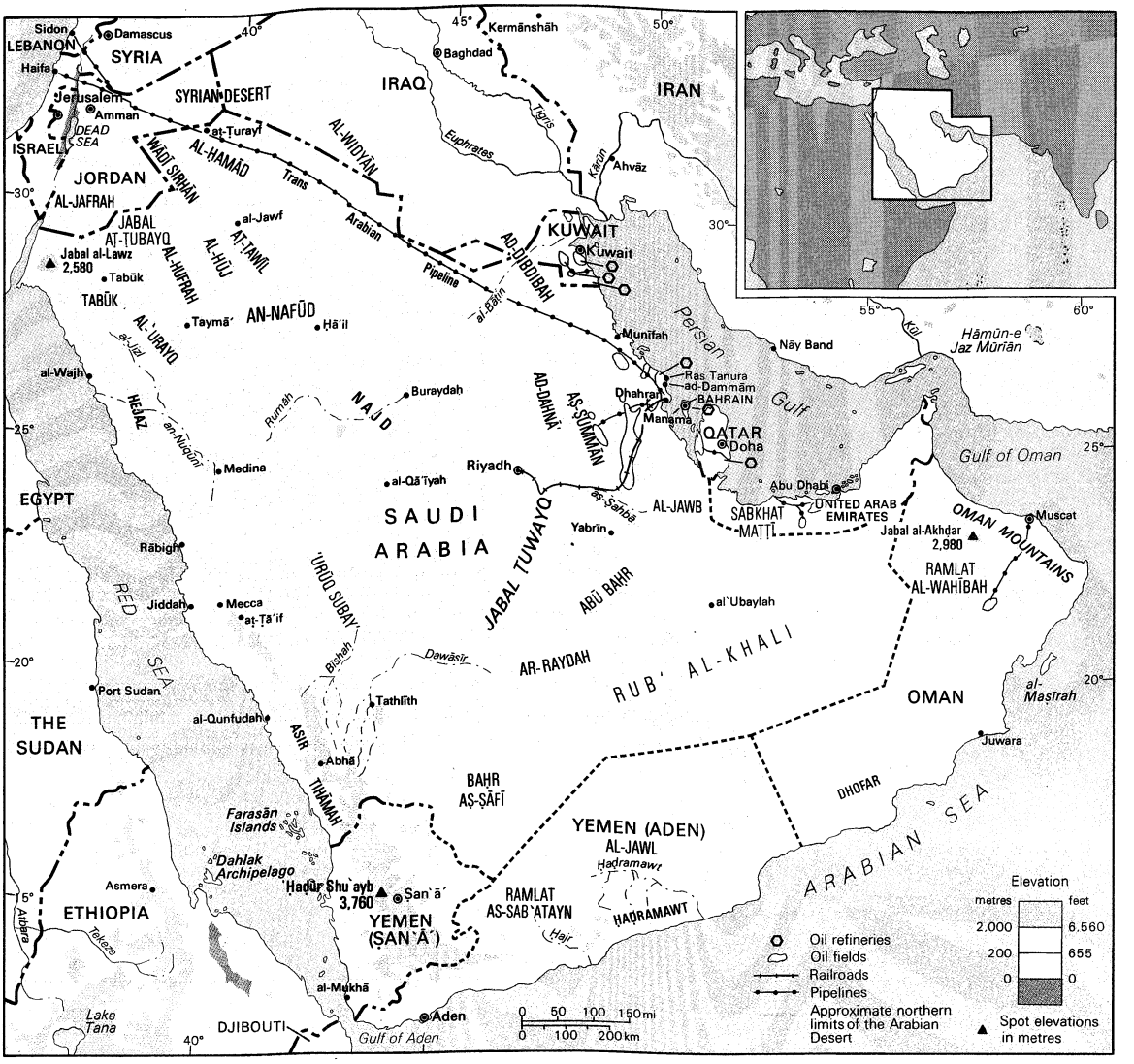
The rest of the peninsula displays a moderate relief characterized by broad plains or steppes. At least a third is covered by sand. The Red Sea escarpment, however, stands in sharp contrast to the moderate relief of the interior plateaus. South of aṭ-Ṭā'if, east of Mecca, the scarp is young, rugged, and dissected into short, steep canyons and ridges. At the foot of the scarp the Tihāmah coastal plain slopes to the sea; at Jabal as-Sawdā', near Abhā in Asir, the drop is about 9,000 feet in six miles. North of aṭ-Ṭā'if the Hejaz and Najd plateaus seldom rise above 3,600 feet (1,100 metres), except where volcanic fields occur or where remnants of the crystalline rocks that underlie the region rise to the surface. The slope to the Persian Gulf averages eight feet per mile.

Geology. The Arabian Desert consists of two major regions: the first, the ancient Afro-Arabian Shield of rocks that are igneous (solidified from a molten state) and metamorphic (transformed by heat and pressure) in the west; and the second, younger sedimentary rocks that pitch gently away from the Shield toward the Persian Gulf basin. Floods of volcanic lava in the west disturbed drainage patterns and created new divides. The highlands of the Hejaz, Yemen, the Ḥaḍramawt, and Oman display structural complexes of the whole rock spectrum. Volcanic activity in Yemen and Oman dates from the Mesozoic Era (from 65,000,000 to 225,000,000 years ago), while more recent volcanic eruptions occurred in Aden, Asir, Hejaz, and Jordan between the Oligocene Epoch, which began 38,000,000 years ago, and the Recent Epoch, which began about 10,000 years ago.

The two major structural regions

Physiography. Mountainous highlands rise in northwest Hejaz, in the Asir region, in Yemen, and in Oman. Lesser ranges have been uncovered by erosion in the interior. Eighteen volcanic fields are scattered through the west, mainly in Hejaz, several of them being more than 10,000 square miles in area.

Plateaus are a common desert feature. Jordan east of the Dead Sea forms a moderately elevated plateau. To the southeast, Jabal Tubaiq rises higher, standing as a mass of sandstone deeply cut into by numerous seasonal watercourses (wadis). Farther southeast the plateaus of Tabūk, Taymā', Ṭawīl, al-Hufrah, and al-Hūj reach to the west edge of an-Nafūd (Great Nafud), a sand desert in the north. Through central Najd, a highland region southeast of an-Nafūd, a series of west-facing scarps mark cuestas (low ridges with steep faces on one side and gentle slopes



The Arabian Peninsula.

on the other) of limestone reaching to highlands of the Ḥaḍramawt in the south, where the plateau of al-Jawl (Jol) is located.

Below the plateaus spread broad plains, stony, cherty (composed of compact microcrystalline quartz), or gravelled, their surfaces well preserved under the arid climate; some plains are duricrusted (covered with a crust of soil formed by salts), having smooth, firm surfaces formed by the cementation of sandy debris at groundwater level. Typical of the stony plains is al-Ḥamād, which stretches from an-Nafūd northward into the Syrian Desert. Chert plains were formed on the post-Eocene to pre-Miocene surface (*i.e.*, between 26,000,000 and 38,000,000 years old) in al-Ḥamad, and in al-Malsūnīyah region east of the Khurayṣ oil field. The gravel plains resulted from deposits left from 10,000 to 2,500,000 years ago by ancient river systems now represented by such wadis as ar-Rummah–Bāṭin, Sahbā, and Dawāsir–Jawb, which carried vast loads of sediment from the interior toward the Persian Gulf. Ad-Dibdibah was the delta of Wādī ar-Rummah–Bāṭin, and al-Ḥadabah was the delta of Wādī aṣ-Ṣahbā. The gravel plains of ar-Raydah and Abū Baḥr, and adjacent areas covered by sand, formed the delta of the Dawāsir–Jawb system. Several of the deltas that were formed by these ancient rivers are as large in area as the delta of the Nile. The plateau of the northern Ṣummān is smooth where it has been uncovered from under ad-Dibdibah gravels. Farther south, the surface is cut by erosion into many closed basins.

When the Arabian Peninsula was shifted away from Africa during the continent-building process, the Afro-Arabian Shield was greatly elevated, thus disturbing prior drainage patterns. The uplifting of the peninsula and the appearance of cracks or faults were accompanied by large upwellings of lavas that filled valleys and covered ridges and mountains. The Arabian platform was tilted northeastward, creating a prominent regional divide along its western rim from Yemen to Jordan. In the south another regional divide separates the coastal drainage of the Ḥaḍramawt from the Wādī Ḥaḍramawt system inland, and a third system, also in the south, divides the al-Jawl region from the system draining into the Rub' al-Khali. The Oman Mountains divide short, steeply graded northeast-flowing wadis from the less steep wadis flowing southwest into the eastern Rub' al-Khali.

The effects of geological uplift

Lesser divides enclose basins in Hejaz, Najd, and Jordan. Four adjacent interior basins are to be found in the shield between latitude 21° and 26° N. Others occur in the Tabūk area, in the Wādī as-Sirḥān, at al-Jafrah in south Jordan, and in the Jordan–Dead Sea depressions. Hundreds of smaller basins, seldom larger than four to eight miles in diameter, are found in volcanic fields, in the shield, and in many poorly drained areas of sedimentary rocks. A few of these small basins are saline, but the vast majority are floored by silts deposited in thin layers.

Most drainage channels in the Arabian Desert are either dry or else are intermittent, flowing only when rains are heavy. Two systems flow perennially in the region—the Tigris–Euphrates rivers and Wādī Ḥajar in the southern Ḥaḍramawt.

The main drainage systems of al-Widyān, ar-Rummah–Bāṭin, Sahbā', and the Dawāsir–Jawb were the scenes of

great floods in the Quaternary Period, which began 2,500,000 years ago. Today floods are infrequent but no less destructive; they seldom, however, reach the desert sands where the channels have been dammed up. The directions taken by several large systems have been altered by stronger streams that have intercepted them; examples are such wadis as the Jizl-Ḥamd in northern Hejaz and the Ḥaḍramawt in the south.

Where they enter large expanses of sand, wadis lose their identities. The alluvium carried in flood time is added to the sand body and redistributed by the wind. About two-thirds of the Arabian Desert is drained by complete wadi systems. The intermittent action of running water is more effective than the erosive action of the winds in shaping the landscape. Most of the exceptions are provided by the formation of sand dunes and the grooving of rocks by wind-borne sand.

The *sabkhah* (saline flat) of Eastern Arabia is usually a gently sloping sandy plain with concentrated salt brine at or just below the surface, located on a coastal plain. It has formed by the filling of embayments with sand—a process that continued as the sea level retreated—while the high rate of evaporation concentrated the trapped seawater to strong brine. Scouring of the sandy surface by *shamāl* (north to northwest) winds exposes the salty crust, which may then be dissolved into brine by winter rains. High tides may spread the saline surface waters inland. The salt crust does not usually attain a thickness of more than three feet. It may be interbedded with sand, silt, mud, or other deposits formed by evaporation, such as gypsum. The surface of Sabkhat Maṭṭī, the largest single exposed *sabkhah,* is sometimes composed of soft, wet, muddy, and salty slush, with a hard crust less than three feet deep. It is very treacherous to the unwary traveller, be he on foot, camelback, or in a motor vehicle. The *sabkhah* surface has no direct analogy with quicksand, but its danger lies in the inability of the traveller to recognize its nature in time to avoid sinking into a morass. Northwest of the oil refinery on Ras Tanura, *sabkhah*s carry a thin crust of salt and sand, underlain by soft calcareous (chalky) mud that is formed in shallow tidal flats by algae and eventually is covered by sand or salt. This calcareous mud has the consistency of custard.

The eastern Rub' al-Khali has a broad *sabkhah* floor that rises to less than 330 feet in elevation, which seems to be the highest altitude that *sabkhah*s attain.

There are saline basins, akin to playas (the sandy, salty, or mud-caked floors of desert basins) in the Americas, which develop in enclosed valleys, either from the evaporation of flood waters bearing dissolved minerals, or—more usually—from the evaporation of saline waters nurtured by nearby outcrops of salt. The Arabic name for this kind of salt flat is *mamlaḥah.* Arabs have quarried crude salt from both *sabkhah* and *mamlaḥah* for hundreds of years.

Sand covers at least a third of the Arabian Desert, occurring as dunes of varying size and complexity or as a thin film on surfaces of lower relief. With few exceptions, sand does not accumulate in flat sheets but in dunes, ridges, or giant complexes. The variety of dune shapes and sizes in the Arabian Desert is legion. Many forms have not been described in print. Contrary to statements by early western explorers that there is nothing but a shapeless sea of sand, the sand desert develops along systematic lines with distinct and characteristic patterns. There are also clear-cut genetic relationships between dunes in adjacent areas. In such huge sand areas as the Rub' al-Khali, the evolution of dune forms can be traced from simple dunes into the more complex types.

Sand-dune formations

The two largest sand bodies in Arabia are an-Nafūd in the northwest, and the Rub' al-Khali (ar-Ramlah) in the southeast. An-Nafūd has an area of 26,000 square miles; the Rub' al-Khali has an area of 229,000 square miles. Between them are two almost parallel arcs of more or less continuous dunes. The outer arc, convex to the east, is ad-Dahnā', about 750 miles long and from six to 50 miles wide. The inner arc is shorter and less continuous and includes six elongated *nafūd*s situated in low areas between the west-facing limestone escarpments of the central Najd. The two major arcs are separated by the great Jabal Ṭuwayq cuesta. Ad-Dahnā' connects with an-Nafūd in the northwest and with the Rub' al-Khali in the southwest. The inner belt grows more disconnected south of 24° N and dies out before reaching Wādī ad-Dawāsir.

An-Nafūd fills an irregular basin surrounded by hills, ridges, and buttes (isolated hills rising abruptly from the surrounding land) or mesas (land formations having relatively flat tops and steep rock walls). Its sand was eroded and carried by winds from outcrops of sandstone that lie west and northwest. An-Nafūd is characterized by a unique dune form, a giant crescent slipface (the slope on the leeside of the dune that approximates the angle of rest of loose sand, usually about 32°) moving through a thick mass of loose sand. The lee hollow in front of the slipface is floored by bedrock, barren of sand. The impression obtained from an aerial view is of hooflike hollows, that extend from the southwest edge, well across the Nafūd. Along other margins of the desert and through the centre are linear and pyramidal dunes. The Nafūd has been a barrier to travel for ages, only rarely attempted by Western travellers. A caravan route runs from northwest to southeast connecting the oases of al-Jawf with those of Ḥā'il. The topographic relief of dunes in the Nafūd may exceed 300 feet.

With an area of about 230,000 square miles—larger than France—Rub' al-Khali displays a variety of dune forms that appears to be unique, although parts of the Sahara show many similar dunes as well as a few that are different. The name Rub' al-Khali (the Empty Quarter) is not usually used by the Bedouins who travel over it, who refer to it instead as ar-Ramlah (the Sand). The Rub' al-Khali has five general types of sand terrains: (1) crescentic or transverse; (2) linear; (3) giant dune complexes, or "sand mountains," that take the shapes of domes, pyramids, huge crescents, and sigmoidal (S-shaped) ridges; (4) hook-shaped dunes; and (5) dunes with vegetation and moisture. The eastern Rub' al-Khali fills a broad shallow basin, floored mainly by *sabkhah*s that slope toward the southern shores of the Persian Gulf. The floor west of the 51st meridian rises to the southwest to nearly 3,600 feet (1,100 metres) and is composed largely of gravel plains. Highlands surround this sand desert on the northeast, south, and west.

"The Empty Quarter"

The greatest volumes of sand lie in the eastern basin and—as has been discovered through the use of aerial photography—are arranged in belts. These belts appear to have accumulated during the Pleistocene Epoch (from 10,000 to 2,500,000 years ago) at several successive stages that were related to the advance and retreat of sea levels.

Ramlat as-Sab'atayn in eastern Yemen, in the lowland south of the western border of the Rub' al-Khali, has an area of about 60 by 150 miles, or about 10,000 square miles. It is composed mainly of transverse and linear dunes.

Ramlat Ahl Wahībah is located near the extreme eastern cape of the peninsula; it is an oblong dune field 60 by 100 miles wide, made up of linear ridges and interdune valleys oriented almost north and south.

An elongated sand body named al-'Urayq lies in the southern part of the Wādī ar-Rummah basin. 'Irq as-Subbay' (*'irq* is an Arab word for linear sand dune) dams the middle Wādī al-Jarab drainage system. A small cluster of broad transverse dunes has formed on the south coast east of the port of Aden.

Climate. The Arabian Desert spreads across 22° of latitude from the 12th to the 34th parallel north; it may thus be considered as a tropical desert. Summer heat is intense, reaching a maximum 124° F (51° C). In the interior the heat is dry and tolerable. Coastal regions and some highlands, however, attain high summer humidity with dews and fogs at night or early morning. Rainfall averages less than four inches (100 millimetres) a year but can range from zero to 20 inches. Interior skies are usually clear except for intermittent winter rains, spring hazes, or dust storms. Torrential rains flood the main drainage basins infrequently. Winters are invigoratingly cool, with the coldest weather occurring at high altitudes and in the far north. A minimum recorded at aṭ-Ṭurayf on Tapline (the Trans-Arabian Pipeline) in 1950 was 10° F (−12° C),

and was accompanied by several inches of snow and an inch of ice on ponds. Summer rains in the Rub' al-Khali accompany the monsoon winds from the Indian Ocean. Winter rains may occur in the northern Rub' al-Khali. The most arid part of the Arabian Desert appears to be on the western margin of the Rub' al-Khali, north of Wādī ad-Dawāsir.

Dominant winds blow from the Mediterranean, swinging to the east, southeast, south, and southwest in a great arc. Two semiannual windy seasons occur from December to January and from May to June. These are called *shamāl*s (from the Arabic: "north"), and last from 30 to 50 days with wind velocities averaging 30 miles per hour. *Shamāl*s, which try the patience of man and beast, are dry, transport huge loads of sand and dust, and alter the shapes of sand dunes. Millions of tons of sand are carried by each storm into the Rub' al-Khali. Blown sand does not rise more than a few feet, except when picked up by whirlwinds, dust devils, or regional sandstorms. Winds box the compass in central Najd and in the southeastern Rub' al-Khali. Strong southeast gales sweep the big sand desert for several days at a time, reversing the effect of *shamāl*s on dunes.

The power of desert winds to excavate basins is much exaggerated. The wind is important as an agent of erosion but, as mentioned above, is never as effective as running water.

"Brown rollers"

The sudden appearance on the horizon of the "brown roller" in spring or fall can be frightening. This is a frontal storm up to 60 miles wide, carrying sand, dust, and debris high in the air; it is followed by a sharp drop in temperature and often by rain. Wind velocities reach gale force for half an hour or so. Hot days produce myriads of dust devils (*jinn*) and the ill-famed mirage.

Soils. Desert soils differ from humid soils in that they undergo less chemical weathering. Mechanical weathering breaks down coarse particles into finer grains. Quartz sand abounds, covering more than a third of the desert surface. Granular debris from the crystalline basement of the Precambrian shield forms pebbly fans about the bases of hills. Sands and silts are washed down to lower levels and are then winnowed out by winds. Fine materials grade down to silt. Smaller particles, such as clays, rarely form. Limestone, when pulverized, forms silt-sized dusts. Waterborne silts eventually are deposited in *khabari,* or silt flats. Irrigated silt flats are farmed, the soils having proved to be fertile. Najd villages that once depended on November rains to raise their winter wheat, now irrigate and farm all year. Highlands of Asir and Yemen are terraced for crops. Salt flats in the desert, though too salty for many crops, if irrigated and drained properly, can be cultivated. Many of the plants that grow in saline soils, and which are called halophytes, are succulents that are suitable as camel fodder. The date palm thrives on salty soils if properly irrigated and drained.

Desert dune sands are generally dry but can hold rainfall to depths of three feet or more, thus nourishing xerophytes (plants adapted to survive under arid conditions). Shrubs unique to the area, called *'abl* and *ghaḍā,* send out long, shallow roots to catch the slightest bit of moisture. The roots make good firewood.

The highlands of Yemen and Asir produce forests of juniper trees. Valleys and lower slopes are extensively terraced for soil and water conservation and produce many crops, of which coffee is important in local markets. The soils are derived from crystalline rocks of high mineral content.

Plant life. There is a great variety of desert flora and fauna. Plants are primarily xerophytic or halophytic. After spring rains long-buried seeds germinate and bloom in a few hours. The normally barren gravel plains turn green. Even chert plains produce late winter and early spring grazing for camel and sheep. The plains were once the home of the famed Arab horse, but grazing was always too poor to support a large horse population. Certainly all the grazing areas were overgrazed, thus contributing to the formation of the present widespread barren tracts. Halophytes grow on the saline flats and include many succulents and fibrous plants that can be eaten by camels. The sedge, which grows in sandy areas, is a tough plant with deep roots that help to hold down the soil. The tamarisk tree is often found on the borders of oases, where it helps to prevent the encroachment of sand.

The rare shrub *rāk,* or *arāk,* is known as the "toothbrush bush," its twigs being used by Arabs to polish their teeth. Many herbs grow throughout the desert and are well known to the Bedouins, who use them for seasoning, preserving food, perfuming clothing, and washing hair. The eastern Rub' al-Khali, generally thought to be dry and barren, supports much plant life on flanks of giant dunes, including a sweet grass called naṣī that provides the main forage for the now-rare oryx (a species of African antelope). There are no cacti in the Old World, except for those imported from the Americas. One of these imports, the prickly pear, thrives and is fed to livestock, the fruit being eaten by people. Spiny and thorny plants are common. The euphorbia, a plant with milky juice and flowers with no petals, grows in Hejaz, and the camelthorn is everywhere. Acacia trees were once abundant in Jibal Tuwayq, but the demand for charcoal decimated them. The few growing in wadis and gardens provide welcome shade. A single clump of four stands in daring isolation on the edge of al-Mijann overlooking the southern Persian Gulf and the feared Sabkhat Maṭṭī (the treacherous saline flat already mentioned). Junipers reach a great size in high Asir and Yemen. The trunks are cut into the beams and pillars that characterize the region's architecture. The milkweed tree (*ishar*) grows to a height of 20 feet in Wādī al-Bāṭin and is common in the wadis of Najd and in Wīdī Bīshah.

The ubiquitous date palm

The date palm is grown in many oases, the dates themselves providing food for humans, camels, donkeys, and horses. There are many varieties. The palm supplies wood for building and for making water well frames and pulley shafts of ancient type; its fronds are used for handicrafts and for thatching roofs. The oases also produce many fruits and vegetables such as rice, alfalfa, henna (a shrub which yields a reddish-orange dye), citrus, melons, onions, tomatoes, barley, wheat, and—in higher regions—peaches, grapes, and prickly pears.

Animal life. The animal life is varied and unique. Desert insects include the fly, the malaria-carrying *Anopheles* mosquito, fleas, lice, ticks, roaches, ants, termites, beetles, and mantids (predatory insects), which camouflage themselves as leaves, twigs, or pebbles. Also found are the scavenging dung beetle, myriads of butterflies, moths, and caterpillars, and the pestiferous locust that periodically plagues the landscape and populace in great swarms.

Arachnida (a class of segmented invertebrates) include large sapulgids (scorpion-killers), scorpions, and spiders. Sapulgids grow to eight inches (20 centimetres) in length. Scorpions also range up to eight inches and are coloured black, green, yellow, red, and off-white. The scorpion's painful sting is deadly to small children.

Pools in oases contain small fish. There are a few amphibious animals, such as newts, salamanders, toads, and frogs. Reptiles include lizards, snakes, and turtles. The *ḍabb,* a fat-tailed lizard, lives on the plains and reaches a length of up to three and a half feet. It is a vegetarian with toothless jaws; its fat tail, roasted, is a Bedouin delicacy. The monitor lizard reaches lengths up to three feet; it feeds on locusts and other insects. Many lizards, including skinks, geckos, agamids, and collared lizards, are found in the sands. A salmon-coloured lizard, the *dammūsa,* is lively and pretty, seeking the black beetle for food, and literally diving and swimming in the slipfaces of the sand dunes. An agamid lizard (*ṭuḥayḥī*) scurries across the sand with its tail coiled like a watch spring, uncoiling when it stops.

Among the snakes, all of which are feared by Arabs, the sand cobra—relative of the sea snake—is slim, sand-coloured, and venomous. Vipers abound in sand and rocks but, being nocturnal, are seldom seen in the heat of day.

Birds of the Arabian Desert include local species as well as migrant groups. The migrants are from northern Europe, Africa, and India. The local birds breed from late winter to early spring. Many of the young display excellent camouflage. The bifasciated (striped) lark, the sand grouse, the Arabian courser, and the lesser bustard live in the desert all year, as also do several falcons, eagles,

and vultures. The peregrine falcon is seen in Asir; saker and lanner falcons (a brown falcon with a golden cap) are found in Najd and the Eastern Province; and the kestrel is everywhere. The saker falcon (an aggressive light-brown falcon) is often captured young and trained by Bedouin falconers to hunt the bustard and sand grouse. Ravens in pairs or flocks may appear anywhere. Three eagles are known—the white-tailed, golden, and tawny eagles. Vultures were more numerous when camels were in greater use. The largest, a black species with a wing spread of up to 12 feet, has nearly disappeared. The Egyptian vulture (*ar-rakhamah*), a medium-sized bird coloured white and black with yellow, is widely distributed. The lammergeier vulture lives in Asir and Yemen. There are several owls, a burrowing species being common.

Migrant birds follow several flyways, one through the central Najd and others on each coast. Water and shore birds migrate in fall and spring between northern Europe and the tropics. Bee-eaters, warblers, babblers, carrion kites, swallows, martins, swifts, wheatears, shrikes, larks, flycatchers, hoopoes, and some exotic species may be seen alone, in pairs, or in flocks. Cranes, heron, flamingoes, ducks, and small wading birds feed on shores and in the lakes which intermittently appear. The ostrich, once abundant in the sand deserts, has been extinct since 1940.

Desert game

Mammals were probably numerous before the Arabs hunted them with rifles, shotguns, and submachine guns from motor vehicles. Gazelles, which roamed the plains in herds of hundreds before World War II, are almost extinct. The oryx had disappeared from the Rub' al-Khali by 1960. The ibex (a species of wild goat), which dwells on cliffs in Jibal Tuwayq, has been decimated. In desert plains, the ratel (a badger-like carnivore), the fox, and the civet cat live in territorial isolation. Wolves are much feared but, although widespread, are not numerous. The hyena lives wherever sheep are herded, preferring escarpments that provide cover. Jackals are to be seen, especially at dusk when they appear for water. There are hares, as well as golden sand rabbits. Small rodents include the jerboa (a mouselike rodent), mice, rats, and porcupines, while small hedgehogs are often found among rocks. Troops of baboons run in Asir.

The people. Man has inhabited the Arabian Desert since early Pleistocene times. His artifacts have been found widely spread throughout the desert but are most abundant in the southwestern Rub' al-Khali, where game was prolific until about 30 years ago. Archaeological research was not encouraged by the government until recently, so little is known about mankind in Arabia in earlier epochs. Remains of cultures from the last 3,000 years occur in Hejaz, Asir, Yemen, and the Ḥaḍramawt. More recent remains—perhaps 1,000 to 2,000 years old—are being examined in the Eastern Province; most of them date from the early years of the Islāmic period.

The Bedouin know little of their history beyond the dawn of Islām. They adapted to nomadic desert life by breeding camels, Arab horses, and sheep, but also conceded to necessity by growing date palms and other crops. The Bedouin seldom abandoned the nomadic life for the oasis longer than was essential to gather crops, generally hiring others for agricultural labour; finding grazing and water were their main concern, in addition to conducting raids to seize horses and camels. Grouped in hereditary tribes, the Bedouin claimed certain lands as their *dirah* (tribal territory), where their flocks could graze and water. After 1925 King Ibn Sa'ūd of Saudi Arabia, who encouraged the Bedouin to settle in oases, was able to use his influence to prevent intertribal raiding, and to make the tribal *dirah* of less importance. Tribal loyalties remain strong and no Bedouin of a noble tribe will marry into a lesser tribe. Islām is their way of life, but modernization has brought much change and has lessened religious influence. Townsmen have kept their identity distinct from that of the Bedouin; each group regards the other with contempt. This feeling has lessened somewhat, however, since more Bedouins have settled in towns.

Western cultural influences

Western cultural influence began with the discovery of petroleum in 1936 and led to the introduction of the motor vehicle, airplanes, telephone, radios, telegraphs, and, recently, television. The cinema has also become popular, despite disapproval by religious leaders.

Schools were few and were only for boys until 1960; since then, however, girls are also taught. The veil is still worn by women, and, though the lot of women has improved, their place remains in the home.

The economy. *Transportation.* Travel today in the Arabian Desert is easy and rapid. Instead of the slow, laborious camel caravan, motor vehicles now roar across desert terrains. Jet planes fly overhead, and the railroad from ad-Dammām on the Persian Gulf to Riyadh, the political capital of Saudi Arabia, covers the distance in a few hours. Arab countries have planned rebuilding the railroad from Medina to Damascus, destroyed in World War I, but conflict with Israel appears to have halted construction. Paved roads cross the desert from the Persian Gulf to Jidda and also follow the Trans-Arabian Pipeline from Dhahran to Sidon in Lebanon. Semi-maintained dirt roads connect many Arab communities. New roads have been built into Asir, making that province and its fine agricultural produce accessible to the markets of Riyadh, aṭ-Ṭā'if, Mecca, Jidda, and Medina. A Saudi Arabian government airline schedules flights between towns in different regions of the desert.

About 80 percent of the Arabian Desert is accessible to motor vehicles, with or without roads. Essential equipment consists of vehicles equipped with sand tires and two-way radios, and having the capacity to carry fuel, food, and water, as well as repair parts and camping equipment.

Resources. The greatest natural resource of the Arabian Desert is its underground water supply, which, as it remains virtually unreplenished because of low rainfall, in effect consists of Pleistocene waters that are now being tapped. Modern techniques are being used by the governments of Arab countries to develop water sources and to irrigate soils for farming.

The oil boom

Petroleum was found in 1936 in the Eastern Province at Dammām Dome, on which the oil-company town of Dhahran is built, but commercial production was not achieved before 1938. Since World War II many new oil fields have been brought into operation, with production exceeding 3,000,000 barrels daily; reserves are enormous and the end is not yet in sight. Prospecting for other minerals has been in progress since 1950.

Building materials in use before the mid-20th century were stone, adobe, a crude cement made from impure calcareous (chalky) rock taken from the floor of the Persian Gulf, the wood of the date palm, the wood of the juniper in the Asir region, and thatches of cane, palm fronds, or other vegetation. Modern construction utilizes steel, concrete, light alloys, imported lumber, local stone—especially granites from Asir and limestone from the Najd—and slates from Hejaz. Salt and gypsum are produced from saline flats.

Study and exploration. Before the 18th century few travellers or would-be conquerors had succeeded in penetrating the Arabian Desert; fewer still had written about it. Scientific exploration began in 1762 with a Danish expedition led by the German surveyor Carsten Niebuhr. In the 19th century, British officers of the Indian government undertook surveys of the surrounding seas and coasts.

The first classical work on the geography of Arabia, *Travels in Arabia Deserta* (1888), was written by the English traveller C.M. Doughty. At the turn of the 20th century, the Czech explorer Alois Musil travelled through northern Hejaz and Najd, mapping topography as he went. In 1917 H. St. John Philby, an official of the British Foreign Office, paid a visit to the Sultan of Najd (later King Ibn Sa'ūd of Saudi Arabia); Philby later became a Muslim, settled in Riyadh as a councillor to Ibn Sa'ūd, and explored the Arabian Desert by camel and motor car, writing detailed and accurate accounts of his travels. Another British official, T.E. Lawrence, who worked with the Sultan of Hejaz against the Turks in World War I, gained fame for his romantic writings about his exploits in the region. Many individuals travelled in limited parts of the desert; the most notable among them is the British traveller Wilfred Thesiger, who crisscrossed the Rub' al-Khali after World War II.

The discovery of oil in 1932 on Bahrain and the subsequent American exploration of the Arab mainland that began in 1933 led, as already mentioned to the opening of the Arabian Desert to Western influence. After World War II, geographical and geological exploration intensified and was accompanied by vast aerial photographic surveys as a result of which the first accurate maps of the peninsula were prepared and published between 1956 and 1965. Since 1950, the U.S. State Department and the U.S. Geological Survey have co-operated with the Kingdom of Saudi Arabia in conducting surveys of mineral resources.

(D.A.H.)

GOBI

One of the great desert and semidesert regions of the globe, the Gobi (from Mongolian *gobi,* meaning "waterless place") stretches across the vastness of Inner Asia over huge stretches of both the Mongolian and Chinese People's republics. Contrary to the perhaps romantic image long associated with what—at least to the European mind—was a remote and unexplored region, much of the Gobi is not sandy desert but bare rock. One may drive over this surface by car for long distances in any direction; northward, say, toward the mountains of the Altai and Hangayn ranges or eastward toward the city of T'ien-shan or southward toward the Pei Shan-mo. To the west, 1,000 miles (1,600 kilometres) from the Gobi's eastern limits, lies the Sinkiang region, a great basin enclosed by the Plateau of Tibet to the south and the Tien Shan ranges to the north. The desert as a whole occupies a vast arc of land 1,000 miles long and 300 to 600 miles wide. The region first became known to Europeans through the remarkable 13th-century descriptions of Marco Polo. Modern geographical study of the Gobi has mainly been undertaken by Russian scholars, though a series of works has appeared in Mongolian and Chinese since the 1960s. For the purposes of the present article, the Gobi is defined as lying between the Altai Mountains and Hangayn Nuruu (Hangayn Mountains) in the north; the eastern Tien Shan in the west; and the A-erh-chin Shan-mo, Pei Shan-mo, and Yin-shan Shan-mo in the south.

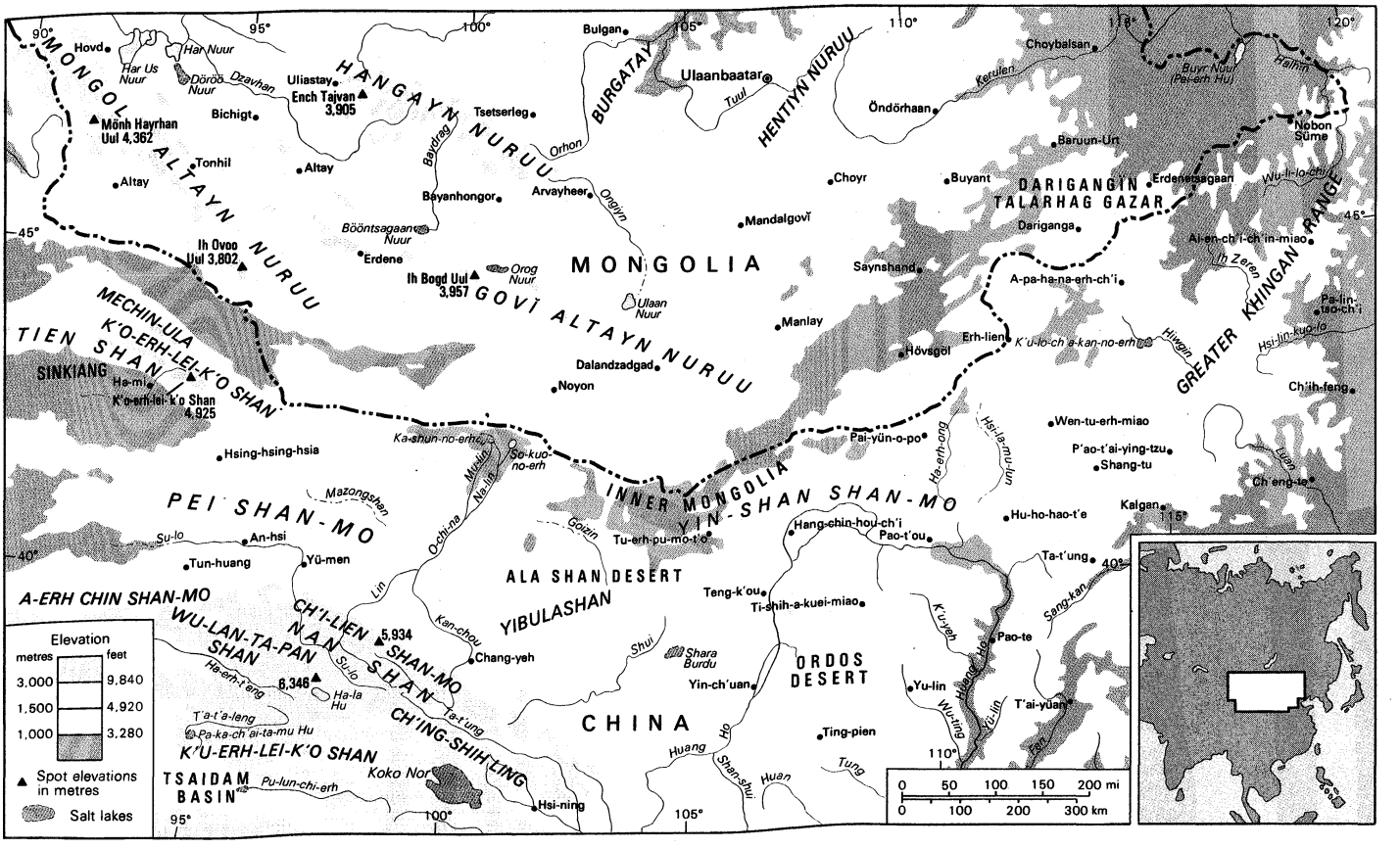
Physical features. *Physiography.* The Gobi consists of the Ka-shun, Dzungarian, and Trans-Altai Gobi (south of the Mongolian Altai Mountains) in the west and the Eastern, or Mongolian, Gobi in the centre and east.

Regional subdivisions

The Ka-shun is bounded by the spurs of the Tien Shan to the west and the Pei Shan-mo in the south and rises as high as 5,000 feet (1,525 metres). It is gently corrugated, with a complex labyrinth of wide hollows separated by flat hills and rocky crests sometimes rising more than 300 feet above the plain. The desert is stony and waterless, though salt marshes lie in the secluded depressions. The soil is grayish brown and contains gypsum. Vegetation is very rare, though richer in the riverbeds, where there are individual shrubs of tamarisk, zaysansky haloxylon and nitre bush (both saltworts), and annual halophytes (plants growing naturally in soils containing certain salts).

The Trans-Altai Gobi is situated between the eastern spurs of the Mongol Altayn Nuruu (Mongolian Altai), the Govĭ Altayn Nuruu (Gobi Altai) in the north and east, and Pei Shan-mo in the south. The plain is elevated, sharp, and rugged. Alongside the plains and the isolated group of low, rounded hills is a fairly extensive mountain area, extending over six miles out into the plain. The mountains are very barren and broken up by dry ravines. The western section of the Trans-Altai Gobi is basically a plain, too, but interspersed with small raised areas and furrowed by dry riverbeds and, again, with extensive salt marshes. In the central portion this fragmentation increases, and mesas (flat-topped, steep-sided hills) appear along with dry gullies ending in flat depressions, occupied by *takyr* (clayey tracts). The Trans-Altai Gobi is quite parched, with annual precipitation of less than four inches, though there is always water underground. There are virtually no wells and springs, however, and vegetation is very sparse and almost useless for livestock.

The Dzungarian Gobi is situated north of the Ka-shun Gobi, between the eastern spurs of the Mongol Altayn Nuruu and the eastern extremity of the eastern Tien Shan (the K'o-erh-lei-k'o Shan). It is like the Trans-Altai Gobi, and its edges are fractured by ravines, alternating with residual hills and low mountain ridges.



The Gobi.

The Eastern Gobi is of similar character, with altitudes varying from 2,300 to 5,000 feet, but enjoys rather more precipitation—up to eight inches (200 millimetres) a year—though with virtually no rivers. The underground waters are relatively abundant and only partly mineralized. They are also near the surface, feeding small lakes and springs. The vegetation, however, is sparse: herb wormwood in coarse, grayish-brown soil. In the moister depressions there are the usual salt marshes and grassy swamps. In the northern and eastern outlying regions, where more precipitation occurs, the landscape of the desert gradually mellows, sometimes even becoming steppes.

Geology. The Gobi's various chalky plains are chiefly Paleocene to Recent (up to 65,000,000 years old), though some of the low, isolated hills are older. The terrain contains small masses of shifting sands. In the central Gobi, Mesozoic remains of dinosaurs (65,000,000 to 225,000,000 years old) and Paleogenic and Neogenic fossils of mammals have been found. The desert also contains Paleolithic and Neolithic sites occupied by ancient man.

Fossils

Climate. The climate is acutely continental and dry, ranging from January's −40° F (−40° C) to 113° F (45° C) in July. Winter is severe; spring is dry and cold; and summer is warm. The annual total precipitation varies from 2.7 inches (69 millimetres) in the west to more than eight inches in the northeast (the maximum falls in summer), and in the eastern regions the impact is quite monsoonal. North and northwestern winds prevail over the Gobi.

Drainage and soils. Aridity in the Gobi depends on the relative chalkiness of the soil and has been aggravated by the strong mountain structure to the west. The lakes have correspondingly shrunk, leaving a series of terraces considerably farther from and higher than the present shorelines. Indeed, Wu-lan Hu (east of the Pei Chan-mo), Orog Nuur, and Böontsagaan Nuur (in the easternmost Mongolian Altai) are but shadows of their former selves.

The drainage of the desert is largely underground; surface rivers have very little constant flow. Mountain streams are confined to the Gobi's fringes and even then quickly dry up, as they disappear into the loose soil or the salty, enclosed depressions. Many rivers flow only in summer. On the other hand, subterranean water is widespread and of sufficient quality to allow cattle raising.

The soil is chiefly grayish-brown and brown carbonaceous (rich in carbon), gypseous (containing gypsum), coarse gravel, often combined with sandy salt marshes and *takyr.*

Plant life. Vegetation is sparse and rare. On the plateau, and on the plains beneath the mountains, small bushlike vegetation occurs: echinochloa (a type of succulent grass found in warm regions), yellowwood bean caper, winterfat (a shrub covered with densely matted hairs), Dzungarian reaumur, nitre bush, and bushlike halophytic vegetation. In the salt marshes, too, halophilous groups prevail: potash bush, Siberian nitre bush, tamarisk, and annual halophytes; in the sands grow zaysansky haloxylon, the sandy wormwood, and sparse perennial and annual herbs such as the Gobi kumarchik and timuriya. In semidesert tracts vegetation is richer, belonging to the herbaceous-wormwood groups: Gobi feather grass; the annual Gobi kumarchik (*Agriophylum gobicum*); the perennials timuriya (*Timouria villosa*) and snakeweed (*Cleistogenes* sp.); and cold wormwood. There are herb meadows with rhizome Mongolian onions and herb salt marshes with sparse beds of bushlike *Caragana.* In the Gobi Altai and other high mountains, desert-grass steppes completely cover the lower slopes, and, on the upper parts, mountain versions of the feather-grass steppes appear.

Areas with fairly rich vegetation

Animal life. Animal life is varied, with such large mammals as the wild camel, the ass kulan (*Equus asinus ferus*), the dzheiran gazelle, and the dzeren (an antelope). Przewalski's horse, which once ranged in the western region of the desert, is probably extinct in the wild. Rodents include marmots and gophers, and there are reptiles.

The people and economy. The population density is small—fewer than three persons per square mile—mostly Mongols with Chinese in Inner Mongolia. The Chinese have increased greatly in recent years. The main occupation of the inhabitants is nomadic cattle raising, though, in regions where the Chinese are concentrated, agriculture is predominant. The living quarters of the Mongol nomads are felt yurts (types of tent), while the Chinese farmers live in *fanzy,* clay homes built from crude brick.

The province of the Gobi and its semidesert sections is mainly a livestock region, sheep and goats constituting more than half of the total herds. Next come the large-horned cattle. Horses make up only a small percentage of the total and, together with the large-horned cattle, are concentrated in the lusher semidesert of the southeastern region. A fair number of the livestock still consists of two-humped camels, still used for transportation in some areas. Pasturage for cattle is available all year round because of underground waters. Livestock raising is nomadic, and herds move ten times a year, migrating as much as 120 miles between extreme points.

Useful mineral deposits are scant, but there are pockets of petroleum around the city of Saynshand (about 250 miles southeast of Ulaanbaatar), and, in the Mongolias as a whole, salt and light metal ores are mined. Agriculture is developed only along the river valleys.

Natural resources

The Gobi is intersected by railroads in the east and west. There are several highways: from the town of An-hsi to the town of Ha-mi across Pei Shan-mo and the Ka-shun Gobi, from the town of Kalgan (northwest of Peking) to Ulaanbaatar, and from Ulaanbaatar to Dalandzadgad (some 300 miles south-southwest of Ulaanbaatar). In addition, various ancient caravan tracks crisscross the Gobi in all directions. (M.P.Pe.)

KARA-KUM

Kara-Kum (Peski Karakumy in Russian, in the transliteration system of the Akademiya Nauk; Russianized Turkmen *karakumy,* "black sand"; *i.e.,* with vegetation) is the name given to a great sandy desert comprising about 60 percent of the area of the Turkmen Soviet Socialist Republic of the U.S.S.R. Another, smaller desert in the Kazakh S.S.R. near the Aral Sea is called the Aral Kara-Kum.

Physical features. *Physiography.* The Turkmen Kara-Kum is approximately 115,000 square miles (300,000 square kilometres) in area. It is bordered on the north by the Sary-Kamysh Depression, on the northeast and east by the Amu Darya Valley, and on the southeast by the Karabil Highlands and Badkhyz semidesert. In the south and southwest the desert runs along the foot of the Kopet-Dag Mountains, and in the west and northwest it borders the course of the ancient valley of the Uzboy River. It is divided into three parts: the elevated northern Unguz Kara-Kum; the central, low-lying Tsentralny Kara-Kum; and the southeastern Kara-Kum, through which runs a chain of salt marshes. Across the border of the Unguz and Tsentralny Kara-Kum there runs the Unguz chain of saline, isolated, wind-eroded hollows.

The relief of the Turkmen Kara-Kum is quite sharp and clear-cut and reflects its origin and historical development. The surface of the Unguz Kara-Kum has been eroded by violent winds. The plain of the Central Kara-Kum runs from the Amu Darya to the Caspian Sea along the same incline as the river. The height of wind-accumulated, half-overgrown, sand ridges ranges from five to 100 feet (1.5–30 metres), depending on age and wind velocity. Somewhat less than 10 percent of the area consists of crescent-shaped dunes (*barkhany*), some of them 250 feet or more in height. There are numerous hollows of clay deposits, called *takyr,* and saline land formed by the evaporation of subsoil waters.

Climate. The climate of the Turkmen Kara-Kum is continental, with long, hot summers and unpredictable but warm winters. The average temperature in July in the north and along the shore of the Caspian Sea ranges from 79° to 82° F (26° to 28° C), and in the central part of the Central Kara-Kum from 86° to 90° F (30° to 34° C). In January, average temperatures are 25° F (−4° C) in the north and 39° F (4° C) in the south, but temperatures may fluctuate from as low as −4° F (−20° C) to 97° F (36° C) within a 24-hour period. The average annual rainfall varies from 2.75 inches (70 millimetres) in the north to six inches (150 millimetres) in the south. Precipitation occurs mainly in autumn and early spring, more than half of it

Climate of the Kara-Kum

between November and April. There is little snow. The prevailing northeasterly and northwesterly winds are mild.

Plant and animal life. The vegetation is quite varied, consisting mainly of grass, small shrubs, bushes, and trees. The humid early spring permits the widespread growth of ephemeral plants—the main animal fodder—while in the barchan dune areas the typical vegetation consists of cereals, the wormwood shrub, and trees of the species *Ammodendron conollyi.* The most common bushes are species of *Astragalus, Calligonum,* and the salt-tree (*Salsola richteri*). In regions of deep underground water, the white haloxylon is the most typical plant, but in regions where water is nearer to the surface, the black haloxylon occurs. The vegetation of the Turkmen Kara-Kum can be used as hay in winter by camels, sheep, and goats.

Animals are not numerous, but they are of many kinds. The insects include ants, termites, ticks, beetles, darkling beetles, dung beetles, and spiders. Various species of lizards, snakes, and turtles also occur. The most common birds are skylarks, haloxylon (sacksal) jays, wagtails, and desert sparrows. Among the rodents are gophers and jerboas. The tolai hare, the hedgehog, barchan cat, and corsac fox and also the dzheran (a gazelle) usually live near *takyr* soils.

The people and economy. The population is composed of Turkmen, among whom some tribal distinctions have been preserved. Formerly nomadic, they settled into agricultural pursuits or fishing on the shores of the Caspian. Collective and state farms have developed permanent settlements, with gas and electricity. Cattle-raising brigades care for the socialized livestock. The development of oil, gas, and other industries has brought new settlements, populated by diverse nationalities.

Irrigation and economic development

Irrigation has made the desert suitable for the raising of livestock on a large scale, especially astrakhan sheep. The Karakum Canal has brought water to Southeastern Kara-Kum, the southern border of the Tsentralny Kara-Kum, and along the foot of the Kopet-Dag Mountains, where fine-fibred cotton is now grown in the oasis areas. The proposed extension of the canal to the shore of the Caspian Sea will permit the cultivation of the subtropical arid regions.

Intensive economic development after World War II brought an industrial revolution to the Kara-Kum. Factories, gas lines, railroads, and highways have changed the face of the region. Hydroelectric stations are under construction at Takhiyatas and Tyuyamuyun. (B.A.F.)

TAKLA MAKAN

A desert of Central Asia and one of the largest sandy deserts in the world, the Takla Makan (Takolamagan in Chinese Pinyin) occupies the central part of the Tarim Basin in China. The desert area extends about 600 miles (960 kilometres) from west to east, and it has a maximum width of 260 miles and a total area of about 105,000 square miles (272,000 square kilometres). The desert reaches altitudes of 3,900 to 4,900 feet (1,200 to 1,500 metres) in the west and south and from 2,600 to 3,300 feet in the east and north.

Physical features. The Takla Makan is flanked by high mountain ranges: the Kunlun in the south, the Tien Shan in the north, and the Pamirs in the west. There is a gradual transition to the marshy Lop Nor (Lop Lake) Basin in the east, and, in the south and west, between the sandy desert and the mountains, lies a band of sloping desert lowland composed of pebble-detritus deposits.

Physiography. Several small mountain ranges and chains, composed of sandstones and clays from the Paleogene and Neogene periods (10,000 to 65,000,000 years ago), rise in the western part of the desert. The arc-shaped Mazar Tagh (Mazar Mountains), located between the Ho-t'ien Ho (Ho-t'ien River) or Khotan Darya (Khotan River) and the Yarkand River valleys, arch toward the southwest. Ninety miles long and from two to three miles wide, with a maximum height of 5,363 feet, they rise an average of only 1,000 to 1,150 feet above the surface of the sandy plain. The Chiao-lo Shan (Chöl Tagh) is also an insular range, surrounded on all sides by massifs of moving sands. Rosstagh, also known as Tokhtakaz Mountain, its highest peak, is 5,117 feet (1,560 metres), and the range rises from 600 to 800 feet above the plain. Both ranges are covered by a shallow mantle of eluvium and rock debris and have sparse, desert-type vegetation. In the north, the sands of the Takla Makan form a clear boundary with the vegetated Tarim River Valley.

The general slope of the plain is from south to north, and thus the rivers running off from the Kunlun Mountains flow in this direction. The Ho-t'ien and K'o-li-ya (Keriya) river valleys have survived up to the present, but a majority of the shallower rivers have been lost in the sands, after which their empty valleys fill up with windborne sand.

Wind-formed surface features

The surface of the Takla Makan is composed of friable alluvial deposits several hundred feet thick. This alluvial stratum has been affected by the wind, and its wind-borne sand cover is as much as 1,000 feet thick. There is a variety of eolian (wind-formed) topographic features, and sand dunes of various shapes and sizes are encountered. These eolian sand dunes were formed through the weathering of the alluvial and colluvial deposits of the basin and of the foothill plains of the Kunlun and eastern Tien Shan. The size of the larger sand-dune chains is considerable: they range from 100 to 500 feet high and 800 to 1,650 feet wide, with a distance between the chains of from one-half to three miles. The highest eolian topographic forms are the pyramidal dunes, rising 650 to 1,000 feet. In the eastern and central parts of the desert, networks of hollow dunes and large, complex sand-dune chains predominate. They are also common in the western portion of the desert (east of the Ho-t'ien Ho Valley), where transverse and longitudinal (with respect to the wind) topographic forms coexist. On the edge of the desert, semipermanent, clustered sand dunes with tamarisk and niter-bushes, as well as clayey regions with disconnected sand dunes, predominate. Such a diversity in wind-formed features is a result of the complex wind conditions of the basin.

Climate. The Takla Makan's climate is moderately warm and markedly continental, with a maximum annual temperature range of 70° F (21° C). Precipitation is very low, ranging from 1.5 inches (38 millimetres) in the west to 0.4 inch in the east. The air temperature in the summer is high, rising as high as 100° F (38° C) on the eastern edge of the desert. In July the average air temperature reaches 77° F (25° C) in the eastern regions. Winters are cold: in January the average air temperature is 16° to 14° F (−9 to −10° C), and the low in winter is generally below −4° F (−20° C). Northerly and northwesterly winds predominate in the summer in the western region. These two air currents, on meeting near the desert's centre at the end of the K'o-li-ya Ho, create a complex circulation system that is clearly reflected in the topography of the sand dunes. In the spring, when the surface sand becomes warm, ascending currents develop, and northeasterly winds become particularly strong. During this period hurricane-force dust storms, filling the atmosphere with dust to altitudes of from 11,500 to 13,000 feet, often occur. Winds from other directions also raise clouds of dust into the air, covering the Takla Makan with a shroud for almost the entire year.

Hydrography. Since the Tarim depression is an internal-drainage basin, the entire runoff from the surrounding mountains collects in the basin itself, feeding the rivers and the groundwater strata. In all probability, the groundwater table under the sands has an unbroken course, flowing west to east to Lop Nor. The importance of rainfall in moistening the sands and feeding the groundwaters is slight, due to its small quantity and high rate of evaporation. The rivers draining the Kunlun Mountains penetrate from 60 to 120 miles into the desert, gradually drying up in the sands. Only the Ho-t'ien Ho crosses the desert and, in July and August, occasionally carries its waters to the Tarim River.

The groundwater table

Plant and animal life. Vegetation is very sparse in the Takla Makan—almost the entire territory is devoid of plant cover. In depressions among the sand dunes, where the groundwaters lie no deeper than 10 to 16 feet from the surface, thin thickets of tamarisk, niter-bushes, and reeds may be found. The thick strata of moving sands, however, prevent the wider spread of this vegetation. Along the edges of the desert, in the area where the sand dunes meet

The Takla Makan Desert.

the river valleys and deltas and where the groundwaters lie comparatively close to the surface, the vegetation is richer. Here, in addition to the plants mentioned above, a number of species characteristic of river valleys are found: Turanga poplar, oleaster, camel thorn, Zygophyllaceae (caltrop family), and saltworts. Sand dunes in hummocks frequently form around the scrub.

The animal life of the Takla Makan is sparse. Only in peripheral regions of the desert, in ancient and modern river valleys and deltas where water and vegetation appear, are the fauna more diverse. Herds of gazelles are found in open spaces, and there are wild boars in river valley thickets. Wolves and foxes are among the carnivores. As late as the beginning of the 20th century, tigers could be found, but they have since been exterminated. Rare animals include the Siberian deer, which inhabits the Tarim River Valley, and the wild camel, which at the end of the 19th century roamed over a considerable portion of the Takla Makan as far as the Ho-t'ien Ho but now appears only occasionally, in the eastern desert region.

There are a large number of rabbits, gerbils, field mice, and jerboas in the sand dunes, and, among insectivores, the long-eared hedgehog and bats are common. Small, tufted larks and the Tarim jay are the commonest birds.

The people and economy. There is no fixed population in the Takla Makan. Hunters make periodic visits, but the area territory is not used by stock breeders because of the almost complete absence of vegetation.

Indications of the presence of petroleum have recently been found in the two regions—Kashgar–Khargalik at the western part of the Tarim Basin and Aksu Kucha at the northern edge of the basin. (M.P.Pe.)

Drainage systems and waterways

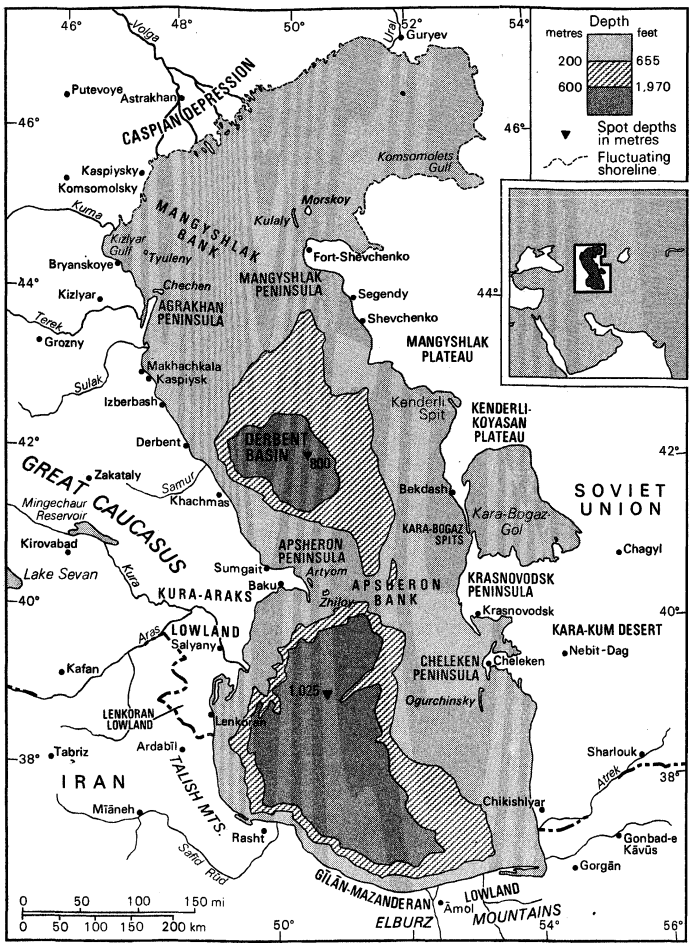
The subsequent sections on the various drainage systems and waterways to be found in Asia are arranged in the following order: Southwest Asia, South Asia, East and Southeast Asia, and Siberia.

CASPIAN SEA

The Caspian Sea lies beyond the Caucasian mountain peaks and dominates the huge, flat expanses of western Central Asia. Its name derives from the ancient Kaspi peoples, who once lived in Transcaucasia to the west; among its other historical names, Khazarsk and Khvalynsk derive from former peoples of the region, while Girkansk stems from Girkanos, "the country of the wolves." Its elongated shape sprawls for nearly 750 miles (1,200 kilometres) from north to south, although its average width is only 200 miles (320 kilometres). It currently covers an area of 143,000 square miles (370,000 square kilometres)—larger than Japan—while its surface lies some 93.5 feet (28.5 metres) below ocean level. The maximum depth, toward the south, is 3,360 feet. Except for its southern shores, which lie at the foot of the giant Elburz Mountains of Iran six-sevenths of the Caspian coast runs through Soviet territory.

It is often stated that the Caspian is the greatest salt lake in the world. But this is not absolutely correct, as scientific studies have shown that, until geologically quite recent times, it was linked, via the Sea of Azov, the Black Sea, and the Mediterranean, to the world ocean. This factor has molded strongly all aspects of its physical geography. The Caspian is of exceptional scientific interest, because its history, particularly in respect to former fluctuations in both area and depth, offers clues to the complex geological and climatic evolution of the region. Man-made changes, notably those resulting from the construction of dams, reservoirs, and canals on the immense Volga River system (which drains into the Caspian from the north), have had their own effect on the contemporary hydrological balance. The Caspian is also of great importance in the transportation networks of the region and in the

Importance of sea-level fluctuations



The Caspian Sea.

production of petroleum and gas. It is anticipated that the use of its splendid sandy beaches for health and recreation resorts will increase considerably in the future.

Physical features. The Caspian possesses as many as 50 islands, mostly small. Chechen, in the northwest, is the largest, followed by Artyom, Zhiloy, Tyuleny, Morskoy, Kulaly, and Ogurchinsky. The basin, as a whole, is usually divided into the North, Middle, and South Caspian, using relief and, in part, hydrological characteristics as the basis.

Shoreline features. The shores of the North Caspian are low and reflect the great accumulation of alluvial material washed down by the Ural, Terek, Sulak, and, above all, the Volga rivers, the deltas of which are extensively developed. The western shore of the Middle Caspian is hilly. The foothills of the Great Caucasus loom close but are separated from the coast by a narrow marine plain. The Apsheron Peninsula, on which the city of Baku is sited, thrusts out into the sea here, while, just to its south, the floodplain of the Kura and Araks rivers forms the Kura-Araks Lowland. The southwestern and South Caspian shores are formed of the sediments of the Lenkoran and Gīlān-Mazanderan lowlands, with the high peaks of the Talish and Elburz mountains rearing up close inland. The eastern shore of the South Caspian is also low and is formed by sediments resulting from wave action; it is broken sharply by the low, hilly Cheleken and Krasnovodsk peninsulas. For the most part, the eastern shore of the Middle Caspian is precipitous, with the sea destroying the margin of the limestone plateau of Mangyshlak and Kenderli-Koyasan. A most important feature of this area is the Kara-Bogaz-Gol, the Caspian's largest, split off from the main body of water by the Kara-Bogaz barrier spits; the Krasnovodsk and Kenderli spits are similar in type.

The major rivers—the Volga, Ural, and Terek—empty into the North Caspian, with their combined annual flow accounting for 88 percent of all river water entering the sea. The Sulak, Samur, Kura, and a number of smaller rivers flow in on the western littoral, contributing about 7 percent of the flow, and the remainder comes in from the rivers of the Iranian shore. The eastern littoral is notable for a complete lack of permanent streams.

Submarine features. The North Caspian, with an area of 30,000 square miles, is the shallowest portion of the sea, with an average depth of 13 to 20 feet, reaching a maximum of 66 along the boundary with the Middle Caspian. The bottom is formed of a monotonously rippling sedimentary plain, broken only by a line of southern bars and shoals—some of which provide foundations for Tyuleny, Kulaly, and Zhemchuzhny islands—reflecting underlying structural rises. Beyond this belt, known as the Mangyshlak Bank, the Middle Caspian, 53,000 square miles in area, forms an irregular depression with an abrupt western slope and a gentler eastern gradient. The shallowest portion—a shelf with depths up to 330–460 feet—extends along both shores, with the westernmost slope furrowed by submerged landslips and canyons. The remains of ancient river valleys have been discovered on the gentler northern slope; the bottom of the depression is formed of a plain that deepens to the west. The Apsheron Bank, a belt of shoals and islands rising from submerged elevations of older rocks, marks the transition to the South Caspian, a depression covering about one-third of the sea. This is fringed by a shelf that is narrow to the west and south but widening to the east. A series of submerged ridges breaks up the relief to the north, but otherwise the bottom of the depression is a flat plain, with the maximum depth of 3,360 feet.

The three main topographic divisions

Geology. The relief of the Caspian Sea reflects its complex geological structure. The northern portion forms a peripheral section of the North Caspian tectonic depression, a vast downwarp in the Earth's surface that is itself part of the great ancient structural block known as the Russian Platform. The Mangyshlak Bank links the mountainous peninsula of the same name to the east with underlying western shore structures; all the rocks concerned reflect an outlying structural uplift of the Hercynian mountain-building movement, which occurred some 290,000,000 years ago. The bottom of the Middle Caspian is very complex. To the west, the submarine shelf is part of the sagging edge of the Great Caucasus Geosyncline, while the submerged Turan Platform on the east swells up in the feature known as the Kara-Bogaz swell. The Apsheron Peninsula regions bear the impress of the Alpine mountain-building and folding period (dating from some 10,000,000–26,000,000 years ago), which created the Caucasian ranges, as do the folded structures of the western South Caspian depression. The entire South Caspian, in fact, has as its base a suboceanic-type basalt crustal structure, although this is overlain in the south by huge accumulations of sedimentary layers many miles thick. The great age thus indicated for the area has led geologists to postulate that this is the remnant of the geologically ancient Tethys Sea. Similarly, the North Caspian Sea bottom is also very old, dating back to Precambrian times, or at least 570,000,000 years ago. The bottom of the North and Middle Caspian has a continental-type crustal structure. It has been suggested that the Middle Caspian depression results from a sagging at the edge of these ancient structures that occurred in Late Paleozoic times, about 250,000,000 years ago; the border between the Middle and South Caspian is, in fact, still experiencing folding activity.

Until the upper Miocene, about 12,000,000 years ago, the sea basin of the Caspian was connected to the Black Sea, through the structural depression known as the Manych Trench. After an upper-Miocene elevation, the Caspian became an enclosed body, with oceanic submarine characteristics preserved today only in the South Caspian. The ocean connection was reestablished in the upper Pliocene, about 2,500,000 years ago, and it is possible that there was also a link northward across the Russian plain to the Barents Sea of the Arctic.

Pre-Miocene link with the Black Sea

In the ensuing Quaternary times, great glaciers advanced and retreated across the Russian plain, and the Caspian Sea itself—in successive phases known as Bakin, Khazar, and Khvalyn—alternately shrank and swelled out. This process left a legacy in the form of peripheral terraces,

marking old shorelines, and can also be traced in the recent underlying sedimentary layers.

The Caspian Sea bottom is now coated with young sediments, finely grained in the shallow north but with shell deposits and oolitic sand—reflecting the high lime content of the Caspian waters—widespread in other coastal areas. Lime also affects the composition of the much deeper bottom layers.

Climate. The North Caspian lies in a moderately continental climatic zone, while all the Middle (and most of the South) Caspian lies in the moderately hot belt. The southwest is touched by subtropical influences, and this remarkable variety is completed by the desert climate prevailing on the eastern shores. Atmospheric circulation is dominated in winter by the cold, clear air of the Asiatic anticyclone, while in summer spurs of the Azores high-pressure and the South Asian low-pressure centres are influential. Complicating factors are the cyclonic disturbances rippling in from the west and the effect of the Great Caucasian ranges. As a result of these factors, northwesterlies (32 percent of occurrences) and southeasterlies (36 percent) dominate circulation patterns. Savage storms are associated with northerly and southeasterly winds.

Summer air temperatures are quite evenly distributed (average July–August figures: 75°–79° F [24°–26° C], with an absolute maximum of 111° F [44° C] on the sunbaked eastern shore), but winter temperatures range from 14° F (−10° C) in the north to 50° F (10° C) in the south. Average annual rainfall varies from 67 inches to eight inches (1,700 to 200 millimetres) over the sea. Most falls in winter and spring. Evaporation from the sea surface is very high, reaching 40 inches a year. Ice formation afflicts the North Caspian, which usually freezes completely by January, and in very cold years floating ice comes as far south as the Apsheron Peninsula region.

Hydrography. Short-term wind-induced fluctuations in the sea level can rise to up to seven feet, though such rises average about two feet. Seiches (rises induced by barometric pressure changes) can cause similar fluctuations. Tidal changes are but a few inches, and seasonal rises induced by high spring water in the rivers are not much more.

One of the more fascinating aspects of the study of the Caspian, however, is the reconstruction of long-term fluctuations over the centuries from archaeological, geographical, and historical evidence. It seems the Caspian reached a level of 72 feet below sea level about 4,000 to 6,000 years ago and again early in the 19th century AD. A still lower level held from the 7th to the 11th centuries, while the lowering that took place between 1929 and 1957 stemmed from the effects of climatic change resulting in lesser river influx and increased evaporation amplified by reservoir construction on the Volga, and from river water consumption for irrigation and industry. The flow of water into the Kara-Bogaz-Gol, now about 12 feet lower than the Caspian, has also had an effect. By the late 20th century, the water level was very close to the −93.5 feet level, reflecting a balance between input (rainfall, river inflow, subterranean upwelling) and consumption (evaporation, flow into the Kara-Bogaz-Gol, human usage) that gave the latter a slight edge and hence a projected annual lowering of the level by three inches (7.5 centimetres). If, however, the north-flowing Vychegda and Pechora were diverted into the Volga, it would seem that the present level could at least be maintained (and possibly even increased, under favourable climatic conditions) until the year 2000. Soviet planners have given serious attention to the requirements and implications of such an ambitious project.

Recent attempts to preserve the sea level

In summer, the average seawater temperature is 75°–79° F (24°–26° C), with the south a little warmer. There are, however, significant winter contrasts, from 37°–45° F (3°–7° C) in the north to 46°–52° F (8°–11° C) in the south. Upwellings of deep water at the eastern littoral—a result of prevailing-wind activity—can also bring a marked drop in summer temperature. Salinity in the Caspian is about 1.27 percent on average, but this conceals a variation from a mere 0.1 percent near the Volga outlet to a high of 32 percent in the Kara-Bogaz-Gol, where intense evaporation occurs. Caspian waters differ from those of the ocean in their high sulphate, calcium, and magnesium carbonate content and—as a result of river inflow—low chloride content.

Water mass circulation occurs, basically, in a north to south movement along the western shore, with a complex pattern developing further south, where there are several subsidiary movements. Currents can be speeded up where they coincide with strong winds, and the sea surface is often ruffled by wave action, with the maximum storm waves being observed near the Apsheron Peninsula.

Marine life. There are about 850 animal and more than 500 plant species represented in the Caspian—a relatively low figure for a body of water of this size. Animal life has been affected greatly by changes in salinity. It includes, among the fish, sturgeons, herring, pike, perch, and sprat; several species of mollusks; and a variety of other organisms including sponges. Some 15 species of Arctic (*e.g.*, the Caspian seal) and Mediterranean types complement the basic fauna. Perch are important among freshwater fish varieties. Some organisms have migrated to the Caspian quite recently: barnacles, crabs, and clams, for example, have been transported by sea vessels; grey mullets have been deliberately introduced by man.

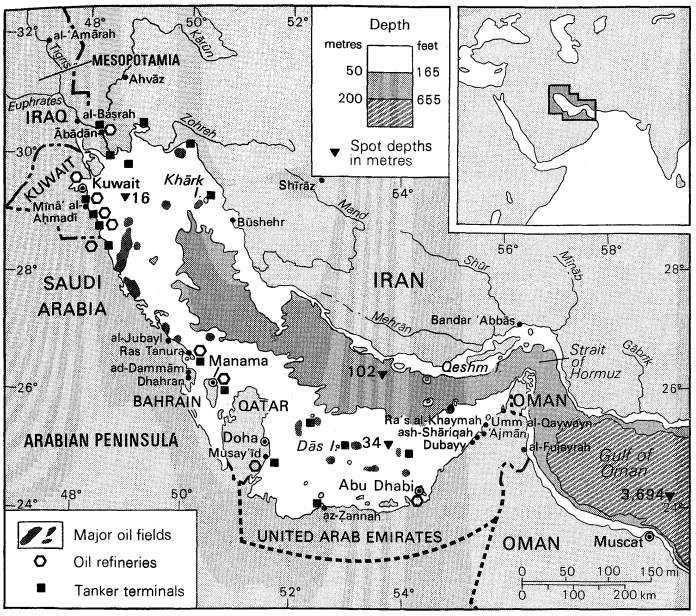
The economy. The Caspian was long famous for its sturgeon catch, but this has been reduced greatly in recent years, as a result of the decline in sea level, and the connected drying up of the most favourable places for spawning. The seal industry is, however, being developed in northern regions. Oil and gas have now become the region's most important resources, following extensive geological surveys in the 1940s and 1950s. Seabed oil is extracted from derricks and artificial islands, most of which are concentrated off the shores of the Azerbaijan S.S.R., supplying half that republic's total oil extraction volume. The extraction of such minerals as sodium sulphate from the Kara-Bogaz-Gol is also of considerable economic importance. Finally, the Caspian is of major importance for transportation in the region: petroleum, wood, grain, cotton, rice, and sulphate are the basic goods carried, while Astrakhan, Baku, Makhachkala, Krasnovodsk, and Shevchenko are the most important ports. They are also connected by regular passenger runs, while railway stock is transported direct, without unloading, on the Baku–Krasnovdsk run. (A.N.K./O.K.L.)

PERSIAN GULF

The Persian Gulf (known to the Arabs as the Arabian Gulf) is the shallow marginal sea of the Indian Ocean that lies between the Arabian Peninsula and southwest Iran. The sea has an area of 92,500 square miles (240,000 square kilometres). Its length is 615 miles (985 kilometres), and its width varies from a maximum of 210 miles to a minimum of 35 miles in the Strait of Hormuz. It is bordered on the north, northeast, and east by Iran, on the northwest by Iraq and Kuwait, on the west and southwest by Saudi Arabia, Bahrain, and Qatar, and on the south and southeast by the United Arab Emirates and part of Oman. The term Persian Gulf is often used to refer not only to the Persian Gulf proper but also to its outlets, the Strait of Hormuz and the Gulf of Oman, which open into the Arabian Sea. The discussion here deals with the Persian Gulf alone.

Physical features. *Physiography.* The Iranian shore is mountainous, and there are often cliffs; elsewhere a narrow coastal plain with beaches, intertidal flats, and small estuaries borders the gulf. The coastal plain widens north of Būshehr on the eastern shore of the gulf and passes into the broad deltaic plain of the Tigris, Euphrates, and Kārūn rivers. Cliffs are rare on the Arabian shore of the gulf, except around the base of the Qatar Peninsula and in the extreme southeast around the Strait of Hormuz. Most of the Arabian shore is bordered by sandy beaches, with many small islands enclosing small lagoons.

The gulf is very shallow, rarely deeper than 300 feet (90 metres), although depths exceeding 360 feet are found at its entrance and at isolated localities in its southeastern part. It is noticeably asymmetrical in profile, with the deepest water occurring along the Iranian coast and a broad shallow area, which is usually less than 120 feet deep, along the Arabian coast. There are numerous is-



The Persian Gulf.

lands, some of which are salt plugs or domes and others merely accumulations of coral and skeletal debris.

The Persian Gulf receives very small amounts of river-borne sediment except in the northwest, where the Tigris, Euphrates, and Kārūn, together with other small streams, empty into the gulf. The rivers reach their peak flow in spring and early summer, when the snow melts in the mountains; disastrous floods sometimes result. There are some ephemeral streams on the Iranian coast south of Būshehr, but virtually no fresh water flows into the gulf on its southwest side. Large quantities of fine dust are, however, blown into the sea by predominant northwest winds from the desert areas of the surrounding lands. Biological, biochemical, and chemical processes lead to the production of considerable calcium carbonate in the form of skeletal debris and fine mud. The deeper parts of the Persian Gulf adjacent to the Iranian coast and the area around the Tigris–Euphrates Delta are mainly floored with gray-green muds rich in calcium carbonate. The shallower areas to the southwest are covered with whitish-gray or speckled skeletal sands and fine carbonate muds. Often the sea floor has been hardened and turned to rock by the deposition of calcium carbonate from the warm, salty waters. Chemical precipitation is abundant in the coastal waters, and sands and muds are produced that mix with the skeletal debris of the local sea life. These sediments are thrown up by the waves to form coastal islands that enclose lagoons. The high salinities and temperatures result in the precipitation of calcium sulfate and sodium chloride to form extensive coastal salt flats.

Seafloor

Geology. The present-day Persian Gulf, together with its northwestern continuation now infilled by the deposits of the Mesopotamian rivers, is thought by some to be the remains of a once much larger basin of deposition aligned northwest to southeast that existed throughout a large part of geological history. In this basin many thousands of feet of sediments accumulated, consisting mostly of limestone and marls (loose, crumbling earthy deposits containing calcium carbonate), together with evaporites and organic matter, which ultimately produced the area's vast oil resources.

Climate. The gulf has a notoriously unpleasant climate. Temperatures are high, though winters may be quite cool at the northwestern extremities. The sparse rainfall occurs mainly as sharp downpours between November and April and is higher in the northeast. Humidity is high. The little cloud cover is more prevalent in winter than in summer. Thunderstorms and fog are rare, but dust storms and haze occur frequently in summer. The wind that blows predominantly from a north–northwest direction—the so-called *shamāl* (Arabic: "north wind")—is seldom strong and rarely reaches gale force. Squalls and waterspouts are common in autumn, when winds sometimes reach speeds of 95 miles per hour in as little as five minutes. Intense heating of the land adjacent to the coasts leads to gentle offshore winds in the mornings and strong onshore winds in the afternoons and evenings.

Hydrography. The small freshwater inflow into the gulf is mostly from the Tigris–Euphrates–Kārūn rivers. The water temperatures range from 75° to 90° F (24° to 32° C) in the Strait of Hormuz to 60° to 90° F (16° to 32° C) in the extreme northwest. These high water temperatures and a low influx of fresh water result in evaporation in excess of freshwater inflow; high salinities result, ranging from 37 to 38 parts per 1,000 in the entrance to 38 to 41 parts per 1,000 in the extreme northwest. Even greater salinities and temperatures are found in the waters of the lagoons on the Arabian shore. The tidal range varies from about four to five feet around Qatar and increases to 10 to 11 feet in the northwest and nine to 10 feet in the extreme southeast. When onshore winds are strong, the level of the coastal waters, particularly in the southern gulf, may rise by as much as eight feet, causing extensive flooding of the low coastal plains. Tidal currents are strong (five miles per hour) in the entrance of the gulf but elsewhere—except between islands or in estuaries and lagoon entrances—rarely exceed one to two miles per hour. The wind affects local currents and sometimes reverses them.

Tides

Waves rarely exceed 10 feet in height and are largest in the southern gulf. The swell from the Indian Ocean affects only the water at the entrance of the gulf; when it is opposed by wind, very turbulent conditions result. The general circulation pattern appears to be caused by water entering from the Indian Ocean, evaporating, becoming denser, and sinking to flow out into the Indian Ocean beneath the inflowing open ocean water.

The waters of the area support many plants and animals, but the high temperatures and salinities lead to a diminution in the variety of forms; many Indian Ocean forms penetrate only a small way into the gulf.

The economy. Until the discovery of oil in Iran in 1908, the Persian Gulf area was important mainly for fishing, pearling, the building of dhows (Arab lateen-rigged boats), sailcloth making, camel breeding, reedmat making, date growing, and the production of other minor products, such as red ochre from the islands in the south. The arid lands surrounding the gulf produced little else and, except for the rich alluvial lands of the Mesopotamian plain, supported only a small population of fishermen, date growers, and nomads. Today these traditional industries have declined, and the economy of the region is dominated by the production of oil. The Persian Gulf and the surrounding countries account for about one-third of the world's total oil production and have more than one-half of the world's proved reserves. The area has been explored only in a preliminary way, and there are still far fewer drill holes in the area than are put down in a single year in the United States. Offshore exploration below the shallow waters of the gulf has revealed the presence of large reserves of oil and gas. These discoveries have led to numerous legal wrangles between states about exact territorial limits. Large quantities of oil are refined locally in Iran, in Bahrain, and elsewhere, but most of it is exported to northwestern Europe and other parts of the world as crude oil.

Oil

Other exploitable mineral deposits appear to be rare, but only cursory surveys have been made. Exploration is actively being pursued in the area.

Fishing is becoming highly commercialized. The traditional pearl-fishing industry has declined since the introduction of the Japanese cultivated pearls in the 1930s. Large fishing industries have been set up in Kuwait, Qatar, and Bahrain, and some countries have become exporters of fish.

There has always been a considerable sea trade carried on by local craft between the Persian Gulf and Africa and India; this is now completely dominated by an incessant flow of large tankers that carry oil from the large marine terminals at Khārk Island, Kuwait, az-Ẓahrān (Dhahran), Bahrain, Musay'īd, az-Zannah, Dās Island, and other locations to all parts of the world. The heavy traffic and

the offshore oil installations have produced many hazards, despite the use of a system of radio-navigational stations. (G.Ev.)

RED SEA

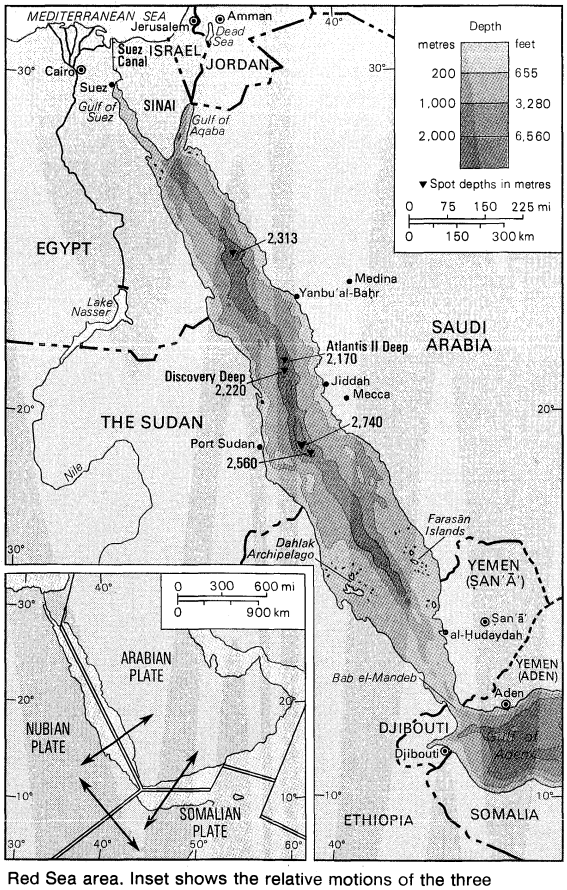
The Red Sea is a narrow strip of water extending southeastward from Suez for about 1,300 miles (2,100 kilometres) to the Straits of Bab el-Mandeb, which connects with the Gulf of Aden and thence with the Indian Ocean. The sea separates the coasts of Egypt, The Sudan, and Ethiopia to the west from those of Saudi Arabia and Yemen (Ṣan'a') to the east. Its maximum width is 190 miles; its greatest depth 9,580 feet (2,920 metres); and its area approximately 169,000 square miles (438,000 square kilometres). The Red Sea contains some of the world's hottest and saltiest seawater. When the Suez Canal is open, it is one of the most heavily travelled waterways in the world, carrying maritime traffic between Europe and Asia. Its name derived from the colour changes observed in its waters. Normally the Red Sea is an intense blue green; occasionally, however, it is populated by extensive blooms of the algae *Trichodesmium erythraeum,* which, upon dying off, turn the sea a reddish-brown colour.

Physiography and submarine morphology. The Red Sea lies in a fault depression that separates two great blocks of the Earth's crust—Arabia and North Africa. The land on either side, inland from the coastal plains, reaches heights of more than 6,560 feet above sea level, with the highest land in the south.

The Gulfs of Suez and Aqaba

At its northern end the Red Sea splits into two parts, the Gulf of Suez to the northwest and the Gulf of Aqaba to the northeast. The Gulf of Suez is shallow—approximately 180 to 210 feet deep—and it is bordered by a broad coastal plain. The Gulf of Aqaba, on the other hand, is bordered by a narrow plain, and it reaches a depth of 5,500 feet. From approximately 28° N, where the Gulfs of Suez and Aqaba divide, south to a latitude near 25° N, the Red Sea's coasts parallel each other at a distance of about 110 miles apart. Here the sea floor consists of a main trough, with a maximum depth of some 4,100 feet, running parallel to the shorelines.

Inset adapted from D.P. McKenzie, D. Davies, and P. Molnar, "Plate Tectonics of the Red Sea and East Africa," *Nature,* vol. 226, p. 244 (April 18, 1970)



Red Sea area. Inset shows the relative motions of the three plates that make up the Red Sea area.

South of this point, and continuing southeast to latitude 16° N, the main trough becomes sinuous, following the irregularities of the shoreline. About halfway down this section, roughly between 20° and 21° N, the topography of the trough becomes more rugged, and several sharp clefts appear in the sea floor. Because of an extensive growth of coral banks, south of 16° N only a shallow narrow channel remains. The sill separating the Red Sea and the Gulf of Aden at Bab el-Mandeb is raised by this growth; therefore, the depth of the water is only about 380 feet, and the main channel becomes very narrow.

The clefts within the deeper part of the trough are unusual sea floor areas in which hot brine concentrates are found. These patches apparently form distinct and separated deeps within the trough having a north-south trend, whereas the general trend of the trough is from northwest to southeast. At the bottom of these areas are unique sediments, containing deposits of heavy metal oxides from 30 to 60 feet thick.

Most of the islands of the Red Sea are merely exposed reefs. There is, however, a group of active volcanoes just south of the Dahlak Archipelago (16° N), as well as a recently extinct volcano on Jebel Tier.

Geology. The Red Sea occupies part of a large rift valley in the continental crust of Africa and Arabia. This break in the crust is part of a complex rift system that includes the East African Rift valley, which extends southward through Ethiopia, Kenya, and Tanzania for almost 2,200 miles, and northward for over 280 miles from the Gulf of Aqaba to form the great Wadi Aqaba–Dead Sea–Jordan Rift, also extending eastward for 600 miles from the southern end of the Red Sea to form the Gulf of Aden.

Relationship to the Great Rift System

The Red Sea Valley cuts through the Arabian-Nubian Massif, a central mass of Precambrian igneous and metamorphic rocks (*i.e.,* formed deep in the earth under heat and pressure from 570,000,000 to 4,600,000,000 years ago), whose outcrops form the rugged mountains of the adjoining region. The massif is surrounded by Paleozoic marine sediments (from 225,000,000 to 570,000,000 years old). These sediments were affected by the folding and faulting that began in the Late Paleozoic Era; the laying down of deposits, however, continued to take place during this time and apparently continued into the Mesozoic Era (from 65,000,000 to 225,000,000 years ago). The Mesozoic sediments appear to surround and overlap those of the Paleozoic and are in turn surrounded by Early Tertiary sediments (from about 54,000,000 to 65,000,000 years old). In many places large remnants of Mesozoic sediments are found overlying the Precambrian rocks, suggesting that at one time a fairly continuous cover of deposits existed above the older massif.

The Red Sea is considered a relatively new sea whose development probably resembles the Atlantic Ocean in its early stages. The Red Sea's trough apparently formed in at least two complex phases of land motion. The movement of Africa away from Arabia began during the lower Eocene Epoch (50,000,000 years ago). The Gulf of Suez opened up during the lower Oligocene Epoch (about 35,000,000 years ago), and the northern part of the Red Sea in early Miocene time (25,000,000 years ago). The second phase began about 3,000,000 to 4,000,000 years ago, creating the trough in the Gulf of Aqaba and also in the southern half of the Red Sea Valley. This motion, estimated as amounting to 0.59 to 0.62 inches (15–15.7 millimetres) a year, is still proceeding, as indicated by the extensive volcanism of the past 10,000 years, by earthquake activity, and by the flow of hot brines in the trough.

Climate. The Red Sea region has very little precipitation in any form, although prehistoric artifacts seem to indicate that there were greater amounts of rainfall in times gone by. In general, the year-round climate makes an active life difficult, for the average temperature varies between 77° and 82° F (25° and 28° C), and there is a very high degree of relative humidity in summer. In the northern part of the Red Sea area, extending down to 19° N, the prevailing winds are north to northwest. Best known are the occasional westerly, or "Egyptian," winds,

which blow with some violence during the winter months and are generally accompanied by fog and blowing sand. From latitudes 14° to 16° N the winds are variable, but during the months of June through August strong northwest winds move down from the north, sometimes extending as far south as the Straits of Bab el-Mandeb; by September, however, this wind pattern retreats to a position north of 16° N. South of 14° N the prevailing winds are south to southeast.

Hydrography. No water enters the Red Sea from rivers, and rainfall is scant; but the evaporation loss, in excess of 80 inches a year, is made up by an inflow through the eastern channel of the Straits of Bab el-Mandeb from the Gulf of Aden. This inflow is driven toward the north by prevailing winds and generates a circulation pattern in which these low salinity waters (the average salinity is about 36 parts of salt per thousand) move northward. Water from the Gulf of Suez has a salinity of about 40 parts per thousand due to evaporation, and consequently a high density. This dense water moves toward the south and sinks below the less dense inflowing waters from the Red Sea. Below a transition zone, which extends from depths of about 300 to 1,300 feet, the water conditions are stabilized at about 72° F (22° C), with a salinity of almost 41 parts per thousand. This south-flowing bottom water, displaced from the north, spills over the sill at Bab el-Mandeb, mostly through the eastern channel. It is estimated that there is a complete renewal of water in the Red Sea every 20 years.

Salinity and temperature

Below this southward-flowing water, in the deepest portions of the trough, there is another transition layer, only 80 feet thick, below which, at some 6,400 feet, lie a number of pools of hot brine. The brine in the Atlantis II Deep has an average temperature of almost 140° F (60° C), a salinity of 256 parts per thousand, and contains no oxygen. There are similar pools of water in the Discovery Deep, and in the Chain Deep (at about 21°18′ N). Heating from below renders these pools unstable so that their contents mix with the overlying waters; they thus become part of the general circulation system of the sea.

The economy. *Resources.* Three major types of mineral resources are found in the Red Sea region: petroleum deposits, evaporite deposits (sediments laid down as a result of evaporation, such as salt, gypsum, and dolomite), and the newly discovered heavy metal deposits in the bottom oozes of the Atlantis II and Discovery deeps, which lie between 21°15′ and 21°30′ N. The oil and gas deposits are being exploited to varying degrees by the nations adjoining the sea. The evaporites are utilized only very slightly, primarily on a local basis. Of the heavy metal deposits, none of which had been touched, those contained in the sediments of the Atlantis II Deep alone were estimated as having a $25,000,000,000 value. The sediment of the Discovery Deep also has a significant metalliferous content but at a lower concentration than that in the Atlantis II Deep. These deposits are in the form of fairly fluid oozes, with an average of about 85 percent brine. The average analysis of the Atlantis II Deep deposit reveals an iron content of 29 percent; zinc 3.4 percent; copper 1.3 percent; lead 0.1 percent; silver 54 parts per million; and gold 0.5 parts per million. The total brine-free sediment estimated to be present in the upper 30 feet of the Atlantis II Deep is about 50,000,000 tons. These deposits appear to extend to a depth of 60 feet below the sediment surface, but the quality of the deposits below 30 feet is unknown.

Mineral deposits

The recovery of sediment located beneath 5,700 to 6,400 feet of water poses problems. But since most of these metalliferous deposits are fluid oozes, it is anticipated that it may be possible to pump them to the surface much the same way as oil. There are also numerous proposals for drying and beneficiating (treating for smelting) these deposits after recovery. It would indeed seem that exploitation is now feasible, provided international agreements can resolve legal difficulties.

Navigation. Navigation in the Red Sea is difficult. The unindented shorelines of the northern half provide few natural harbours; in the southern half the growth of coral reefs has restricted the navigable channel and blocked some harbour facilities. At Bab el-Mandeb the channel is kept open by blasting and dredging. Atmospheric distortion (heat shimmer), sandstorms, and highly irregular water currents add to the navigational hazards.

Study and exploration. The Red Sea is one of the first large bodies of water mentioned in recorded history. It was important in early Egyptian maritime commerce (2000 BC) and was used as a water route to India by about 1000 BC. It is believed that it was reasonably well charted by 1500 BC, because at that time Queen Hatshepsut of Egypt sailed its length. Later the Phoenicians explored its shores during their circumnavigatory exploration of Africa in about 600 BC. Shallow canals were dug between the Nile and the Red Sea before the time of Christ but were later abandoned. A deep canal between the Mediterranean and the Red Sea was first suggested about AD 800 by the caliph Hārūn ar-Rashīd. It was not until 1869 that Ferdinand Marie de Lesseps completed the Suez Canal connecting the Red and Mediterranean seas. (B.C.S./W.B.F.R.)

Early use as a route to India

TIGRIS AND EUPHRATES RIVERS

The two greatest rivers of southwestern Asia have widely separated sources in the mountains of eastern Turkey and travel in a southeasterly direction through north Syria and Iraq to the head of the Persian Gulf. The total length of the Euphrates (Akkadian Purattu; Turkish Firat; Arabic Furāt; Biblical Perath) is approximately 1,700 miles (2,700 kilometres). The total length of the Tigris (Akkadian Idiklat; Biblical Hiddekel; Turkish Dicle; Arabic Dijla) is about 1,180 miles (1,890 kilometres). Both rivers traverse two-thirds of their courses before reaching the fringes of the Mesopotamian Plain—the silt-filled depression that must be regarded as their combined delta. To the north of this an upper and a middle course is usually distinguished for each river; each upper course is restricted to the valleys and gorges of east Anatolia, at altitudes diminishing from those of their sources at 6,000–10,000 feet (1,830–3,050 metres) above sea level; each middle course proceeding more tranquilly through the uplands of north Syria and Iraq at elevations varying from 1,200 feet at the foot of the so-called Kurdish Escarpment to 170 feet where the alluvium begins at the head of the delta.

The Tigris and Euphrates derive the bulk of their water from the winter rains and snow of the Güneydoğu Toroslar and Zagros mountains. Both, in their middle courses, are joined by substantial left-bank tributaries. Tributary streams of the Tigris, the Great and Little Zab, which are fed in the spring with snow-melt waters from the high ranges of Iraqi and Iranian Kurdistan, are both more powerful and more unpredictable than the Balīkh and the (western) al-Khābūr, which flow into the Euphrates from springs beneath the Escarpment. The Tigris, therefore, is a more copious and swifter stream than the Euphrates, its character being reflected in the Arabic name Dijla, meaning arrow. Geologically, the most striking feature of both streams is the heavy content of silt which they carry in flood-time. At this season their mean discharge in Iraq is reckoned at about 175,000 cubic feet (4,960 cubic metres) per second, and their water carries as much as 3,000,000 tons of eroded material from the highlands in a single day. The fact that little of this sediment reaches the sea explains the vast tract of alluvial soil of which the Mesopotamian Plain is composed.

Physical features. After reaching the Syrian frontier 250 miles apart, the rivers in their middle courses gradually converge until they are separated only by a triangle of barren limestone desert, known as al-Jazīrah (the Island). On either side of this they have cut deep and permanent beds in the Tertiary rock, so that their courses have remained virtually unchanged since prehistoric times and are punctuated at frequent intervals by the ruins of ancient cities. The Tigris here traverses the homeland of the Assyrians, whose three great capitals—Nineveh, Calah (modern Nimrūd), and Ashur—still overlook the river. By contrast, when the alluvial plain is reached, the rivers are found to have so frequently bifurcated or altered their beds that their earlier courses, and the canals dependent upon them, can sometimes be traced only by the lines of *talls* ("mounds") representing ancient settlements. The *talls* increase in number as the territory of Babylonia is reached,

Tigris–Euphrates river system.

and the mounds marking the sites of the great Sumerian cities in the extreme south suggest a pattern of occupation almost unrelated to the present system of waterways.

Use of embankments

The two rivers, in their lower courses, tend to build up their beds to a level considerably above the plain across which they flow. They are, furthermore, confined between artificial bunds (embankments) to prevent the inundation of the adjacent agricultural lands in the highwater season. Consequently, when, in the past, these bunds have proved ineffectual or have been damaged intentionally, not only have vast areas of surrounding country been flooded but on occasion the rivers have changed their main courses. The high level of the rivers has nevertheless made possible the system of irrigation to which the country owes its prodigious agricultural potential.

The Euphrates has no major right-bank tributaries in its delta, but is joined on its left bank by the Gharrāf Channel, which branches off from the Tigris at al-Kūt. The Tigris has two major left-bank tributaries. It is joined near Baghdad first by al-'Uzaym and then by the even larger Diyālā. This produces a volume of water in excess of the river's capacity, which dissipates itself in extensive marshes on either side of its lower course. To the west the marshes merge into the waters of a great shallow lake, the Hawr al-Ḥammār. The Euphrates flows through this lake before uniting with the Tigris in a single channel, the Shaṭṭ al-'Arab, which reaches the sea over 100 miles further south.

The Euphrates

Physiography. The headwaters of the Euphrates comprise two confluent streams known as the Murat and Kara Su. From their sources near Erzurum, Turkey, they traverse the high lava plain to the northwest of Lake Van and meet eventually at Keban, near Elâzığ, where the rockfill Keban Dam, completed in 1974, spans a deep gorge. The stream, after breaking through the range of Güneydoğu Toroslar, drops down through the foothills of the ancient

district of Commagene. Approaching to within 100 miles of the Mediterranean, it flows by Birecik in Turkey, where it is spanned by road, and Jarābulus on the Syrian side of the frontier, where it is crossed by the Iraqi Railway. From here its southeasterly course across Syria is through barren and thinly populated country, where its cultivable valley is no more than a few miles wide. After the western al-Khābūr joins the Euphrates, it flows through a broader agricultural province until it reaches Abū Kamāl, near the Iraqi frontier, after which the valley narrows again to a strip of alluvium between limestone escarpments. At intervals in the reach between the frontier and Hīt, small towns such as ʿĀnah and Rāwah occupy islands in mid-stream, subsisting upon a riparian system of cultivation that uses ingenious waterwheels for irrigation. Below Hīt, the river cliffs recede and from this point onward irrigation begins on a large scale. Just south of the river below ar-Ramādī lies Lake Ḥabbānīyah, a large depression, enclosed on three sides by low hills, that has been converted into a controlled escape, to minimize the danger of floods and to be used for storage during the low-water period. Between ar-Ramādī and al-Hindīyah, over a distance of about 140 miles, the mouths of all the main controlled irrigation canals as well as most of the pumping installations are to be found. At al-Hindīyah itself, until early in the present century, the river split into two branches, al-Ḥillah and al-Hindīyah, each of which, over the centuries, had alternately assumed importance. In 1908 a barrage was built to canalize and control al-Ḥilah branch, thus making al-Hindīyah the main river channel. Below al-Kifl the river enters an unstable area where effective control is difficult. It bifurcates twice before reappearing above as-Samāwah as a single stream, which still maintains its elevation above the surrounding marshes. Below an-Nāṣrīyah, numerous channels dispose of the waters, which make their way into the Hawr al-Ḥammār, from which, in turn, the outflow finds its way to the Shaṭṭ al-ʿArab.

The Tigris

The Tigris rises in a small mountain lake, Hazar Gölü, southeast of Elâziğ, Turkey. Minor tributaries, flowing into it, drain a wide area extending eastward into Hakkâri *il* (province). Passing beneath the great basalt walls of Diyarbakir, and by a troglodyte settlement at Hasankeyf, it reaches the Syrian frontier at its junction with the eastern al-Khābūr near Cizre (Jazīrat ibn ʿUmar), entering Iraq a few miles beyond at Faysh Khābūr. After passing Mosul, with the ruins of Nineveh on its left bank, the Tigris is joined by its two main tributaries, the Great and the Little Zab, which drain the mountain area of Iraqi Kurdistan. During the season of melting snows, the Zabs, laden with silt, double the volume of the main stream. The river next passes through the al-Fatḥah gorge, where rapids impede navigation, after which a further reach of 60 miles brings the Tigris to the head of the alluvial plain near Sāmarrāʾ. Here there is a barrage and a regulated escape that diverts surplus water into the Tharthār Depression to the west. From Baghdad onward, after receiving the waters of al-ʿUẓaym and the Diyālā, the Tigris is increasingly confined between artificial embankments, from which the southward overspill keeps the great Hawr Dalmaj (Dalmaj Marsh) almost perennially supplied with water. At al-Kūt, 200 miles downstream from Baghdad, there is a barrage and the channel of the Shaṭṭ al-Gharrāf (Shaṭṭ al-Hai) branches off southward. The Shaṭṭ al-Gharrāf is a very ancient bed of the river, irrigating a wide interfluvial area before merging its surplus water with that of the Euphrates where it enters the Hawr al-Ḥammār. The main stream, meanwhile, both above and below al-ʿAmārah, again splits into a number of channels that disperse their water over rice-growing areas and extensive marshes on either bank. At al-Qurnah it joins an old bed of the Euphrates, now fed by the outflow of these same marshlands to become Shaṭṭ al-ʿArab. At Qarmat ʿAlī, a little above Basra, the main stream is joined by more waters from the Euphrates that have filtered through the Hawr al-Ḥammār.

From Baghdad to Basra

Shaṭṭ al-ʿArab

On either side of the Shaṭṭ al-ʿArab, both above and below Basra, close settlement is limited to a belt of cultivation between one and three miles wide. These lands are irrigated by a network of creeks that fill with water from al-ʿArab when the sluices are opened at high tide and are famous for their groves of date palms, from which the fruit is exported all over the world. At Khorramshahr, al-ʿArab is joined by the Kārūn River, which flows southward from Khuzistan in Iran. From here to the sea an uneasy political agreement gives equal rights of navigation to Iraq and Iran. Basra on the right bank is, of course, the principal seaport of Iraq, while Khorramshahr, on the left bank, is Iranian. Two further ports on the lower course of al-ʿArab, Ābādān (Iran) and al-Fāw (Iraq) are no more than specialized discharge points for exporting oil.

Climate. The countries traversed by the Tigris and Euphrates have a continental sub-tropical climate, with extremes of heat and cold in summer and winter, respectively, and a scanty rainfall. In the alpine regions, where the rivers have their upper courses, winter winds are weak and variable, and much of the light precipitation falls in the form of snow, which lies for four or more months in the higher valleys. At this season the mean temperature is much below freezing, so that agriculture is at a standstill and communications are restricted. After the snowfields melt in the spring the rivers are in spate. The mounting volume of their waters is augmented in their middle courses by seasonal rainfall, which reaches its maximum between March and May. Although in the rivers' lower courses intermittent rains continue during the winter months, the total rainfall rarely exceeds eight inches (200 millimetres) annually, and from May onward rain ceases altogether. In the alluvial plain, the most conspicuous climatic feature is the extreme heat of the summer months, with day temperatures rising as high as 120° F (49° C), and with the relative humidity as low as 15 percent. The fertility of the delta is therefore entirely dependent upon the seasonal flooding of the Tigris and Euphrates.

Plant and animal life. Dense communities of common reed and the narrowleaf cattail are found in the wide areas of marshland through which the rivers flow in south Iraq. The young shoots of both are used by Marsh Arabs for fodder. Along the rivers themselves in their lower courses the Euphrates poplar and a species of willow grow in small belts. The poplar is used for practical purposes such as boat building. Undergrowth in these riverain thickets is composed chiefly of five-stamen tamarisk and mesquite, which also extends northward to the middle course of the Tigris and to its tributaries, up to an altitude of 3,000 feet. It is often accompanied by licorice, from whose roots an exportable product is obtained.

Wild pigs are widespread in the marshes adjoining the rivers throughout their courses. In southern Iraq, jackals and occasional hyenas are to be seen among the riverside gardens, and a very large form of Indian jungle cat inhabits the more remote tamarisk thickets. The smaller Eastern wild cat is less common. Among smaller animals are several species of gerbils and the agile jerboa, which may be seen as far north as central Anatolia. Buxton's mole-rat, which covers the entrance to its burrow with a mound of earth, is also found in the riverbanks. Many birds migrating between Europe and Asia fly along the rivers' course.

Locally resident birds include babblers, bulbuls, scrub warblers, and sandgrouse. Among the larger species that frequent the marshes and occasionally breed there are pelicans, flamingoes, herons (including the huge giant heron), storks, spoonbills, and two kinds of bustard. Winter visitors are geese, white-fronted or graylag, and innumerable ducks, including all the common European kinds as well as the red-crested pochard and marbled duck. Gray cranes breed on the upper reaches of the Euphrates. Brilliant European bee-eaters visit the riverbanks on passage, and Persian bee-eaters remain to breed. Both the European and the Indian rollers nest in the vicinity of the rivers, particularly in the vertical banks of water channels.

Among the freshwater fish characteristic of the Tigris-Euphrates system, the carp family is the most conspicuous. The rivers harbour a variety of genera in this family, including small-scale forms recalling the Indian mahseer. A barbel genus is recorded as attaining a weight of 300 pounds (136 kilograms). Besides carp, there are a few varieties of catfish as well as the spiny eel with its curious tubular nostrils.

The people. With the exception of city dwellers, the Arab population on the rivers' banks live either by stock breeding or by agriculture. Their way of life varies from the nomadism of the desert bedouin to the settled condition of the peasantry (fellahin) in the agricultural districts. Both bedouin and fellahin, together with semi-settled Arabs, may be included within the organization of a single tribe. Over the past half century, however, tribal associations have begun to disintegrate. Among the fellahin an improved system of land settlement and legal reforms are breaking the authority of feudal landlords and encouraging the syndication of small holdings.

The Marsh Arabs

The uniform simplicity of village life among the fellahin extends to the limits of the delta. Almost the only variation is to be found among the Maʿdan, or Marsh Arabs, who occupy the vast triangle of swampland between an-Nāṣirīyah, al-ʿAmārah, and Basra. A distinctive culture emerges from their seminomadic life. They raise water buffalo and hunt wildfowl or pigs from their mashuf canoes. The giant mardi reeds found in these swamps reach a height of up to 25 feet, providing them with building material for their characteristic architecture.

North of al-Fatḥah gorge, the Tigris and its tributaries pass through country in which pure Arabs are increasingly in a minority. In the 17th century this region provided winter pasture for the Kurdish tribes but was then settled with Turkmen by Sultan Murad IV, in order to secure his lines of communication with Baghdad. To the north the Turkmen merge with the Iraqi Kurds, whose orchards and vegetable gardens are found along the Tigris tributaries and their smaller affluents. In their upper courses, both the Tigris and the Euphrates pass through mountain country whose inhabitants were once predominantly Kurdish. Today, however, the Kurds have become assimilated into the Anatolian population. Until recently the highland provinces, where both rivers have their sources, were underpopulated as a result of the departure of the Armenians early in the 20th century.

The economy. The rivers have two well-marked flood periods; the first, an irregular rise mainly due to the rain, that lasts from November to the end of March; the second, that of the main flood in April and May. During the second period, the rivers' combined discharge may reach 175,000 cubic feet (4,960 cubic metres) per second, but the volume drops rapidly in June and rarely exceeds one-tenth of that amount in later summer and autumn. From the point of view of agriculture, the rivers are high at the wrong time of year for most crops (except rice), so that cultivation by direct inundation cannot be generally practiced.

Furthermore, the sheer volume of flood water endangers the bunds within which the rivers are confined in their lower courses. The primary requirement of river control is therefore to maintain an effective system of diversion and storage, both as a precaution against the kind of inundation that threatened the existence of Baghdad as recently as 1954 and as a means of retaining the flood waters for distribution in the hot season. To this end, in the early decades of the present century, barrages were built at al-Hindīyah and al-Kūt, and an important weir constructed on the Diyālā east of Baghdad. The diversionary escapes at ar-Ramādī on the Euphrates and at Sāmarrāʾ on the Tigris were both completed in 1956. An even larger storage basin has been created by the construction of a dam at Dukān on the Little Zab in 1958, and yet another is being built at Bakhma on the Great Zab.

Irrigation. The districts of northern Iraq and Syria, together with the piedmont area of southeastern Anatolia, through which the two rivers flow in their middle courses, have a milder climate than the plains in the south and sufficient rainfall to raise a crop of winter grain without irrigation. North of ʿĀnah on the Euphrates and of Tikrīt on the Tigris date palms disappear and are replaced by vines, olives, tobacco, and temperate fruits. There are no major canal systems in this area, and such irrigation as is required is supplied by mechanical lifts. By contrast, in the south the principal crops are wheat, barley, millet, and rice, as well as dates, all of which are dependent upon irrigation. Irrigation takes three forms. Water is distributed from rivers and canals by direct flow onto the land through small channels; by lifting water mechanically into such channels; and by direct inundation. The second of these methods is limited to land relatively close to rivers and canals; the third is useful only for rice cultivation. The relative importance of each method may be judged by the figures for 1942, an average and comparatively recent year; out of a total area cropped by irrigation, 1,700,000 acres (688,000 hectares) were dependent on lifts, 1,500,000 on canals, and 250,000 on inundation. Canals are of two kinds; controlled canals, which receive water from regulators on the main river at all seasons, and uncontrolled canals, which are fed only when the river level reaches the canal head—an eventuality that unfortunately occurs mainly at the season when water is least in demand.

The three systems of irrigation

While the Tigris, at the northern extremity of the delta, flows at a slightly higher level than the Euphrates, before reaching Baghdad the relationship is reversed, and the Euphrates becomes the higher of the two by about 30 feet. There is a second reversal before the Tigris reaches al-Kūt. In antiquity, advantage was taken of these characteristics, which permitted irrigation from one river and drainage into the other, and the situation is still partially reflected in the operation of the controlled canal zones of the present irrigation system.

Areas dependent upon irrigation

Five areas are dependent upon irrigation from large canals, controlled by barrages and sluice gates. The canal systems are the following: (1) the five left-bank Euphrates canals, running generally eastward between ar-Ramādī and al-Musayyib. Irrigation is perennial and is by free flow. (2) The Euphrates canals, depending directly on the Hindīyah Barrage. (3) The Diyālā canals, whose flow is maintained in summer by the Diyālā Weir. (4) The canal system on the Kūt Barrage, including the Gharrāf Canal, and the Dujaylah (an old bed of the Tigris). (5) Canals and spillways from al-ʿAmārah to Qalʿat Ṣāliḥ, on the left bank of the Tigris. Uncontrolled canal irrigation and inundation are predominantly practiced in rice-growing areas, such as in ash-Shāmīyah district (through which run the two branches of al-Hindīyah channel of the Euphrates) along the lower Euphrates around as-Samāwah, from an-Nāṣirīyah to the Hawr al-Ḥammār, and along the Lower Tigris below al-Kūt. Water-lifting machinery, consisting either of oil pumps or more primitive types of apparatus, is in use throughout the middle and lower courses of both rivers. Among the traditional devices in use, particularly on the middle Euphrates, are tall waterwheels, driven by the force of the current, which raise the water in earthenware jars attached to their rims. On the Tigris the height of lifts vary from six feet in the lowest reaches to 40 feet at Baghdad.

In the great irrigated areas of lower Mesopotamia, where the system of extensive cultivation has traditionally been practiced, a primary threat is the gradual salinization of the soil. The irrigation water from the rivers is slightly saline and the heavy residue left by evaporation in hot weather is augmented by salt forced to the surface by groundwater. The resultant deterioration of the soil has been promoted by the pattern of irrigation canals, whose dikes, created by dredging, have divided the low-lying ground into small units, making it difficult to drain off water before evaporation occurs. Philologists have found evidence in cuneiform texts to suggest that the deterioration of land through salinization was already well understood as early as the third millennium BC, when the centres of agricultural prosperity began to shift northward from the original homeland of the Sumerians.

Navigation. In 1836 a British expedition, led by Col. Francis Chesney, navigated the Euphrates by steamship from an-Nāṣirīyah to al-Qurnah, north of the Hawr al-Ḥammār. Today the Euphrates is not navigable except by local craft. From Basra to Baghdad, however, the Tigris is navigable by steamers with draughts of 4 feet 6 inches in the high water season and 4 feet at low water. Above Baghdad small steamers of three foot draught can, with some difficulty, reach Mosul. Much of the downstream traffic on both rivers consists of *kalaks*—rafts of timber and brushwood supported on inflated skins. *Kalaks* carry loads of up to 35 tons and take three or four days to cover

Types of river craft

the 275 miles from Mosul to Baghdad. Upon arrival, the rafts are dismantled, the timber is sold, and the skins are returned by road to their starting-point. Sailing craft include *muhaylahs* and *safīnahs* 30–80 feet long, with a capacity of up to 50 tons. Carrying smaller loads are *balams*—long, narrow double-ended, flat-bottomed craft with a shallow draft. Until recently *guffahs* —huge circular coracles of basketwork, coated with bitumen and capable of carrying up to 20 men—were in regular use in the vicinity of Baghdad.

The ancient trade route from the Persian Gulf to the Mediterranean followed the right bank of the Euphrates almost to Aleppo. The modern road and railway, after crossing the river at as-Samāwah, are diverted at al-Ḥillah toward Baghdad, but the road returns from Baghdad to the important Euphrates crossing at al-Fallūjah, after which it runs across the desert to Damascus. From Baghdad, with its four modern bridges over the mainstream, the railway follows the right bank of the Tigris to Mosul, then turns northwestward across the corner of Syria known as the "Bec-de-Canard" to follow the Turco-Syrian frontier to the Euphrates crossing at Jarābulus. On its way, it passes near the ruins of Haran (modern Harran), the "crossroads city" of the ancient world, and intersects a road that runs due north from ar-Raqqah to cross the headwaters of Euphrates once more at Samsat.

Among the mountains of east Anatolia, railways follow ancient paths, in part, along the upper course of the Tigris and both confluents of the Euphrates, bridging the Euphrates near Malatya.

Study and exploration. History has little to record regarding physiological changes in the upper and middle courses of the two rivers. In the delta, by contrast, the pattern created by complex hydrological developments over a period of several millennia is available for study, even if it is as yet only partially understood. Different explanations, for example, have been given for the way in which the plains were formed and the present coastline created. Until recently it was thought that the head of the Persian Gulf once extended as far as the apex of the delta but must have receded southward until, in early historic times, it corresponded roughly to a line drawn through al-ʿAmārah and an-Nāṣirīyah. Nearer to the present coastline, meanwhile, the westward outflow from the Kārūn River had built up a barrier of silt, behind which a great lake formed. The lake had then gradually filled with alluvium, until only the Hawr al-Ḥammār and the marshes around it remained. In 1952, this long-accepted interpretation was refuted by geologists who detected a gradual subsidence in the basal rocks beneath the Euphrates estuary at a rate that was enough to maintain the alluvial deposit at a constant level. They accordingly concluded that the coastline had remained almost unchanged since historic times—a contention that still requires to be reconciled with conflicting archaeological evidence.

The tangle of waterways and ancient irrigation channels with which the Mesopotamian Plain appears to be covered, especially when seen from the air, gives an exaggerated impression of ancient prosperity, modified only if one remembers that at no time in the past were all these channels in simultaneous use. Of the major canal systems on which the country depended in early medieval times, for instance, some are today re-used and extended, while others are obsolete and forgotten. The five great Euphrates canals (Isa, Sarsar, Malik, Kutha, and Nil), on which the fertility of central Iraq depended in ʿAbbāsid times, have modern counterparts, but their overflow can no longer drain into the Tigris because of a change in level. East of the Tigris, now long neglected and dry, is the great Nahrawān Canal, which irrigated an area extending almost to the Persian frontier. Leaving the Tigris just north of Sāmarrāʾ and utilizing an ancient bed of the main river, this canal collected the waters of al-ʿUẓaym and the Diyālā, carrying them southeastward almost to al-Kūt. It was abandoned after the Mongol invasions, and only a small part of its former dependent territory is now supplied with water by the Diyālā Weir. The old Ishaqi-Dujaylah Canal that irrigated the farmlands west of the Tigris in the same era also now lies derelict. (S.H.F.L.)

The ancient canals

ARABIAN SEA

The Arabian Sea lies in the northwestern section of the Indian Ocean. Situated within the trade wind latitudes, it forms part of the principal sea route between Europe and India. It is bounded to the east by India, to the north by Pakistan and Iran, to the west by the Arabian Peninsula and the Horn of Africa, and to the south by the remainder of the Indian Ocean. To the north the Gulf of Oman connects the sea with the Persian Gulf via the Strait of Hormuz. To the west the Gulf of Aden connects it with the Red Sea via the Strait of Bab el-Mandeb. Its total area is about 1,490,000 square miles (3,859,100 square kilometres); it has a mean depth of 8,750 feet (2,734 metres). In Roman times its name was the Mare Erythraeum (the Erythraean Sea).

Political units bordering the sea—apart from India, Iran, and Pakistan—are the Sultanate of Oman, Yemen (Aden), and Somalia. Islands in the sea include Socotra (a part of the Yemen [Aden]) off the Horn of Africa, the Kuria Muria Islands off the coast of Oman, and the Laccadive Islands (a part of India), the latter of which are a group of coral atolls lying between 100 and 250 miles (160–400 kilometres) off the southwest (Malabar) coast of India. The Indus River is the principal river draining into the sea.

Physical features. Most of the Arabian Sea has depths that exceed 9,600 feet, and there are no islands in the middle. Deep water reaches close to the bordering lands except in the northeast, off Pakistan and India. To the southeast the Laccadive Islands form part of the submarine Maldive Ridge, which extends farther south into the Indian Ocean where it rises above the surface to form the Maldive Islands. On the western side of the sea, the plateau island of Socotra, about 70 miles long and with an area of about 1,200 square miles, is an insular extension of the Horn of Africa, lying 160 miles east of Cape Guardafui.

Submarine morphology and geology. Stretching southeastward from Socotra is the submarine Carlsberg Ridge, which coincides with the belt of seismic activity in the Indian Ocean that divides the Arabian Sea into two major basins—the Arabian Basin to the east and the Somali Basin to the west. The deepest soundings, 15,900 feet, have been recorded in the Somali Basin. The Carlsberg Ridge is longitudinally split by a central valley that reaches depths of 10,800 feet below the sea's surface. The coastal escarpments of the Gulf of Aden are formed by rift faults that converge toward the southwest to continue into Africa as the boundary scarps of the Ethiopian Rift Valley, which forms part of the East African Rift Valley.

The Carlsberg Ridge

The Arabian Basin is separated from the Gulf of Oman Basin by the Murray Ridge, a narrow submarine ridge that extends northeast to southwest to meet the Carlsberg Ridge.

A deep submarine canyon has been discovered off the mouth of the Indus River, while an abyssal cone (a cone in the lowest depths of the ocean) and an associated abyssal plain occupy much of the northeast floor of the Arabian Sea. To the east of the Somali coast is another large abyssal plain.

The coastal shelf is narrow along the coast of the Arabian Peninsula and is even narrower along the Somali coast. No true coral reefs are found along the Arabian coast. Sediments near Raʾs al-Ḥadd (the easternmost headland of the Arabian Peninsula), where upwelling of deep water occurs in summer, consist of fine greenish mud with a high organic content containing hydrogen sulfide. The region, which contains many fish remains, is known as a fish cemetery. Deposits derived from the land cover the major part of the continental slope bordering the sea, down to a depth of about 8,250 feet. Below this, deposits consist of *Globigerina* (a genus of microscopic animals with calcareous shells belonging to a group known as Foraminifera), while basins lying below 12,000 feet are covered by red clay. Manganese nodules have been discovered in the midst of foraminiferal ooze; a single dredge obtained as much as 275 pounds (125 kilograms) of these nodules. Sediment thickness decreases from 8,200 feet in the north to about 1,600 feet in the south of the Arabian Basin.

The Arabian Sea is believed to have been formed during the Mesozoic and Cenozoic eras (*i.e.*, within the last

225,000,000 years). Geophysical studies conducted in the Gulf of Aden suggest that the continental blocks of Africa and Arabia separated, with new oceanic crust forming in between, in the pre-Miocene epoch (*i.e.*, more than 26,000,000 years ago).

Climate and hydrography. Minimum surface temperatures of about 75° to 77° F (24° to 25° C) occur in the central Arabian Sea in January and February, while temperatures higher than 82° F (28° C) occur both in June and in November. During the rainy season, which occurs when the southwest monsoon winds blow, from April to November, salinities of less than 35 parts per 1,000 have been recorded in the upper 150 feet of the sea, while during the dry season from November to March, when the northeast monsoon winds blow, salinities of more than 36 parts per 1,000 have been recorded at the surface over the entire Arabian Sea north of latitude 5° north, except off the Somali coast.

The Somali Current

The complex Somali Current, which attains speeds of about seven knots off the coast of Socotra, becomes part of a clockwise circulation system that continues to the northeast along the coast of Arabia and thence south along the coast of India to 10° north. At this point it merges with the southwestern drift current, flowing east between 5° and 10° north, that is associated with the southwest monsoon winds. Pronounced upwelling of deeper waters occurs along the Somali and Arabian coasts in summer. Of the five water masses that have been distinguished in the upper 3,000 feet of the north Indian Ocean, three have been identified as originating in the Red Sea, the Persian Gulf, and the Arabian Sea, respectively. The paths of flow of these water masses are to the south and east.

Marine life. High levels of inorganic nutrients, such as phosphate concentrate, which produce a rich fish life, have been observed in the western Arabian Sea and off the southeastern coast of the Arabian Peninsula. Occurring in the euphotic zone (zone of light, which is found in the upper 450 feet of the sea), this concentration of nutrients is undoubtedly due in part to coastal upwelling, which plays a major role in fertilizing the water.

Pelagic fish (*i.e.*, those living at or near the surface far from the land) in the Arabian Sea include tuna, sardine, billfish (a species having a long sharp bill or snout), wahoo (a large, swift game fish), sharks, lancet fish (a large species having dagger-like teeth), and moonfish (a species of silvery fish with very compressed bodies).

A periodic phenomenon in the Arabian Sea, however, is mass mortality of fish. In 1957, for example, in an area of approximately 77,000 square miles about 20,000,000 tons of fish were believed to have perished. Fish mortality is attributed to a subsurface layer of water of tropical origin that is poor in oxygen but rich in phosphate; under certain circumstances this layer is brought to the surface by strong upwelling, resulting in the death of fish from lack of oxygen.

Study and exploration. To medieval Arabs the Arabian Sea was known as the Sea of India or as part of the "Great Sea," from which smaller gulfs such as the Sea of Faris (Persian Gulf) or Sea of Kolzum (Red Sea) were distinguished. From about the 8th or 9th centuries onward.

The early Arabian and Persian navigators

Arab and Persian seafarers learned to use the surface currents propelled by the summer and winter monsoon winds. Detailed navigational instructions for sailing between southern Arabian, East African, and Red Sea ports, as well as ports in India, Malaysia, and China, were written down by pilots from Oman and the Ḥaḍramawt region of southern Arabia between the 9th and 15th centuries. Some of these works, entitled in Persian *rahmangs* (books of routes), contain useful information on navigating by the stars, on winds, currents, soundings, descriptions of coasts, approaches, and islands. Among flourishing medieval ports mentioned in these works are Diu and Surat in India, Hormuz in Persia, and Muscat and Aden on the Arabian Peninsula. Landmarks mentioned included Raʾs (cape) al-Ḥadd and Raʾs Madrak—capes on the southeast coast of the Arabian peninsula—and Cape Guardafui in Somalia, which is the cape of the Horn of Africa.

In recent times the Arabian Sea has been studied by several oceanographic expeditions, of which the most important has been the Mabahiss, or John Murray, Expedition of 1933 to 1934, which reported findings concerning hydrography, chemistry, currents, water masses, bottom topography, and sediments. Further information was obtained during the International Indian Ocean Programme (1960–65) in which British, U.S., Soviet, and German ships participated, studying currents, biological productivity, seismology, and geology. (An.A.A.)

BAY OF BENGAL

The Bay of Bengal, lying roughly between latitude 5°–22° N and longitude 80°–95° E, forms a relatively shallow embayment of the northeastern Indian Ocean. The bay occupies an area of 839,000 square miles (2,173,000 square kilometres) and is bordered by India and Sri Lanka (Ceylon) on the west, Bangladesh to the north, and Burma and the northern part of the Malay Peninsula to the east. According to the definition of the International Hydrographic Bureau, the southern boundary of the bay extends from Dondra Head at the southern end of Sri Lanka to the northern tip of Sumatra. The bay is about 1,000 miles (1,600 kilometres) wide, with an average depth of more than 2,600 feet (790 metres). The maximum depth is 14,764 feet (4,500 metres). A number of large rivers, namely, the Ganges and Brahmaputra on the north, the Irrawaddy on the east, and the Godāvari, Mahānadi, Krishna, and Cauvery on the west, flow into the Bay of Bengal. The Andaman and Nicobar groups are the only islands. Among the principal ports on the bay are the Indian ports of Calcutta, Cuddalore, Kākināda, Machilīpatnam, Madras, Paradip, and Vishākhapatnam.

A number of expeditions have traversed the area since the latter part of the 19th century, and, in connection with the International Indian Ocean Expedition, the "Vityaz" from the Soviet Union, and the "Pioneer" and "Anton Bruun" from the United States, have more recently carried out detailed investigations. Extensive work has been done on the western coast of the bay by research workers of Andhra University.

Physical features. *Physiography.* The Bay of Bengal is bordered to the north by a continental shelf 100 miles wide, but narrower to the south, and by slopes of varying gradient on the west, north, and northeast. Except for the submarine canyons around Sri Lanka, the trough-like valley off the Ganges Delta, and suspected north–south turbidity channels, the deep floor of the bay until recently was believed to be occupied by a vast plain sloping to the south and in places dissected by underwater valleys, perhaps caused by turbidity currents. As a result of the International Indian Ocean Expedition, however, the oceanography is now better understood, and many new physiographic features—mountain chains, submarine canyons, and deep channels—have been brought to light. One of the main submarine features is the north–south Indonesian Trench near the Nicobar-Sumatra mainland, which extends into the bay at a maximum depth of 14,800 feet. A large submarine canyon beginning at the head of the bay and called "Swatch of No Ground" extends across the continental shelf for 100 miles with a uniform breadth of eight miles. It appears to begin at about 60 to 180 feet and to cut into the shelf to depths of as much as 3,000 feet below the adjacent shelf level.

Bottom topography

In 1963 the Andhra, Mahadevan, and Krishna canyons were discovered off the Andhra coast. A number of other canyons have been identified off the Kākināda–Madras (Coromandel) coast and named after rivers in the vicinity of their locations, namely, Swarnamukhi Canyon, Pennar Canyon, Madras Canyon, Nagarjuna Canyon, Godāvari Canyon, Gautami Canyon, and another south of Puri. Some of them are believed to have been formed during the Pleistocene (10,000 to 2,500,000 years ago) and to have been maintained since then by progressive slumps, density flows, and slow creep.

Hydrography. A unique feature of the bay is the extreme variability of its physical properties. Temperature in the offshore areas, however, is very uniform at all seasons, with decreasing temperatures toward the north. Surface densities are considerably greater in spring than in fall. The sea presents alternately slick and ruffled surfaces over

shallow internal waves all along the east-coast shelf. Surface movements of the waters change direction with the season, the northeast monsoon giving them a clockwise circulation, the southeast monsoon a counterclockwise circulation. Severe storms occur at the change of monsoon, particularly to the south in October.

In addition to water-level changes resulting from waves and tide, the average sea level varies throughout the year.

Marine life and bottom deposits. Near its shores the Bay of Bengal is rich in phytoplankton and the related zooplankton. The coastal strips near the outlets of the Vamsadhāra, Nāgāvali, Vasishta Godāvari, and Vainateyam Godāvari rivers have deposits of heavy mineral sands, especially rich in manganese. The source of the manganese has been traced to the rocks in the drainage basins of the rivers emptying their waters and sediment load into the bay. The amount of organic matter present in the continental-shelf sediment of the northern part of the east coast is poor compared with the world's average for nearshore sediments.

Mineral assemblages

On the basis of the available data, the Bay of Bengal can be divided into two areas, one characterized by clay minerals such as illite and kaolinite (Andaman and Nicobar Islands) and the other containing montmorillonite—another clay mineral characterized by swelling in water (eastern and southern parts of the bay). These minerals appear to be derived mainly from the Indian peninsula and from the Himalayas, where they occur abundantly in sedimentary rocks and soils of the river basins. Modern geophysical methods of exploration employed in offshore prospecting are likely to open new vistas into the bay's structure and possible mineral resources under the ocean bed.

The economy. The bay faces on the fertile deltas of many large and navigable rivers: the Ganges and Brahmaputra on the north, the Irrawaddy on the east, and the Mahānadi, Godāvari, and Cauvery on the west. The Deccan coast provides no harbours, Madras having a mere open roadstead; but on the east the drowned coast has given many good ports, such as Sittwe, Moulmein, and Rangoon. (S.B.)

BRAHMAPUTRA RIVER

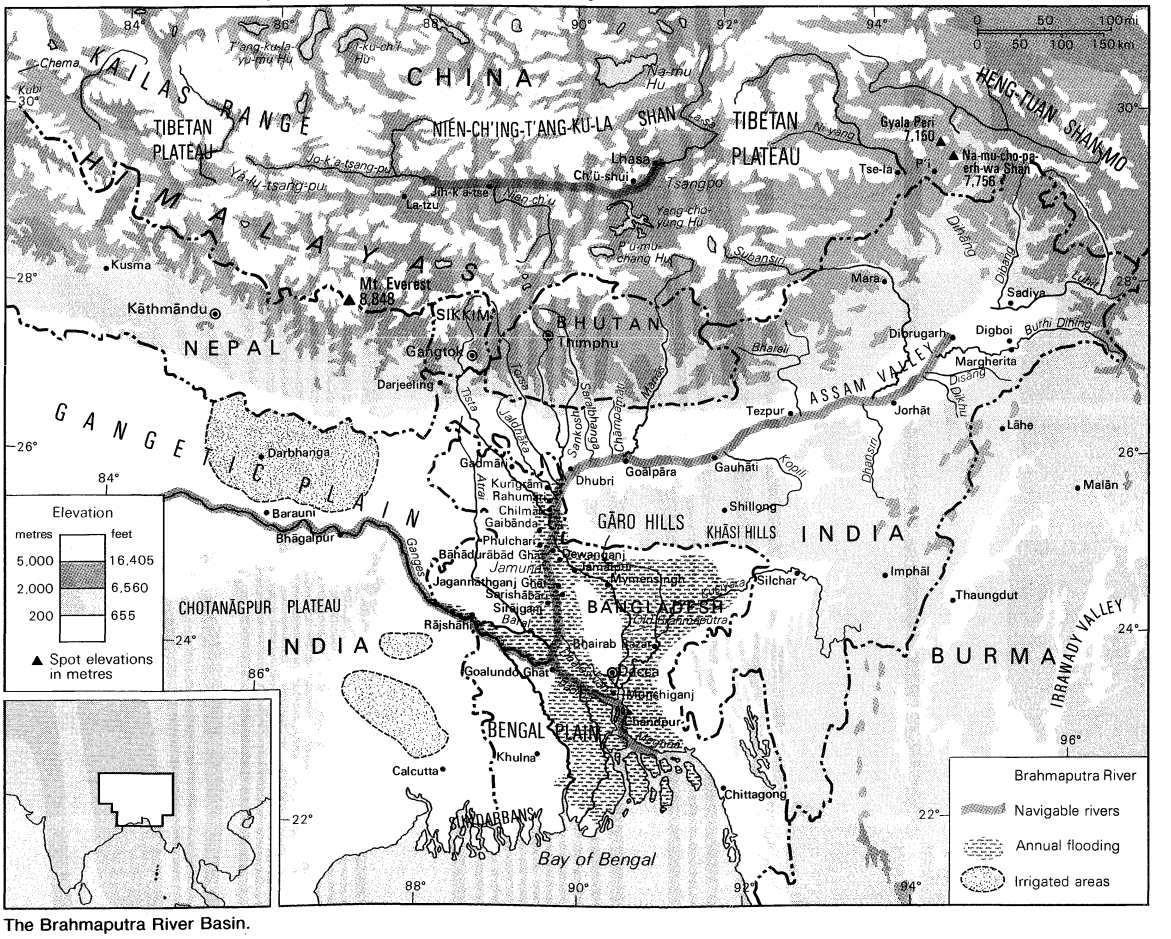
The mighty Brahmaputra River flows for 1,800 miles (2,900 kilometres) from its source in the Himalayas to its confluence with the Ganges, after which the mingled waters of the two rivers empty into the Bay of Bengal. Along its course it passes through the Tibetan Autonomous Region of China; the union territory of Arunachal Pradesh and the state of Assam, India; and Bangladesh. It is known to the Tibetans as the Tsangpo, to the Chinese as the Ya-lu-ts'ang-pu, to the Indians as the Brahmaputra, and to the people of Bangladesh as the Jamuna. For most of its length, the river serves as an important inland waterway; it is not, however, navigable between the mountains of Tibet and the plains of India. In its lower course, the river is both a creator and a destroyer—depositing large amounts of fertile alluvial soil but also causing disastrous and frequent floods.

Physical features. The river's source lies in the Chema-Yungdung Glacier, which covers the slopes of the Himalayas about 60 miles (100 kilometres) southeast of Lake Manasarowar in southwestern Tibet. The three headstreams that arise there are the Kubi, the Angsi, and the Chema-Yungdung. From its source the river runs for nearly 700 miles (1,120 kilometres) in a generally easterly direction between the main Himalayan range to the south and the Nien-ch'ing-t'ang-ku-la Shan (Nyenchen Tangla) to the north. Throughout its upper course the river is generally known as the Tsangpo (the Purifier); it is also known by the Chinese name Ya-lu-tsang-pu Chiang and by other local Tibetan names at various points along its course.

The river's upper course

In Tibet the Tsangpo receives a number of tributaries. The most important left-bank tributaries are the Jo-k'a tsang-pu (Raga Tsangpo), which joins the river west of Jih-k'a-tse (Zhikatze), and the La-sa Ho (Kyi Chu), which

Flooding data from R. Rawson, *The Monsoon Lands of Asia;* irrigation data from L. Cantor, *A World Geography of Irrigation* (1970), Oliver and Boyd, Edinburgh



The Brahmaputra River Basin.

flows past the Tibetan capital of Lhasa and joins the Tsangpo at Ch'ü-shui. The Ni-yang Ho (Gyamda Chu) joins the river from the north at Tse-la (Tsela Dzong). On the right bank the Nien-ch'u Ho (Nyang) meets the river at Jih-k'a-tse.

After passing P'i (Pe) in Tibet, the river turns suddenly to the northeast and cuts a course through a succession of great, narrow gorges between the mountainous complex of Gyala Peri (23,458 feet, or 7,150 metres, in height) and Na-mu-cho-pa-erh-wa Shan (Namcha Barwa) (25,446 feet, or 7,756 metres, in height) in a series of rapids and cascades. Thereafter, the river turns south and forces its way through the eastern extremity of the Himalayas to enter the Assam Valley of northeastern India as the Dihāng River.

At the town of Sadiya, India, the Dihāng turns to the southwest and is joined by the two mountain streams of Luhit and Dibang. After the confluence, about 900 miles from the Bay of Bengal, the river is known as the Brahmaputra (the Son of Brahmā [the Creator]). In Assam the river is mighty, even in the dry season, and during the rains its banks are more than five miles apart. As the river follows its braided, 450-mile course through the valley, it receives several rapidly rushing Himalayan streams, including the Subansirī, Bhareli, Dhansiri, Manās, Chāmpamāti, Saralbhānga, and Sankosh rivers. The main tributaries from the hills and from the plateau to the south are the Burhi Dihing, the Disāng, the Dikhu, and the Kāpili.

The complex lower course

The Brahmaputra enters the plains of Bangladesh after turning south around the Gāro Hills below Dhubri, India. After flowing past Chilmāri, Bangladesh, it is joined on its right bank by the Tīsta River, following a 150-mile course due south as the Jamuna River. (South of Gaibānda, the Old Brahmaputra leaves the left bank of the main stream and flows past Jamālpur and Mymensingh to join the Meghna River at Bhairab Bāzār.) Before its confluence with the Ganges River, the Jamuna receives the combined waters of the Baral, Atrai, and Hurāsāgar rivers on its right bank and becomes the point of departure of the large Dhaleswari River on its left bank. A distributary of the Dhaleswari, the Burhi Ganga, flows past Dacca and joins the Meghna River above Munshiganj.

The Jamuna joins with the Ganges north of Goalundo Ghāt, after which, as the Padma, their combined waters flow to the southeast for a distance of about 65 miles. The Padma reaches its confluence with the Meghna River near Chāndpur and then enters the Bay of Bengal through the Meghna Estuary and lesser channels.

Climate. The climate of the Brahmaputra Valley varies from the harsh, cold, and dry conditions found in Tibet to the generally hot and humid conditions prevailing in the Assam Valley and Bangladesh. Tibetan winters are severely cold, with minimum temperatures below 32° F (0° C), while summers are mild and sunny. The river valley lies in the rain shadow of the Himalayas; there is little rain in the summer, but some snow and rain falls during the winter. In the Indian and Bangladesh part of the valley, the monsoon climate is somewhat modified; the hot season is shorter than usual, and the average temperature is 82° F (28° C). Precipitation is relatively heavy, and humidity is high throughout the year. The annual rainfall of between 70 and 150 inches (1,780 and 3,810 millimetres) falls mostly between June and early October; light rains also fall from March to May.

Plant and animal life. Large areas in Assam are covered with sal (valuable timber trees that yield resin) forests, and tall reed jungle grows in the swamps and depressed, water-filled areas (*jhīls*) of the immense floodplains. Around the settlements in the Assam Valley, the many fruit trees yield plantains, papayas, mangos, and jackfruit. Bamboo thickets abound everywhere.

The most notable animal of the swamps in Assam is the one-horned rhinoceros, which has become extinct in other parts of the world. Tigers and elephants are also found. Numerous varieties of fish include the betki, pabda, ruhi, chital, and mrigal.

Hydrography. Constant changes of the river's course constitute a significant factor in the hydrology of the Brahmaputra; the most spectacular of these changes was the eastward diversion of the Tīsta River and the ensuing development of the new channel of the Jamuna, which occurred in 1787 with an exceptionally high flood in the Tīsta. The waters were suddenly diverted eastward into an old abandoned course, causing the river to join the Brahmaputra opposite Bāhādurābād Ghāt in Mymensingh District. Until the late 18th century the Brahmaputra flowed past the town of Mymensingh and joined the Meghna River near Bhairab Bāzār (the path of the present-day Old Brahmaputra Channel). At that time, the course of the Jamuna River (now the main Brahmaputra Channel) was a minor stream called the Konai-Jenai, which was probably a spill channel of the Old Brahmaputra. After being reinforced by the Tīsta flood of 1787, the Brahmaputra began to cut a new channel along the Konai-Jenai and gradually converted it after 1810 into the main stream, now known as the Jamuna.

Changes in the river's course

Along the lower courses of the Ganges and Brahmaputra and along the Meghna, the land is subjected to constant erosion and deposition of silt because of the shifts and changes in these active river courses. Vast areas are subject to large-scale inundation during the monsoon months from June to September. The oscillations of the Jamuna since 1787 have been considerable, and the river is never in exactly the same place for two successive years. Islands and sizable newly deposited lands (*chars*) in the river appear and disappear seasonally. The *chars* are valuable to the economy of Bangladesh as additional cultivable areas.

In Tibet the waters of the Brahmaputra are clear because little silt is carried downstream. As soon as the river enters the Assam Valley, however, the silt charge becomes heavy. Because of the speed and volume of water in the northern tributaries that flow down from the rain-soaked Himalayan slopes, their silt charge is much heavier than that carried by the tributaries that cross the hard rocks of the old plateau to the south. In Assam the deep channel of the Brahmaputra follows the southern bank closer than the northern. This tendency is reinforced by the silt-laden northern tributaries pushing the channel south.

Another important hydrographic feature of the river is its tendency to flood. The quantity of water carried by the Brahmaputra in India and Bangladesh is enormous. The Assam Valley is enclosed by hill ranges on the north, east, and south and receives over 100 inches of rainfall annually, while in the Bengal Plain heavy rainfall (70 to 100 inches) is reinforced by the huge discharge of the Tīsta, Torsa, and Jaldhāka rivers. This results in heavy annual floods and an estimated discharge during the rainy season of 500,000 cubic feet (14,200 cubic metres) per second. The Assam earthquake of 1950 led to numerous landslides on the Himalayan slopes of the Assam Valley; the resulting enlarged discharge, combined with enormous amounts of silt, caused extraordinary floods. Between 1950 and 1970 there were heavy flood conditions every year except 1951, 1952, and 1958.

The people. The people living in the different sections of the Brahmaputra Valley are of diverse origin and culture. North of the Himalayan rampart, the Tibetans practice Buddhism and speak the Tibetan language. They engage in animal husbandry and cultivate the valley with irrigation water taken from the river.

The Assamese are a mixture of Mongolian–Tibetan, Aryan, and Burmese ethnic origins. Their language is akin to Bengali, which is spoken in West Bengal, India, and Bangladesh. From the late 19th century a vast number of immigrants from the Bengal Plain of Bangladesh entered the valley; they settled there to cultivate the almost empty lands, particularly the low floodplains. In the Bengal Plain itself the river flows through an area that is densely populated by the Bengali people, who cultivate the fertile valley. The hilly margins of the plain are inhabited by the hill tribes of the Garos, Khāsis, and Hajangs.

The economy. *Irrigation and flood control.* Flood-control schemes and the building of embankments were initiated after 1954. In Bangladesh the Brahmaputra embankment running west of the Jamuna River from north to south helps to control floods. The Tīsta Barrage Project is both an irrigation and a flood-protection scheme.

The Tīsta Barrage Project

Little power has been harnessed along the Brahmaputra or the Assam Valley, although the estimated potential is great. The tributaries that drain the Shillong plateau have deep stretches and a series of rapids and falls that are suitable for hydroelectric development. Actual development is so far limited to the Barpani Dam and the Umiam Project near Gauhāti, which have a total capacity of almost 60,000 kilowatts and which serve areas of the Khāsi Hills in Meghalaya. There is little demand for electric power in the Brahmaputra Valley because manufacturing is limited and there is relatively little urbanization. Power needed on the tea estates in the Assam Valley is locally generated by thermal sources, which mainly use crude oil.

Navigation and transport. The river is navigable in Tibet for about 400 miles between La-tzu (Lhatse Dzong) and Lhasa. Coracles (boats made of hides and bamboo) and large ferries ply its waters at 12,000 feet above sea level. The Tsangpo is spanned in several places by suspension bridges.

Because it flows through a region of heavy rainfall in Assam and Bangladesh, the Brahmaputra is more important for inland navigation than for irrigation. The river has long formed a waterway between West Bengal and Assam. It is navigable throughout the Bengal Plain and Assam upstream to Dibrugarh, 800 miles from the sea. Besides all types of local craft, powered launches and steamers may easily ply up and down the river, carrying bulky raw materials, timber, and crude oil. After the conflict between India and Pakistan in 1965, however, heavy traffic was suspended.

The Brahmaputra remains unbridged throughout its course in the plains. Roads and railroads run along the river but never cross it; ferries carry traffic between its banks. Sadiya, Dibrugarh, Jorhāt, Tezpur, Gauhāti, Goālpāra, and Dhubri are important towns and ferry-crossing points in Assam, while Kurīgrām, Rahumāri, Chilmāri, Bāhādurābād Ghāt, Phulchari, Sarishābāri, Jagannāthganj Ghāt, Nagarbāri, Sirājganj, and Goalundo Ghāt are important crossing points in Bangladesh. The railheads are located at Bāhādurābād Ghāt, Phulchari, Jagannāthganj Ghāt, Sirājganj, and Goalundo Ghāt.

Study and exploration. The upper course of the Brahmaputra was unknown until the 19th century. The explorations of the Indian surveyor Kinthup in 1884 and of J.F. Needham in Assam in 1886 established the Tsangpo as the upper course of the Brahmaputra. Various British expeditions in the first quarter of the 20th century explored the Tsangpo upstream to Jih-k'a-tse, as well as the river's mountain gorges.

GANGES RIVER

The great river of the North Indian plains is officially as well as popularly called the Ganga, both in Hindi and in other Indian languages. Internationally, however, it is known by its anglicized name, the Ganges. From time immemorial it has been the holy river of the Hindus. For most of its course it is a wide and sluggish stream, flowing through one of the most fertile and densely populated tracts of territory in the world. Despite its importance, its length of 1,557 miles (2,506 kilometres) makes it only the 15th longest river in Asia, and the 39th longest river in the world.

Rising in the Himalayas and emptying into the Bay of Bengal, it drains a quarter of the territory of India, while its basin supports a concentration of about 300,000,000 people, a population larger than that of any state on earth with the exceptions of China and India. The Gangetic Plain, across which it flows, is the heartland of the region known as Hindustān and has been the cradle of successive civilizations from the Kingdom of Aśoka in the 3rd century BC, down to the Mughal Empire, founded in the 16th century.

For most of its course the Ganges flows through Indian territory, although its large delta in the Bengal area, lies mostly in Bangladesh. The general direction of the river's flow is from north-northwest to southeast. At its delta, the flow is generally southward.

Physical features. *Physiography.* The Ganges rises in the southern Himalayas on the Indian side of the Tibet border. Its five headstreams—the Bhāgirathi, the Alaknanda, the Mandakini, the Dhauli Ganga, and the Pindar—all rise in the Uttarakhand region (the northern mountainous districts), a division of the state of Uttar Pradesh. Of these, the two main headstreams are the Alaknanda (the longer of the two), which rises about 30 miles north of the Himalayan peak of Nanda Devi, and the Bhāgirathi, which originates about 10,000 feet (3,050 metres) above sea level in an ice cave at the foot of the Himalayan glacier known as Gangotri. Gangotri itself is a sacred place for Hindu pilgrimage. The true source of the Ganges, however, is considered to be at Gaumukh, about 13 miles south of Gangotri.

The Himalayan headstreams

After the Alaknanda and Bhāgirathi unite at Devaprayāg, they form a main stream known as the Ganga, which cuts through the outer (southern) Himalayas to emerge from the mountains at Rishikesh. It then flows onto the plain at Hardwār, a place that is held sacred by the Hindus.

Although there is a seasonal variation in the river's flow, its volume increases markedly as it receives more tributaries and enters a region of heavier rainfall. From April to June the melting Himalayan snows feed the river, while in the rainy season from July to September the rain-bearing monsoon winds cause floods. Within the state of Uttar Pradesh, the principal right bank tributaries are the Jumna (Yamuna) River that flows past Delhi, the capital of India, to join the Ganges near Allahābād and the Tons that descends from the Vindhya Range in the state of Madhya Pradesh and joins it soon after. The left bank tributaries in Uttar Pradesh are the Rāmganga, the Gomati, and the Ghāghara.

The Ganges next enters the state of Bihār, where its main tributaries from the Himalayan region to the north are the Gandak, the Burhi Gandak, the Ghugri, and the Kosi and its most important southern tributary is the Son. In West Bengal, the last Indian state that the Ganges enters, the Mahānanda joins it from the north. (Throughout West Bengal in India, as well as in Bangladesh, the Ganges is locally called the Padma). The river then skirts the Rājmahāl Hills to the south, and flows southeast to Farakka, at the apex of the delta. The westernmost distributary of the delta is the Hooghly, on the east bank of which stands the city of Calcutta. The Hooghly itself is joined by two tributaries flowing in from the west, the Dāmodar and the Rupnarayan. In Bangladesh, the Ganges is joined by the mighty Brahmaputra (which for about 150 miles before the junction is called Yamuna) near Goalundo Ghāt. The combined stream, now called the Padma, joins with the Meghna River above Chandpur. The waters then flow to the Bay of Bengal through innumerable channels, the largest of which is known as the Meghna Estuary.

Dacca, the principal city of Bangladesh, stands on the Burhi Ganga, a tributary of the Dhaleswari. Apart from the Hooghly and the Meghna, the other distributary streams which form the Ganges Delta are: in West Bengal, the Jalangi; and in Bangladesh, the Matabhanga, the Bhairab, the Kobadak, the Gorai (Madhumati), and the Ariāl Khān.

The lengths of the major tributaries of the Ganges, in miles, are as follows: the Jumna (860), the Rāmganga (428), the Ghāghara (600), the Son (487), the Gandak (263), the Burhi Gandak (378), the Kosi (450), the Mahānanda (180), the Dāmodar (368), and the Brahmaputra (1,800).

The delta. The Ganges, as well as its tributaries and distributaries, is constantly vulnerable to changes in its course in the delta region. Such changes have occurred in comparatively recent times, especially since 1750. In 1785, the Brahmaputra flowed past the city of Mymensingh; it now flows more than 40 miles west of it before joining the Ganges.

The delta, the seaward prolongation of silt deposits from the Ganges and Brahmaputra river valleys, covers an area of about 22,000 square miles (57,000 square kilometres) and is composed of repeated alternations of clays, sands, and marls, with recurring layers of peat, lignite, and beds of what were once forests. The new deposits of the delta, known in Hindi and Urdu as the *khādar,* naturally occur in the vicinity of the present channels.

The southern surface of the Ganges Delta has been

Silt deposits in the delta

formed by the rapid and comparatively recent deposition of enormous loads of silt. To the east, the seaward side of the delta is being changed at a rapid rate by the formation of new lands, known as *chārs*, and new islands. So much silt is deposited here that the 100 fathom (183 metre) line lies much farther out to sea than it does, for example, off the mouth of the Indus in the Arabian Sea. The western coastline of the delta has, however, remained practically unchanged since the 18th century.

The rivers in the West Bengal area, being sluggish, have been described as dead or dying; little water passes down them to the sea. In the Bangladesh delta region, the rivers are broad and active, carrying plentiful water; they are connected by innumerable creeks. During the rains, from June to October, the greater part of the region is flooded to a depth of several feet, leaving the villages and homesteads, which are built on artificially raised land, isolated above the flood waters. Communication between settlements during this season can be accomplished only by boat.

To the seaward side of the delta as a whole there is a vast stretch of tidal forests and swampland. The forests, which are called Sundarbans (Sanskrit meaning "beautiful forest"), are protected by India and Bangladesh. For conservation purposes, no permanent settlement is permitted in them.

In certain parts of the delta, there occur layers of peat, composed of forest vegetation and rice plants. In many natural depressions, known as *bil,* peat, still in the process of formation, is used as a fertilizer by local farmers. In recent years, it also has been dried and used as a domestic and industrial fuel.

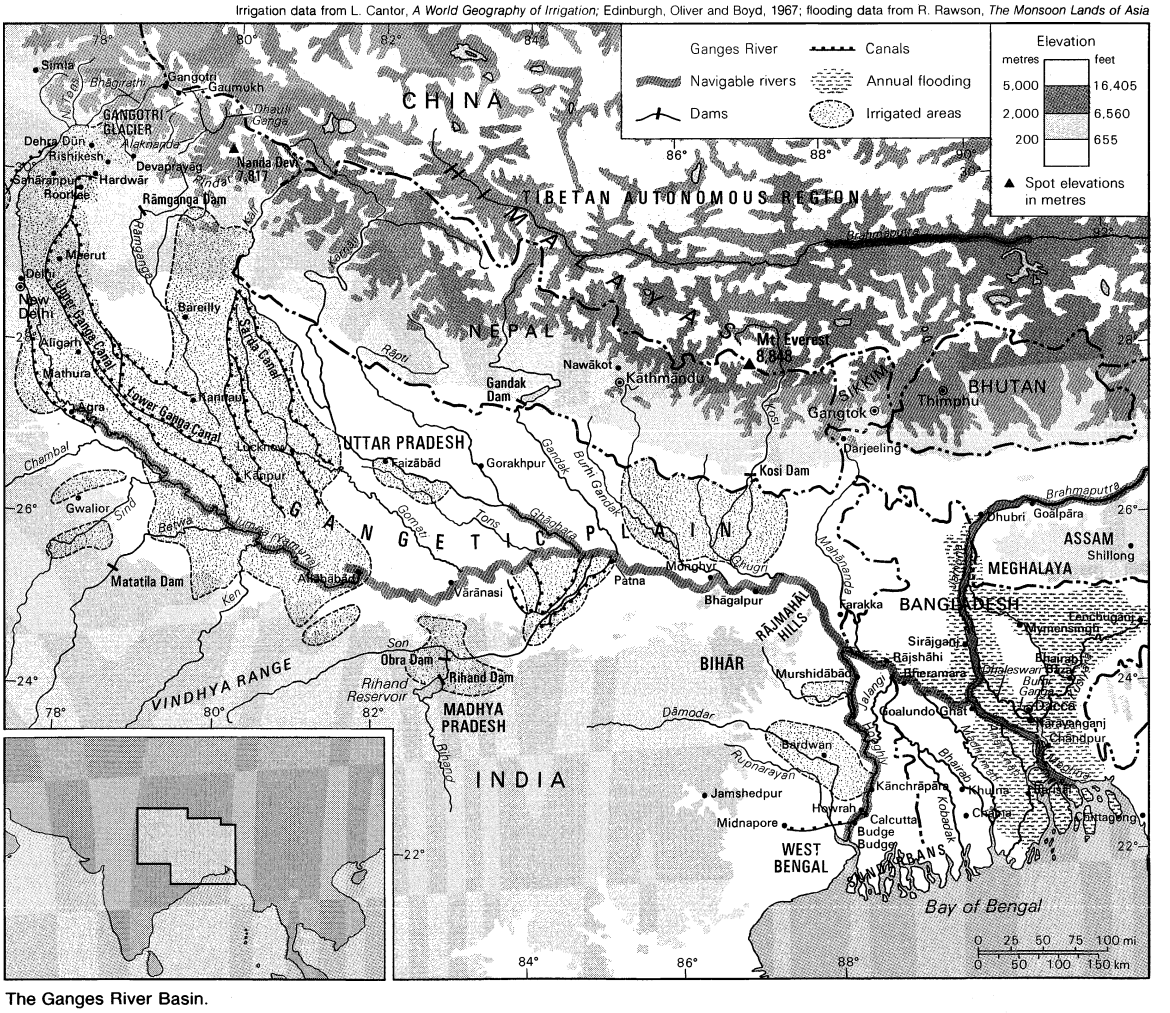
Climate and hydrography. The Ganges Basin contains the largest river system on the subcontinent. The water supply is dependent partly on the rains brought by the monsoon winds from July to October, as well as on the flow from melting Himalayan snows, in the hot season from April to June. Precipitation in the river basin accompanies the southwest monsoon winds, but is also related to cyclones that originate in the Bay of Bengal between June and October. Only a small amount of rainfall occurs in December and January. The average annual rainfall varies from 30 inches (760 millimetres) at the western end of the basin to over 90 inches at the eastern end. (In the Upper Gangetic Plain in Uttar Pradesh rainfall averages about 30 to 40 inches, in the Middle Plain of Bihār from 40 to 60 inches, and in the delta region between 60 and 100 inches.) The delta region experiences strong cyclonic storms both before the commencement of the monsoon season, from March to May, and at the end of it, from September to October. Some of these storms result in much loss of life and destruction of homes, crops, and livestock. One such storm, which occurred in November 1970, was of catastrophic proportions, resulting in deaths of at least 200,000 people.

Since there is little variation in relief over the entire surface of the Gangetic Plain, the river's rate of flow is slow. Between the Jumna River at Delhi and the Bay of Bengal, a distance of nearly 1,000 miles, the elevation drops nearly 700 feet. Altogether the Ganges-Brahmaputra plains extend over an area of 300,000 square miles. The alluvial mantle of the plain, which in some places is more than 6,000 feet thick, is possibly not more than 10,000 years old.

Plant and animal life. The Ganges-Jamuna area was once densely forested; historical writings indicate that in the 16th and 17th centuries wild elephants, buffalo, bison, rhinoceroses, lions, and tigers were hunted there. Most of the original natural vegetation has disappeared from the Gangetic Basin as a whole, and the land is now intensely cultivated to meet the needs of an ever-growing population. Wild animals are few, except for deer, boars, and

Disappearance of original vegetation

Irrigation data from L. Cantor, *A World Geography of Irrigation;* Edinburgh, Oliver and Boyd, 1967; flooding data from R. Rawson, *The Monsoon Lands of Asia*



The Ganges River Basin.

wildcats, and some wolves, jackals, and foxes. Only in the Sundarban area of the delta are some Bengal tigers, crocodiles, and marsh deer still found. Fish abound in all the rivers; especially in the delta area where they form an important item of diet. Many varieties of birds are found, such as mynah birds, parrots, crows, kites, partridges, and fowls. In winter, duck and snipe migrate south across the high Himalayas, settling in large numbers in water-covered areas. In the Bengal area common fish include featherbacks (Notopteridae), barbs (Cyprinidae), walking catfish, gouramis (Anabantidae), and milkfish (Chanidae).

The people. Ethnically, the people of the Ganges Basin are of mixed origin. In the west and centre of the basin they were originally descended from Aryan ancestors. Later, Turks, Mongols, Afghans, Persians, and Arabs came from the west and intermingled with them. To the east and south, largely in the Bengal area, an admixture of Tibetan, Burmese, and miscellaneous hill people has also occurred. The Europeans, arriving still later, did not settle or intermarry to any extent.

Historically the Gangetic Plain has constituted the heartland of Hindustān and has cradled its successive civilizations. The centre of the pre-Christian empire of Aśoka was Patna (Pāṭaliputra), standing on the banks of the Ganges in Bihār. The centres of the great Mughal Empire were at Delhi and Āgra, on the western peripheries of the Gangetic Basin. Kannauj on the Ganges, north of Kānpur, was the centre of the feudatory Empire of Harṣa, which covered most of North India in the middle of the 7th century. During the Muslim era, which began in the 12th century, Muslim rule extended not only over the plain, but over all of Bengal as well. Dacca and Murshidābād in the delta region were centres of Muslim power.

The British, having founded Calcutta on the banks of the Hooghly in the late 17th century, gradually advanced up the valley of the Ganges, reaching Delhi in the mid-19th century.

A great number of cities have been built on the Gangetic Plain. Among the most notable are Roorkee, Sahāranpur, Meerut, Āgra (the city of the famous Tāj Mahal mausoleum), Mathura (esteemed as the birthplace of the Hindu god Krishna), Alīgarh, Kānpur, Bareilly, Lucknow, Allahābād, Vārānasi (Benares; the holy city of the Hindus), Patna, Bhāgalpur, Rājshāhi, Murshidābād, Bardwan, Calcutta, Howrah, Dacca, Khulna, and Barisāl.

In the delta, the area down the Bhāgirathi-Hooghly Channel was urbanized even before the arrival of the British; subsequent growth of commerce and industry resulted in further urbanization. Calcutta and its satellite towns, known as the Calcutta Metropolitan District, stretch along both banks of the Hooghly for about 50 miles, extending from Kānchrāpāra-Kalyāni in the north to Budge Budge in the south.

The holy river

The religious importance of the Ganges may exceed that of any other river in the world. It has been revered from the earliest times, and today is regarded as the holiest of rivers by Hindus. While places of Hindu pilgrimage, called *tirath,* are located throughout the subcontinent, those that are situated on the Ganges have particular significance. Among these are the confluence of the Ganges and the Jumna at Allahābād, where a bathing festival, or *melā,* is held in January and February, during which about 400,000 persons immerse themselves in the river. Other holy places for immersion are at Vārānasi (Benares), or Kāśī, and at Hardwār.

The Hooghly River at Calcutta is also regarded as holy. The places of pilgrimage on the Ganges also include Gangotri and the junction of the Alaknanda and Bhāgirathi headstreams. The Hindus cast their dead upon the river, believing that they will thus go straight to heaven, and cremation ghats (temples at the summit of riverside steps) for burning the dead have been built in many places on the banks of the Ganges.

The economy. *Irrigation.* Use of the Ganges water for irrigation, either when the river is in flood, or by means of gravity canals, has been common since early times. Such irrigation is described in scriptures and mythological books written long before the Christian Era. Megasthenes, a Greek ambassador who was in India, recorded the use of irrigation in the 4th century BC. Irrigation was highly developed during the period of Muslim rule from the 12th century onward, and the Mughal kings later constructed several canals. The canal system was further extended by the British.

The cultivated area of the Ganges Valley in Uttar Pradesh and Bihār benefits from a system of irrigation canals which have increased the production of such cash crops as sugarcane, cotton, and oilseeds. The older canals are mainly in the Ganges-Jumna Doab, a word which means "land between two rivers." The Upper Ganga Canal (which with its distributaries is 5,950 miles long), which begins at Hardwār, was opened in 1856; the Lower Ganga Canal (5,120 miles with distributaries), opened in 1880, begins at Naraura. The Sarda Canal irrigates land in Oudh, in Uttar Pradesh. The land north of the Ganges, being higher, is difficult to irrigate by canal, and water in the subsoil must be pumped to the surface by electricity. The expansion of irrigation into this area has, therefore, depended on the availability of power in the northern part of Uttar Pradesh. Large areas in Uttar Pradesh and in Bihār are also irrigated by channels running from handdug wells.

The Ganges-Kobadak scheme

The Ganges-Kobadak scheme in Bangladesh, largely an irrigation plan, covers parts of the districts of Khulna, Jessore, and Kushtia, lying within the moribund part of the delta where silt and overgrowth choke the rivers.

Total annual rainfall in this region is generally below 60 inches, and winters are comparatively dry. The system of irrigation is based on both gravity canals and lifting devices. Power is provided by a thermal plant situated at Bheramara, on the Ganges in the Kushtia District.

Navigation. In ancient times the Ganges and some of its tributaries, especially in the east, were navigable. According to the Greek Megasthenes, navigation took place on the Ganges and its main tributaries in the 4th century BC. In the 14th century, inland river navigation in the Ganges Basin was still flourishing. In the 19th century, irrigation-cum-navigation canals formed the main arteries of the water-transport system. The advent of paddle-steamers revolutionized inland transport, stimulating the growth of the indigo industry in Bihār and Bengal. Regular steamer services ran from Calcutta up the Ganges to Allahābād and far beyond, as well as to Āgra on the Jumna, and up the Brahmaputra River. Altogether, these services covered about 5,000 miles of waterways; they are still continued today under governmental auspices in both India and Pakistan. Much of the inland water traffic is carried on by various types of rural rivercraft.

The decline of large-scale water transport began with the construction of railways during the mid-19th century. The increasing withdrawal of water for irrigation has also affected navigation. Today, the river traffic is insignificant beyond the middle Ganges Basin around Allahābād.

West Bengal and Bangladesh, however, continue to rely on the waterways to transport jute, tea, grain, and other agricultural and rural products. Principal river ports are Chālna, Khulna, Barisāl, Chāndpur, Nārāyanganj, Goalundo Ghāt, Sirājganj, Bhairab Bāzār, and Fenchuganj, in Bangladesh; and in India, Calcutta, Goālpāra, Dhubri, and Dibrugarh. The partition of India and Pakistan in 1947 produced far-reaching changes, virtually halting the large trade in tea and jute formerly carried to Calcutta from Assam by inland waterway.

In Bangladesh, inland water transport is the responsibility of the Inland Water Transport Authority (IWTA). In India, the comparable authority is the Ganga Brahmaputra Water Transport Board.

The dispute over the Ganges waters

The construction of the Farakka Barrage at the head of the delta, just inside Indian territory in West Bengal, has been a bone of contention between India and Bangladesh. According to the Indian view, the port of Calcutta was deteriorating due to the deposit of silt and the intrusion of saline seawater. In order to ameliorate the condition of Calcutta by flushing away the seawater and raising the water level, India sought to have quantities of fresh water diverted from the Ganges at the site of the Farakka Barrage. The water there is now carried by means of a large canal into the Bhāgirathi River, which joins the Hooghly River, on the banks of which stands Calcutta.

According to Bangladesh, all riparian countries should exercise joint control over the waters of international rivers for the sake of mutual prosperity. The Ganges waters are also vital to irrigation, to navigation, and to the prevention of saline incursions in Bangladesh. Bangladesh has maintained that the Farakka Barrage deprives it of a valuable source of water upon which its prosperity depends. An uneven agreement on the Ganges waters dispute has been reached, however, between India and Bangladesh.

Hydroelectric power. The largest number of streams in the Ganges Basin originate in the Himalayas. Power generation from these rivers depends upon the extent to which, by using water storage, the river flows can be regulated.

The hydroelectric potential of the Ganges Basin has been estimated at about 4,800,000 kilowatts, representing an annual output of 25,400,000,000 kilowatt hours.

INDUS RIVER

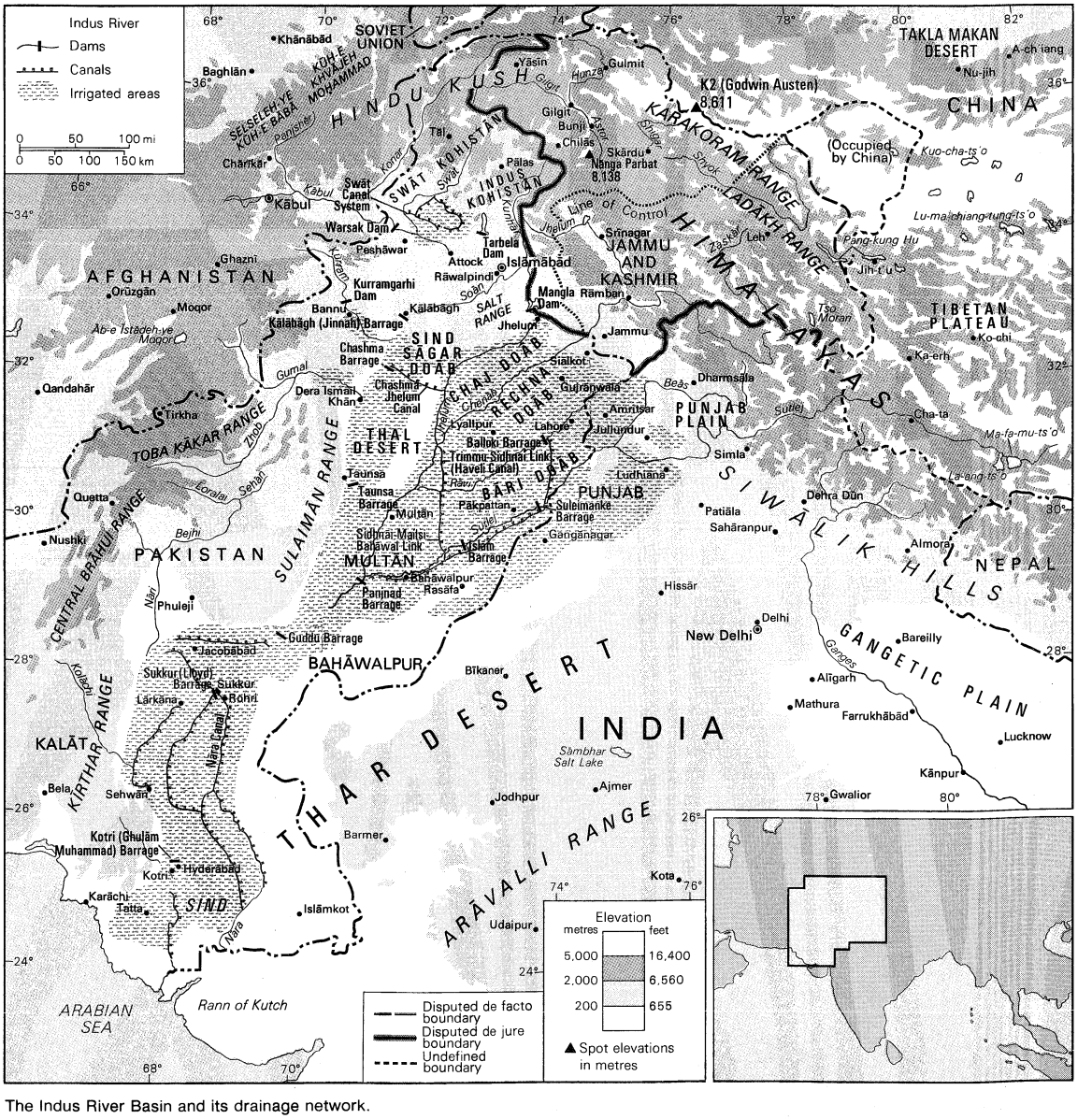
The Indus is a great trans-Himalayan river and one of the longest rivers in the world, having a length of 1,800 miles (2,900 kilometres). It has a total drainage area of about 450,000 square miles (1,165,500 square kilometres), of which 175,000 square miles lie in the Himalayan mountains and foothills and the rest in the semiarid plains of Pakistan. The river's annual flow is about 274,055,000,000 cubic yards (209,542,000,000 cubic metres)—twice that of the Nile and three times that of the Tigris and Euphrates combined. The river's name comes from the Sanskrit word *sindhu.* It is mentioned in the Rigveda, the earliest (*c.* 1500 BC) chronicles and hymns of the Aryan peoples of ancient India, and is the source of the country's name.

Physical features. The river rises in southwestern Tibet at an altitude of 16,000 feet (4,875 metres). For about 200 miles it flows northwest, crossing the southeastern boundary of Jammu and Kashmir at about 15,000 feet (4,600 metres). Eleven miles beyond Leh, in Ladākh, it is joined on its left by its first tributary, the Zāskār. Continuing for 150 miles in the same direction the Indus is joined by its notable tributary the Shyok on the right bank. After its confluence with the Shyok, and up to the Kohistān Mountains, it is fed by mighty glaciers on the slopes of the Karakoram Range, the Nānga Parbat Massif, and the Kohistān Ranges. The Shyok, Shigar, Hunza, Gilgit, and other streams carry the glacial waters into the Indus. Since the present-day precipitation of snow in this region is not sufficient to feed these great rivers of ice, it seems clear that the giant ice streams of the Karakoram are survivors of the last Ice Age of the Himalayas.

Glacial meltwater influx in the Upper Indus

The Shigar joins the Indus on the right bank near Skārdu in Baltistān. The Gilgit, farther down, is another right-bank tributary, joining it at Bunji. Some miles farther downstream, the Astor River joins as a left-bank tributary. The Indus then flows west, crosses the Kashmir border, and turns south and southwest to enter Pakistan. There

Irrigation data adapted from N. Ginsberg (ed.), *The Pattern of Asia* (© 1958); Prentice-Hall, Inc.



The Indus River Basin and its drainage network.

it skirts around the Nānga Parbat Massif (26,660 feet; 8,126 metres) in gorges as deep as 15,000 to 17,000 feet and 12 to 16 miles wide. Trails cling grimly to precipitous slopes overlooking the river from elevations of 4,000 to 5,000 feet.

After emerging from this region of high altitude, the Indus flows as a rapid mountain stream between the Swāt and Hazāra areas in Pakistan until it reaches the reservoir of Tarbela Dam, which was completed in 1975. The Kābul River joins the Indus just above Attock, where the Indus flows at an elevation of 2,000 feet and is crossed by the first bridge carrying rail and road. Finally, it cuts across the Salt Range near Kālābāgh to enter the Punjab Plain.

The Indus receives its most notable tributaries from the Punjab Plains to the east. These five rivers—the Jhelum, the Chenāb, the Rāvi, the Beās, and the Sutlej—give the name Punjab ("land of five rivers") to the land shared between Pakistan and India.

After receiving the waters of the Punjab rivers, the Indus becomes much larger and, during the flood season (July–September), is several miles wide. It flows here at an elevation of 259 feet. Its slow speed at this stage results in its accumulated silt being deposited on its bed, which is thus raised above the level of the sandy plain; indeed, most of the plain in Sind has been built up by alluvium laid down by the Indus. Embankments have been built to prevent flooding, but occasionally these give way, and large areas are destroyed by inundation. Such floods occurred in 1947 and 1958. During heavy flooding the river sometimes changes its course.

The Indus Delta

Near Tatta the Indus begins its deltaic stage and breaks into distributaries that join the sea at various points south-southeast of Karāchi. The delta covers an area of 3,000 square miles or more, and extends along the coast for about 130 miles. The uneven surface of the delta area is marked by a network of existing and abandoned channels. The coastal strip, from about five to 20 miles inland, is flooded by high tides.

Hydrography. The principal rivers of the Indus River system are snow fed. Their flow varies greatly at different times of the year: the discharge is at a minimum during the winter months, there is a rise of water in spring and early summer, floods occur in the rainy season (July–September), and occasionally there are devastating flash floods. The Indus and its tributaries receive all their waters in the upper hilly parts of their catchments. Therefore, their flow is at a maximum where they emerge out of the foothills, and little surface flow is added in the plains, where much loss of water occurs because of evaporation and seepage. On the other hand, some water is added by seepage in the period after the monsoon months. In the main stream of the Indus, the water level is at its lowest from mid-December to mid-February. After this the river starts rising, slowly at first and then more rapidly at the end of March. The high-water level usually occurs between mid-July and mid-August. The river then falls rapidly until the beginning of October, when the water level subsides more gradually. Annually, the Indus carries about 144,283,000,000 cubic yards—one-half of the total supply of water in the Indus River system. The Jhelum and Chenāb each carries about one-fourth as much as the Indus; and the Rāvi, Beās, and the Sutlej combined comprise about one-fifth of the total supply of the system (271,316,000,000 cubic yards).

Shifting course of the Indus

There is considerable physiographic and historical evidence to prove that since the dawn of civilization—at least since the days of Mohenjo-daro culture, 4,000 years ago—the Indus, from the southern Punjab to the sea, has been shifting its course. It is confined between limestone ridges at Rohri-Sukkur, but thereafter it has wandered, shifting generally to the west, particularly in its deltaic sector, so that about 200 years ago it began to flow into the Rann of Kutch. In upper Sind the Indus has shifted westward a distance of about 10 to 20 miles in the last seven centuries. The river is now held back to some extent by higher ground from Sehwān to Tatta at the head of the delta, but the possibility of future shifting cannot be ruled out. There is also evidence of the shifting of the Chenāb, Rāvi, Beās, and Sutlej rivers during the historical period.

Climate. From its source to its mouth, the annual rainfall in the Indus region varies between five and 20 inches (125 to 500 millimetres). Except for the mountainous section of Pakistan, the Indus Valley lies in the driest part of the subcontinent. Northwestern winds sweep the Indus Valley in winter (December to February) and bring four to eight inches of rainfall—vital for the successful growing of wheat and barley. The mountain region of the valley receives precipitation largely in the form of snow. A large amount of the Indus' water is provided by melting snows and glaciers of the Karakoram, Kohistān, and Himalayan mountains. The monsoon rains (July to September) provide the rest of the flow. The climate of the Indus Valley ranges from that of the dry semidesert areas of Sind and lower Punjab to the severe high mountain climate of Kohistān, Hunza, Gilgit, Ladākh, and western Tibet. January temperatures here are below freezing point in the north, while July temperatures reach a maximum of about 100° F (38° C) in Sind and Punjab. Jacobābād, one of the hottest spots on Earth, is situated west of the Indus River in upper Sind and often records summer maximums of 120° F (49° C).

Plant and animal life. There is a close relationship between climate and vegetation in the Indus Valley. In the Lower Indus region of Sind, desert conditions prevail 10 to 25 miles away from the river, and the area is dominated by poor grass and sand. Irrigation by floods or canals permits some cultivation. In upper Sind and Punjab, overgrazing and felling timber for fuel has led to destruction of both trees and vegetation. Further, prolonged human interference with natural drainage and deforestation on the Siwāliks has led to marked deterioration in groundwater conditions and so in vegetation. It appears that in prehistoric and earlier historic times the middle Indus region was more wooded than it is at present: accounts of Alexander the Great's Indian campaigns (*c.* 325 BC) and records of Mughal hunts in the 16th century and after suggest considerable forest growth. Even today, in the Indus Plains not far away from the river, there are thorn forests of open acacia and bush and undergrowth of poppies, vetch, thistles, and chickweed. Near the river are stretches of tall pampa-like grass, and streams and canals are often lined with tamarisk trees and some dense scrub, but there is nowhere a natural forest. In recent years afforestation of some parts of the Thal area in the Punjab east of the Indus has been commenced. Cultivated areas close to the river have many trees, and the strip below the mountains has something of the appearance of parkland. Coniferous trees abound in the Pakistan and Kashmir areas of the mountainous parts of the Indus Valley.

The Indus is moderately rich in fish. The best known variety is called *palla* and is the most important edible fish found in the river. Tatta, Kotri, and Sukkur, all in Sind, are the most important fishing centres. Between the Swāt and Hazāra areas the river is noted for trout fishing. In recent years, since the establishment of Pakistan, fish culture has been practiced in the reservoirs of dams and barrages. Close to the mouth of the Indus—for about 150 miles along the coast—there are numerous creeks and a shallow sea beyond. The area is rich in marine fish, the most important catches including pomfrets and prawns, which are obtained from November to March. A modern fish harbour has been built near the port of Karāchi, providing cold storage and marketing. In recent years an export trade in prawns has developed and sea fish are marketed in different parts of Pakistan.

The economy. *Irrigation.* Irrigation from Indus waters has provided the basis for successful agriculture since time immemorial. Modern irrigation engineering work commenced around 1850, and large canal systems were constructed by the British administration. In many cases old canals and inundation channels of Muslim and Sikh times were revived and modernized; thus the greatest canal irrigation system in the world was created. At partition in 1947, the international boundary between India and West Pakistan cut the irrigation system of the Bāri Doāb and the Sutlej Valley Project, originally designed as one scheme, into two parts. The headwork fell to India while the canals ran through Pakistan. This led to a disruption in the water

The Indus Waters Treaty

supply in some parts of Pakistan. The dispute that thus arose and continued for some years was resolved through the mediation of the World Bank by a treaty between Pakistan and India (1960) known as the Indus Waters Treaty. According to this agreement, the flow of the three western rivers of the Indus Basin—the Indus, Jhelum, and Chenāb (except a small quantity used in Kashmir)—will be utilized exclusively for Pakistan, whereas the entire flow of the three eastern rivers—the Rāvi, Beās, and Sutlej—will be used by India. In accordance with the settlement, Pakistan was permitted to build storage dams, canals, and barrages and to operate a drainage scheme, using tube wells that make underground water available for surface irrigation.

The Mangla Dam on the Jhelum River, near the town of Jhelum, has a crest length of about 11,000 feet and a maximum height of 380 feet and is one of the largest rolled earth-fill dams in the world. The reservoir created by the dam is 40 miles long and has an area of 100 square miles. The project also includes a powerhouse. Mangla Lake is being developed as a fishing centre and a tourist attraction as well as a health resort. The Mangla Watershed Management Organization has been formed to control the silting of the lake by using various techniques of soil conservation and afforestation.

A second gigantic project is the Tarbela Dam on the Indus, 80 miles northwest of Rāwalpindi. The dam, of the rock-filled type, is 9,000 feet long and 485 feet high, and its reservoir is 50 miles long.

Eight new link canals and five barrages have been completed by the Pakistan Water and Power Development Authority. The biggest of these link canals is the Chashma-Jhelum link joining the Indus River with Jhelum, with a discharge capacity of 21,700 cubic feet per second. Water from this canal is used to feed the Haveli Canal and Trimmu-Sidhnai-Mailsi-Bahawal link canal systems, which provide irrigation to the Multān and Bahāwalpur divisions in the lower Punjab.

On the Indus itself, there are five important headworks, or barrages, after the river reaches the plain. In the hilly region, the principal canals west of the Indus are the Swāt Canals, which flow from the Swāt River, a tributary of the Kābul River. These help in the irrigation of the two chief crops of the area, sugarcane and wheat. The Warsak multipurpose project on the Kābul River, about 20 miles northwest of Peshāwar, provides irrigation for food crops and fruit orchards in the Peshāwar Valley and is designed to produce 240,000 kilowatts of electricity. In the plain region, the Kālābāgh or Jinnah Barrage controls the system of canals in the Thal Project, the development authority for which was set up in 1949. The project irrigates a former desert area. It is an integrated project aiming at the extension of agriculture, the development of rural industry, and the settlement of population in villages and towns. Farther downstream in the Dera Ghāzi Khān district is the Taunsa Barrage, designed for the irrigation of land in the Dera Ghāzi Khān and Muzaffargarh districts. It was opened in 1959 and produces about 100,000 kilowatts of electricity. The power helps to operate tube wells. Within the Sind there are three major barrages on the Indus—Gudu, Sukkur, and Kotri, or Ghulām Muḥammad. The Gudu Barrage is just inside the Sind border and is 4,445 feet (1,355 metres) long; it irrigates cultivated land in the region of Sukkur, Jacobābād, and parts of Lārkāna and Kalāt districts. The project has greatly increased the cultivation of rice, but cotton is rapidly becoming the major crop on the left bank of the river and has replaced rice as a cash crop. The Sukkur Barrage was built in 1932 and is about a mile long. Four canals originate on the right bank and three on the left. They serve a cultivable area of about 5,000,000 acres (2,023,500 hectares) of land producing both food and cash crops. The Kotri Barrage, also known as the Ghulām Muḥammad Barrage, was opened in 1955. It is near Hyderābād and is nearly 3,000 feet long. The right-bank canal provides additional water to the city of Karāchi. Sugarcane cultivation has been extended, and crop increases have been achieved in the cultivation of rice and wheat.

Experience in the Indian subcontinent and elsewhere has shown that canal irrigation, unless carefully controlled, can cause much damage to the cultivated land. The water in the unlined canals seeps through the soil and raises the water table so that the soil becomes waterlogged and useless for cultivation. As irrigation by canals has expanded in the Indus and its tributary lands, in some areas underground water has appeared on the surface to form shallow lakes. Elsewhere the water has evaporated in the intense summer heat, leaving behind layers of salt that make crop production impossible. Several careful surveys have been made, including one under the Colombo Plan, and have revealed that, in the past, some of the Indus Plain has been poorly drained or waterlogged and that parts have been affected by salinity. To check this, adequate drainage and the construction of a network of tube wells is necessary. The water drawn out of the tube wells lowers the water table and flushes the surface salts down into the earth. An effective drainage system is therefore essential.

Waterlogging and salinity

Navigation. Until about 1880 the Indus and the other Punjab rivers carried some navigation, but the advent of the railways and expansion of irrigation works has eliminated all but small craft that ply the Lower Indus in Sind. There are fishing boats on the Lower Indus, and the upper reaches of rivers and canals above the first railway crossing are now used for floating timber down from the foothills of Kashmir. (N.A.)

IRRAWADDY RIVER

Irrawaddy, the principal river of Burma and its most important commercial waterway, is about 1,300 miles (2,090 kilometres) long. Its name is believed to have derived from the Hindu *airāvatī* meaning "elephant river." From its sources to its numerous mouths, the river flows wholly within Burmese territory. Its total drainage area is about 158,700 square miles (411,000 square kilometres). Its valley forms the historical, cultural, and economic heartland of Burma.

Physical features. The river is formed by the confluence of the Nmai Hka and the Mali Hka (*hka* being the Kachin word for river). Both branches rise in the glaciers of the high and remote mountains in northern Burma in the vicinity of 28° north. The eastern branch, the Nmai, rises in the Languela glacier and has the greater volume of water but is virtually unnavigable because of its strong current. The Mali, the western branch, has a gentler gradient and, although interrupted by rapids, has some navigable sections.

The river. About 30 miles south of the confluence is Myitkyina, the point of the northernmost limit of seasonal navigation by the Irrawaddy steamers. Bhamo, about 150 miles south of the confluence, marks the northern limit for year-round navigation. Between the confluence and Bhamo, the width of the river during the low-water season varies between a quarter of a mile and half a mile. The depth of the main channel averages about 30 feet (nine metres).

Between Myitkyina and Mandalay, the Irrawaddy flows through three well-marked defiles (narrow passages or gorges). Just south of Sinbo the river enters the upper, or third, defile—a navigational hazard, with a channel only 50 yards wide at its narrowest. Below Bhamo the river makes a sharp westward swing, leaving the Bhamo alluvial basin to cut through the cretaceous limestone rocks of the second defile. The second defile is about 300 feet wide at its narrowest and is flanked by vertical cliffs about 200 to 300 feet high. South of Thabeikkyin the river enters the first defile, which is broader and easier to navigate. Between Katha and Mandalay, the course of the river is remarkably straight, flowing almost due south, except near Kabwet, where a sheet of lava has caused the river to bend sharply westward. Leaving the first defile at Kyaukmyaung, in the Shwebo District, the river follows a broad, open course through the Dry Zone—the ancient cultural heartland—where large areas consist of alluvial flats. From Mandalay, formerly the capital of Upper Burma, the river makes an abrupt westward turn before curving southwest to unite with the Chindwin River in the Pakokku District, after which it continues in a southwesterly direction. It is probable that the upper Irrawaddy originally flowed south from Mandalay, discharging its water through the present

The Dry Zone

Sittang River to the Gulf of Martaban, and that its present westward course is geologically recent. After its confluence with the Chindwin, the Irrawaddy continues to meander through the densely populated Dry Zone to the vicinity of Yenangyaung, after which it flows generally south. In its lower course, between Minbu and Prome, it flows through a narrow valley between forest-covered mountain ranges—the Arakan Yoma ridge to the west and the Pegu Yoma ridge to the east.

The delta. The delta of the Irrawaddy may be said to begin about 58 miles above Henzada. The apex of the delta is about 180 miles from its curved base facing the Andaman Sea. The sides of the delta are formed by the southern extremities of the Pegu Yoma on the east and the Arakan Yoma on the west. The westernmost distributary of the delta is the Bassein River. The easternmost stream is the Rangoon River, on the left bank of which the Burmese capital city of Rangoon is built. As the Rangoon River is only a minor channel, the flow of water is insufficient to prevent Rangoon Harbour from silting up, and dredging is consequently necessary.

Hydrography. The volume of flow of the Irrawaddy and its tributaries fluctuates greatly through the year, chiefly due to the monsoon rains, which occur between June and September, but also to the rapid melting of glaciers during the summer, which adds still further to the volume. The average discharge of the river near the head of the delta varies between a low of 82,000 and a high of 1,152,000 cubic feet (2,300 and 32,600 cubic metres) per second; the annual average discharge is 460,000 cubic feet per second. The range between high and low water is also great. Annual variations between low-water level and flood level of 31.7 feet and 37.3 feet have been recorded at Mandalay and Prome respectively. The lowest water level occurs in February, and the highest in August. In general, from December to March the river varies between the lowest level and five feet above it, while from mid-June to mid-October the river is 20 to 30 feet (six to nine metres) above the lowest level. The river ports therefore find it necessary to have separate high-and low-water landing points.

The people. The peoples living on the river's banks are culturally diverse. On the upper reaches, the Kachins, who practice shifting agriculture, predominate. In the river valley itself, the Burmese are the dominant group, cultivating wheat, cotton, and oilseeds in the Dry Zone, and rice to the south and in the delta region, where rainfall is more plentiful. Also to the south, and particularly in the delta proper, a considerable minority of Karens and some Indians are to be found among the Burmese majority.

The river ports

The economy. *Navigation.* The main river ports on the Irrawaddy, from north to south, are Myitkyina, Bhamo, Katha, Mandalay, Myingyan, Chauk, Yenangyaung, Minbu, Magwe, Thayetmyo, Prome, Henzada, and Yandoon. Of these, Mandalay, Chauk, Prome, and Henzada have good landing facilities. The remaining ports have landing facilities for only one or two barges or lighters—the vessels mooring alongside the riverbank in most places. Despite the fact that Mandalay is the chief rail and highway focus in Upper Burma, a considerable amount of passenger and goods traffic moves by river. The Chindwin valley has no railroad and relies heavily on river transport. Chauk, downstream from the confluence in the oil-field district, is a petroleum port. Prome, about 140 miles to the south, is linked to Rangoon by road and rail. Henzada, near the apex of the delta, is the rail junction for lines leading to Kyangin and Bassein. A ferry operates between Henzada on the west bank and the railway station at Tharrawaw on the east.

Regular steamer service on the Irrawaddy is maintained by the Inland Water Transport Board (IWTB), which has improved service by introducing diesel engines in place of coal-fired ones. Commercial transportation is maintained for about 800 miles. From Henzada to Bhamo (670 miles), commercial traffic is maintained throughout the year, but from Bhamo to Myitkyina (125 miles), for only seven months. Over 2,000 miles of navigable waterways exist in the Irrawaddy delta. On the Chindwin River transportation is carried on by steam or diesel vessels throughout the year up to Homalin—about 400 miles from its confluence with the Irrawaddy. Seasonal navigation is carried on into Tamanthi, which is 57 (river) miles (91 kilometres) above Homalin.

As the Irrawaddy delta is one of the world's major rice-growing areas, rice is a major item of commerce on the river. Also transported are other foodstuffs, petroleum, cotton, and local commodities. Teak logs—of which Burma is the world's major exporter—are floated downstream in the form of large rafts. In the delta region, the rice is carried in small country boats to local markets, from where it is shipped to Rangoon for export.

The Irrawaddy is crossed mainly by the Ava Bridge, which spans the river below Mandalay.

Irrigation. Although the Irrawaddy has been little used for irrigation in the Dry Zone, its tributary, the Mu River, has been used for this purpose since the 9th century. The current Mu Valley Irrigation Project, financially supported by the UN Development Programme, is the largest in the country. It permits the dry-season cropping of maize (corn), groundnuts (peanuts), sesame, millet, and other dry crops.

The soils of the delta are enriched every year by the fertile alluvium carried down by the river. The river waters themselves are used to flood the delta rice paddies. "Irrawaddy" is also the name of an administrative division of the delta, which includes the districts of Bassein, Henzada, Ma-ubin, Myaungmya, and Pyapon. (S.M.B.)

AMUR RIVER

The Amur River (Wade-Giles romanization Hei-lung Chiang; Pinyin Heilong Jiang) in eastern Asia forms part of the frontier between the Union of Soviet Socialist Republics and the People's Republic of China. It rises on the northern border of Northeast China and Inner Mongolia with the Soviet Union and flows generally east and southeast along that border to Khabarovsk, where it turns northeastward and flows to the Tatar Strait, separating Siberia from the island of Sakhalin. It is the longest river in the Soviet Far East. Its Chinese name, Hei-lung Chiang, means Black Dragon River; its Mongol name, Kharamuren, means Black River; among local tribes it is known as Mamu.

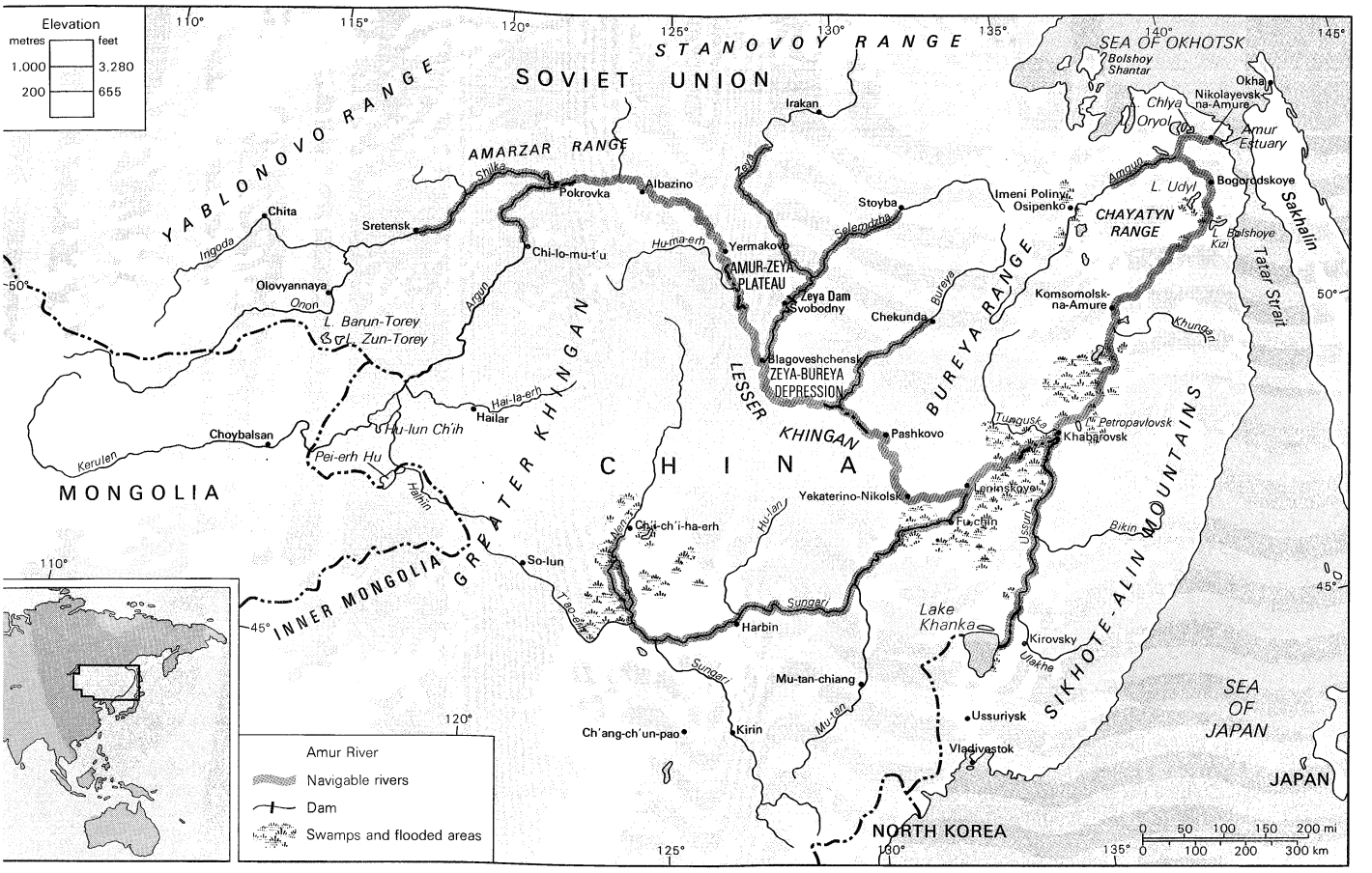
Physical features. *Physiography.* The Amur begins at the confluence of the Shilka and the Argun rivers, at a point 1,755 miles (2,824 kilometres) from its mouth. The Shilka rises more than 340 miles farther inland at the junction of the Siberian Ingoda River and the Mongolian Onon River. The Argun River rises in the Chinese autonomous region of Inner Mongolia, about 1,000 miles from its confluence with the Shilka. The entire river system of the Amur includes numerous rivers and lakes; among the rivers are five major ones—the Zeya, Bureya, Sungari, Ussuri, and Amgun.

It is customary to treat the river as divided into three sections: the Upper, Middle, and Lower Amur. The Upper Amur begins at the juncture of the Shilka and Argun and ends at the mouth of the Zeya (Blagoveshchensk), about 550 miles (880 kilometres) downstream. The Middle Amur extends about 600 miles (960 kilometres) from the Zeya north to Khabarovsk. The Lower Amur, from Khabarovsk to the mouth, is also about 600 miles in length.

The Upper Amur River

The Upper Amur flows through a mountain valley between the spurs of the Greater Khingan, which is covered by thick larch woods, and the pine-clad slopes of the Amarzarsky mountain ridge. Near Albazino the mountains part, and the river enters open-plateau country. The winding river has built up sandbanks, many of them really spits of land formed at looping bends of the river, which in places has cut deeply into the earth. The terraced slopes of the valley testify that in remote times the land was flat and the Amur wandered through wide floodlands. Below Yermakovo lie the Burning Mountains, rocky precipices made up of complex layers of spontaneously igniting carbonaceous, clayey shales that are continually steaming and sometimes give off flames.

Middle Amur flows into the Zeya-Bureya Depression. The left slope of the valley merges imperceptibly with the plain, while the right slope—steep and high—borders the Lesser Khingan Range. Below the Bureya River the plain



The Amur River Basin and its drainage area.

narrows gradually, and near Pashkovo the river runs past spurs of the Bureya Range. Farther on it flows along a narrow ravine-like passage through the Lesser Khingan mountains, its depth and speed sharply increasing.

The Lower Amur runs between low, overflowing banks into a vast marsh, the surface of which is broken by channels and dotted with lakes and ponds; the riverbed branches often, and the channel becomes very wide. Near Novoye the Sungari, the Amur's largest tributary, pours in its yellow, muddy waters. Near Khabarovsk it is joined by the Ussuri. With these accessions to its waters, it overflows widely over the flat, marshy ground of the tidal valley. The riverbed becomes a labyrinth of branches, channels, offshoots, former riverbeds, islands, sandbanks, and spits. During high-water seasons these sections overflow, and the region becomes an enormous lake. At Khabarovsk the Amur is only 230 miles from the shores of the Sea of Japan; but, diverted by the Sikhote-Alin Mountains, it runs northward for 600 miles before it finds the sea. Near Komsomolsk-na-Amure the plain gradually narrows, and the river flows for 90 miles among mountains into a scenic forest valley. It then enters the Udil Kizinsky hollow with wide, flat, marsh-ridden ground. In the hollow lie two large lakes—the Big Kizi and the Udyl. Near Bogorodskoye the hollow is closed in by the Chayatyn Range, and the river flows out onto a low-lying plain, where the Amgun, the last of its important tributaries, joins the Amur on its left bank. It enters the sea through a wide, bell-shaped estuary, which is about 30 miles long.

Climate. The Amur Basin has a monsoon climate—a seasonal alternation of winds from the mainland and from the ocean. In winter the dry, cold air moving down from Siberia brings clear, dry weather with strong frost. Snowfall is generally light. In summer the warm, moist ocean winds predominate, bringing heavy rains and thus raising the water level in the Amur and its main tributaries.

Autumns are warm and dry. Average temperatures, in January, range from −11° F (−24° C) in the south to −27° F (−33° C) in the north. The average July temperature in the south (at Blagoveshchensk) is 70° F (21° C) and in the north about 64° F (18° C). Precipitation in the Amur Basin is uneven, being heaviest in the maritime section, where it ranges between 24 and 36 inches (600 and 900 millimetres) annually. In the middle regions the annual rate does not exceed 24 inches, and in the western, continental regions it averages 12–16 inches. The peak months are July and August, when there are four or five inches of rainfall.

Wide range of precipitation

Marine life. The Amur is rich in fish. The lower river has about 100 species of fish, the upper river 60, surpassing even large European rivers such as the Volga and the Danube. About 25 or 30 species are of commercial value. Northern forms include the Siberian salmon, the sig, and the burbot; southern forms include the Chinese perch and the white amur. There are about 20 indigenous species of carp and broadhead. A biological peculiarity of the Amur fishes is the large number of species that develop in the sea and therefore escape exposure to the sharp changes in water level that occur in the river in summer.

Hydrology. The river is fed principally by the monsoon rains that fall in summer and autumn. The rainwater finds its way quickly into the river, resulting in a period of flooding that extends from May to October. During that period there are usually several times of high water when the river is from 16 to 26 feet (5 to 8 metres) above its usual level. In particularly rainy years the high-water level may be as much as 45 feet above the usual level in the Upper Amur. Farther downstream the variations are less; in the lower reaches, the maximum rise is only about 8 feet. In August and September the floodwaters begin to abate. The lowest level of the river is reached in March and April, before the spring flood, which is fed mainly by runoff of melted snow and which is much smaller than the monsoon floods that occur in summer and autumn. Those floods often cause serious economic loss.

Seasonal flooding

The mean discharge of the Amur is 380,000 cubic feet (10,900 cubic metres) per second. The rate of flow near Komsomolsk-na-Amure is about 348,200 cubic feet (9,860

cubic metres) per second and at the mouth about 383,800 (10,900). The rate varies from as low as 5,300 cubic feet (150 cubic metres) per second to 7,000 (200) in winter near Khabarovsk; the highest rate ever recorded was 1,400,000 (40,000) in 1897.

Ice forms in the Amur in the second half of October. The Upper Amur becomes icebound at the beginning, the Lower Amur in the second half of November. The lower reaches of the river open at the end of April, the upper reaches at the beginning of May. Ice jams often occur in the sharp bends of the river, raising the water level by as much as 50 feet.

At the mouth of the Sungari, the water in the Amur is comparatively clear, but lower down it becomes muddy. In a year the river carries down about 20,000,000 tons of sediments.

The people. The territories of the Amur were originally populated by hunting and cattle-breeding nomadic tribes, such as the Buryats, Yakuts, Nanais, Gilyaks, Udegeys, and Oroks, From the 18th century onward the area north of the Amur was settled by Russians along with Ukrainians, Belorussians, Tatars, Latvians, and others. South of the Amur live Chinese, Mongols, Manchus, and many other peoples.

The Russians began to explore the Amur Basin in the 17th century. Maksim Perfilyev passed through the region in 1638. In the spring of 1644 Vasily Poyarkov set off on a three-year exploration of the Amur; he covered much of the river basin and estuary. Yerofey P. Khabarov followed in 1649–51; Khabarovsk—the second largest city in the Soviet Far East—is named for him.

In 1849–55 an expedition under the leadership of a Russian naval officer, Gennady I. Nevelskoy, made the discovery that Sakhalin is an island and that, therefore, the Amur is accessible from the south and not alone from the north as the Russians had previously supposed. This was the beginning of the systematic study of the water system of the Amur. A series of large expeditions was planned and carried out in the first part of the 20th century (occupying the years 1901–09, 1927–29, 1932, 1935–36). In 1952–58 the Academy of Sciences of the U.S.S.R. sponsored a thorough exploration of the Amur.

The economy. The Amur is navigable throughout its course for about half the year. The large tributaries—the Zeya, Bureya, Sungari, Tunguska, Ussuri, and Amgun—are navigable, as are some of the lakes. The main ports include Pokrovka, Blagoveshchensk, Leninskoye, Khabarovsk, Komsomolsk-na-Amure, and Nikolayevsk-na-Amure. The potential hydroelectric power of the Amur River Basin is estimated at roughly 45,000,000 kilowatts, but little of this potential has so far been developed.

(A.P.M.)

CHINA SEA

The China Sea is that part of the western Pacific Ocean bordering the east-southeast Asian mainland. It consists of two seas, the South China Sea and the East China Sea, which connect through the shallow Taiwan Strait between Taiwan and the People's Republic of China.

East China Sea. The East China Sea, or Eastern Sea, known in Chinese as the Tung Hai (Pinyin Dong Hai), extends northeastward from the South China Sea. Eastward it includes waters out to the Ryukyu Island chain; north to Kyūshū, which is the southernmost of Japan's main islands; northwest to the island of Cheju-do off South Korea; and hence west to the mainland of China. This northern boundary separates the East China Sea from the Yellow Sea.

Boundaries of the East China Sea

Physiography. The East China Sea, which has an area of 290,000 square miles (752,000 square kilometres), is largely shallow, with 71 percent of the area less than 650 feet (200 metres) and an average depth of only 1,145 feet. The deeper part is the Okinawa Trough, extending alongside the Ryukyu Island chain, with a large area over 3,300 feet (1,000 metres) deep and a maximum depth of 8,912 feet (2,717 metres). The western edge of the sea is a continuation of the shelf that extends from the South China Sea up to the Yellow Sea. A large number of islands and shoals dot the eastern boundary as well as the area near the Chinese mainland. The shallow shelf areas are covered with sediments from the bordering landmasses deposited mainly by the Yangtze and other rivers near the northern part of the sea. Coarser sediments of sand occur farther out, and rocks, muds, and oozes are also found in scattered areas. Seismic profiling indicates that the geological subbottom structure comprises nearly parallel folds, with rock ridges near the northern limits of the East China Sea, near the edge of the continental shelf, and along the Ryukyus. These have afforded barriers for sediment brought down by the great Huang (Yellow) and Yangtze rivers. Between the ridges the sediment is up to one mile thick.

Most of the shelf belongs to the stable Neo-Cathaysian Geosyncline (or Cathaysian Platform), dating back at least 300,000,000 years or longer. The Okinawa Trough appears to be perhaps 10,000,000 years old. The Ryukyus are a double island chain, with several volcanic islands on the East China Sea side. Many of the volcanoes are still active. Epicentres of earthquakes are found along the axis of the Okinawa Trough and the Ryukyu Island arc.

Climate. Weather is dominated by the monsoon wind system, the result of differential heating between land and water. In summer, the high Asian landmass is much warmer than the sea; in the winter, it is much colder, particularly in the Plateau of Tibet. Summer heating of air masses over Asia builds areas of low pressure and creates the monsoonal winds, which in this season blow predominantly from the southeast. This brings in warm, moist air from the western Pacific Ocean, producing a rainy summer season that is accompanied by typhoons. In winter the situation is reversed: monsoon winds blow predominantly from the north, bringing with them cold, dry air from the continent.

The China Sea.

Sea currents and tides

Hydrography. Winds also influence water circulation of the Kuroshio (Japan) current, a north-flowing branch of the warm North Equatorial Current that flows near Taiwan. Some of the Kuroshio enters the eastern part of the East China Sea, then diverts eastward back out into the Pacific, and flows east of Japan. Strengthened by monsoon winds, it is at its widest and fastest in summer, and the axis is displaced well into the East China Sea. This warmed surface water varies from 86° F (30° C) in the south to 77° F (25° C) in the north. In winter, northerly monsoon winds modify the circulation, and the north-flowing Kuroshio, though still important, is reduced in strength, while southerly flowing coastal currents are strengthened. This brings in colder water, with temperatures of 41° F (5° C) in the north to 73° F (23° C) in the south.

Because of the constricting nature of the adjoining Yellow Sea and the funnel shape of some of the inlets on the mainland, tidal ranges are especially high along the coast of China. For example, the spring tide range, which is highest in summer and winter, is as much as 23 feet at San-sha Bay and 36 feet at Hangchow Bay.

The economy. The East China Sea is a highly marine-organic productive region, with China, Japan, Taiwan, and North and South Korea actively fishing in the area. Most of the fishing is done by small local boats, although larger trawlers are also used. Tuna, mackerel, shrimps, sardines, milkfish, sea breams, croakers, shellfish, and seaweeds are the main resources harvested. The records of fishing intensity correlate well with the preference of bottom fishes for areas of fine-grain, organic-rich sediments that are probably the result of the greater abundance of worms and other small, mud-digesting organisms in the area.

In addition to the local shipping traffic in and out of Chinese and Korean ports, the East China Sea serves as the main shipping route from the South China Sea to Japanese and other North Pacific ports.

Boundaries of South China Sea

South China Sea. The South China Sea, known in Chinese as Nan Hai (Pinyin Nan Hai), is bounded on the west by the Asian mainland, on the south by the southern limit of the Gulf of Thailand and the east coast of Malay Peninsula, and on the east by Taiwan, the Philippines, and Borneo. The southern boundary is a rise in the seabed between Sumatra and Borneo, and the northern boundary stretches from the northernmost point of Taiwan to the coast of Fukien Province, China (the southern limit of the East China Sea). It embraces an area of about 848,400 square miles, with an average depth of 3,740 feet.

M. Gifford—De Wys Inc.

A portion of the delta of the Mekong River as it flows through southern Vietnam and empties into the South China Sea.

Physiography. The major topographic feature is a deep rhombus-shaped basin on the eastern part, with reef-studded shoal areas rising up steeply within the basin to the south (Reed, Tizard, Nanshan banks) and northwest (Paracel Island and Macclesfield banks). The deep portion, called the China Sea Basin, has a maximum depth of 16,452 feet, and an abyssal plain with a mean depth of 14,100 feet. The continental shelf falls off sharply near Luzon and Palawan islands and forms the Palawan Trough near the latter island.

Along the northwest side of the basin to the mainland is a broad, shallow shelf as wide as 150 miles. It includes the Gulf of Tonkin (maximum depth of 269 feet) and the Taiwan Strait. The large islands of Hainan and Taiwan are situated on this shelf. To the south, off Vietnam, the shelf narrows and connects with the Sunda Shelf, one of the largest in the world, which covers the area between Borneo, Sumatra and Malaysia, and includes the southern part of the South China Sea, the Gulf of Thailand, and the Java Sea. This broad trough is about 130 feet deep at its periphery and up to 330 feet in its central part. On the bottom of the shelf is a network of submerged river valleys that converge into the Sunda Depression and then into the China Sea Basin. These valleys and tributaries vary in width up to three miles.

The Sunda Shelf is covered with littoral sediments contributed by submerged valleys. The inner zone of mud is characteristic of the continental shelf near the Mekong and Red River deltas, while the sediment of the deeper parts of the South China Sea is mainly composed of clay. A characteristic part of the sediments in both deep and shallow water is volcanic ash. This is found in layers, derived from large volcanic eruptions in the East Indies, notably the enormous eruption of Krakatoa in 1883, when ash was transported through the entire area by both wind and currents.

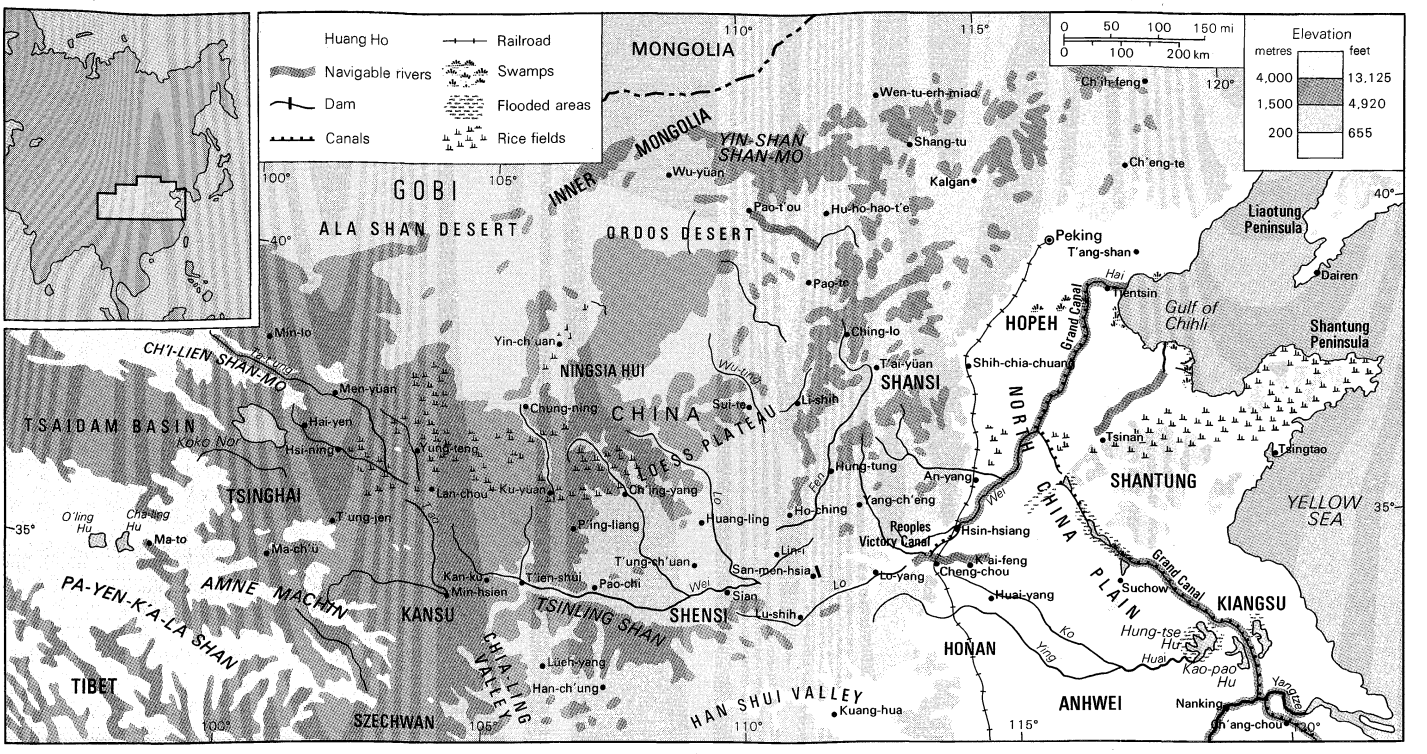
Connecting channels

The South China Sea has connecting channels. The Formosa Strait on the north is about 90 miles wide, with a depth of about 230 feet. The main deep channel connecting the South China Sea with the Pacific Ocean lies between Taiwan and the Philippines and has a depth of about 8,500 feet. Shallow channels are found on the east along the Philippine Island chain and on the south between Borneo and Sumatra. The western connection to the Indian Ocean is the long Strait of Malacca. At its narrowest part it is 19 miles wide and about 100 feet deep. The South China Sea is the largest "marginal sea" of the western Pacific. Some 1,000,000 to 60,000,000 years ago, it was rifted and collapsed as a result of seafloor spreading. The China Sea Basin is believed to have dropped 2.5 miles, leaving residual plateaus studded with numerous coral reefs, islets, and banks, some of which are drowned atolls.

Climate. Weather is tropical and largely controlled by monsoons. In summer, monsoonal winds blow predominantly from the southwest; in winter, winds blow from the northeast. Annual rainfall approximates 12.3 feet (3,750 millimetres), and summer typhoons are frequent.

Hydrography. Monsoons control the sea-surface currents as well as the exchange of water between the South China Sea and adjacent bodies of water. In August, the surface flow into the South China Sea is from the south through Karimata and Kelasa (Gasper) straits. Near the mainland the general flow is northeasterly, passing out through the Taiwan and Luzon straits. There is a weak countercurrent on the eastern side of the sea. In February, the flow is generally to the southwest; the strongest flow occurs along the bulging part of Vietnam, with speeds of up to three knots (nautical miles per hour).

The near-surface waters are relatively warm (about 84° F [29° C] in the summer) because of the low latitude and a tendency for the equatorial current to feed warm water into the area. In early summer, wind from the southwest not only moves the surface water to the northeast but causes it to be displaced off the coast. As a result,



The Huang Ho Basin and its drainage network.

upwelling areas having colder surface temperatures and higher nutrient content are found off central Vietnam. In winter, the general surface temperature is colder, ranging from about 70° F (21° C) in the north to 81° F (27° C) in the south.

The major rivers draining into the South China Sea are the Hsi Chiang, which enters near Macau; the Red River at Haiphong, Vietnam; and the Mekong River, near Ho Chi Minh city (Saigon), Vietnam. The wet summer season causes the Mekong River to triple its annual average flow, and it causes an even greater relative change in the flow of the Red River.

The economy. The heavily fished South China Sea is the main source of animal protein for the people living near its shores, providing as much as 50 percent for the densely populated Southeast Asian area. Most abundant are the various species of tuna, mackerel, croaker, anchovy, shrimp, and shellfish. Nearly the entire catch is consumed locally, either fresh or preserved.

Fishing

The main transport route to and from Pacific and Indian ocean ports is through the Strait of Malacca and the South China Sea. In the main, oil and minerals move up the coast, and food and manufactured goods move down. Some areas in the central South China Sea are not well sounded, and nautical charts bear the notation "dangerous ground." (E.C.LaF.)

HUANG HO

The Huang Ho (Huang He in Pinyin), or Yellow River, one of the most northerly and the second longest of China's rivers, flows from the eastern highlands of the Tibetan Autonomous Region along a course of 3,011 miles (4,845 kilometres) to the Yellow Sea. It passes through the provinces of Tsinghai and Kansu, the autonomous regions of Ningsia Hui and Inner Mongolia, and then into the provinces of Shensi, Shansi, Honan, and Shantung. For much of its length it is a shifting, turbulent, muddy stream that often overflows its banks and sends floodwaters across the North China Plain. For this reason it has been called "China's Sorrow" and "The Ungovernable." The name Huang (Yellow) comes from the fine, loess soil that it brings down with it. More than 110,000,000 people live in the basin of the Huang Ho, which flows past the cities of Lan-chou, Pao-t'ou, Sian, T'ai-yüan, Lo-yang, Cheng-chou, K'ai-feng, and Tsinan. It drains an area of 288,000 square miles (745,000 square kilometres).

Physical features. The Huang Ho is divided into two distinct parts: the mountainous upper basin and the plains section, which is itself subdivided into the middle and lower basins. The broad highlands of the mountainous part rise 1,000 to 1,300 feet (300 to 400 metres) above the river and its tributaries. The highlands are built upon crystalline rocks that are sometimes visible as eroded outcroppings on the surface. On the plateau these rock systems are covered with thick layers of friable deposits, consisting mainly of loess. The loess strata have thicknesses of 160 to 200 feet and in some places as much as 500 feet, extending eastward from the highlands of Tibet all the way to the North China Plain. Through this loose loam the river has cut deep valleys, carrying away with it huge quantities of silt; the easily eroded loess soil accounts for the instability of the riverbed.

Downstream the Huang Ho Basin becomes fan shaped as it broadens out across the North China Plain, which is broken only by the low Shantung Hills. The plain consists of fine silt, brought down by the river, laid over sand and gravel that was deposited when the sea receded in the geological past. Along the river are occasional areas of sand dunes 15 to 30 feet (4.5 to nine metres) in height. The plain, which contains a number of old beds of the river, is China's rice granary.

Physiography. The Huang Ho originates at an altitude of about 15,000 feet in the Pa-yen-k'a-la Shan (Pa-yen-k'a-la Mountains), in the eastern Tibetan Highlands. In its upper reaches the river crosses two large lakes, Cha-ling Hu and O-ling Hu. These are shallow lakes covering an area of about 40 square miles, rich in fish and frozen in winter. The Huang Ho in this region flows generally from west to east. It enters deep gorges winding its way along the southern slopes of the Amne Machin, somewhat above the city of Lan-chou. With many rapids, its fall exceeds 10 feet per mile. Populated areas are few. Past the gorges it leaves the Tibetan Highlands and flows northward across the Ordos Desert, in northern China, having flowed about 726 miles from its source. The basin upstream from that point covers an area of about 48,000 square miles, consisting chiefly of inaccessible, highly mountainous terrain with a cold climate. The few inhabitants engage in cattle breeding.

The upper basin

The middle part of the Huang Ho consists of a great northward loop through the Ordos Desert and then a southward course through a long trench forming the bor-

The middle basin

der between Shensi and Shansi provinces to the city of Cheng-chou, a distance of more than 1,800 miles. Most of the middle basin is cut through the Loess Plateau at altitudes ranging between 3,000 and 7,000 feet. The plateau contains terraced slopes as well as alluvial plains and a scattering of peaks sometimes exceeding 1,500 feet in height. Within its large loop, the Huang Ho drains an area of about 23,000 square miles. The river at first flows in a northeasterly direction for about 550 miles among bare loess hills; it has many rapids, and in a number of places it narrows. It then turns eastward and flows for another 500 miles through alluvial plains; in this stretch it has many branches, and its fall is less than four inches per mile. If it continued eastward it would flow to Peking, but instead it turns sharply to the south and flows for about 445 miles through narrow gorges with steep slopes several hundred feet in height. The river's width usually does not exceed 150–200 feet in this section. The Huang Ho gradually widens and then turns sharply to the east for another 300 miles. Here it flows through inaccessible gorges between the eastern peaks of the Tsinling Shan. The average fall in this stretch is about 14 inches per mile and increasingly rapid in the last 100 miles before the river enters the North China Plain.

This middle basin of the river, between the Tibetan Highlands and the North China Plain, contains the two longest tributaries: the 537-mile Wei Ho of Shensi Province and the 430-mile Fen Ho of Shansi Province.

The lower basin

The lower part of the Huang Ho is about 435 miles long with an average fall of about three inches per mile. This is an area of great floods because the riverbed in many places lies above the surrounding land. In the section above the city of K'ai-feng, the low-water level is between six and 10 feet above the surrounding countryside; the mid-water level is between 19 and 23 feet; and the high-water level sometimes as much as 33 feet above that of the land. From K'ai-feng to the Grand Canal (Yün Ho), the riverbanks rarely exceed three to six feet in height. Marshes are common. Below the Grand Canal the height of the banks increases to 13–16 feet and in some places to 25. Several centuries ago the Chinese built dikes along the river, which in places are broad enough to accommodate villages.

The delta of the Huang Ho begins approximately 160 miles from its mouth, spreading out over an area of about 950 square miles. Built up by silt brought down by the river, the delta is swampy and covered with reeds. A sand bar at the mouth impedes navigation by boats drawing more than four feet of water at low tide; at high tide the depth on the bar is eight or nine feet.

Hydrography. In the past 4,000 years several radical changes have occurred in the Huang Ho's course; at different times the river has entered the Yellow Sea at points varying by as much as 500 miles. From 2278 BC to 602 BC, when it occupied its northernmost course, it flowed through the city of Tientsin and entered the nearby Gulf of Chihli (Po Hai). From 602 BC to AD 70, both the river and its mouth shifted to the south of the Shantung Peninsula. From AD 70 to 1048 the Huang Ho again shifted north, taking up a course much along its present bed.

From 1048 to 1194 changes in the course of the river occurred farther inland, where the river enters the North China Plain. In 1194 the river occupied its southernmost course, the mouth having shifted to the southern edge of the delta. In that year, as a result of the rupturing of the protecting dikes, a second arm of the Huang Ho was created in the southern part of the Shantung Peninsula. From 1289 to 1324 the river took over the bed of the Ko Ho and a large part of the Huai Ho, later returning to its old bed. It was stable for more than 500 years until 1854, when it again began to move farther to the north of the Shantung Peninsula, where it now is located.

Upstream alterations in course

In the upper and middle sections of the Huang Ho, alterations of the riverbed are comparatively slight; concave banks have been washed away, and the alluvial deposits have formed radiating convex banks, or arms, some of which are under cultivation. Numerous outcroppings of rock on the bottom of the river also cause changes in the bed. In the downstream areas, however, where the riverbed is higher than the surrounding land, changes in the bed sometimes lead to a sudden rupturing of the dikes and the flooding of extensive areas. This flooding, which occurs during periods of high water, covers the surrounding territory with huge amounts of silt.

Breaks in the dikes have occurred throughout history. Between 960 and 1048 there were 38 breaks, and 29 from 1048 to 1194. In later years such breaches were less frequent as a result of systematic construction. The slackening of these efforts during the Taiping Rebellion led to a significant change in the course of the river in 1852–54.

In 1887 the Huang Ho burst the dikes near the city of K'ai-feng and began to flow into the Huai Ho, but engineering efforts succeeded in returning it to its former location in 1889. The flood of 1887 covered 30,000 square miles, completely burying many villages under silt. In 1889 another flooding injured nearly 1,000,000 persons and destroyed 1,500 villages. The next major flood, in 1921, wiped out hundreds of populated places, mainly near the mouth of the delta. In the flood of 1933 more than 3,000 populated places were submerged, 3,600,000 people injured, and 18,000 killed. Other floods occurred in 1938 and 1949.

The delta of the Huang Ho is one of the most active in the world. In the century from 1870 to 1970 it pushed outward into the sea an average of more than 12 miles. Some outlying parts have been expanding even more rapidly: one area grew six miles during the period 1949–51, and another grew more than 15 miles in 1949–52.

The Huang Ho carries an average annual volume of about 11.6 cubic miles (48.2 cubic kilometres) of water down to the sea—as much as 16.8 cubic miles (70 cubic kilometres) in high-volume years and as little as 4.8 to 6.0 (20 to 25) in low-volume years. There are also seasonal variations in its volume. More than half of the annual precipitation falls during the rainy season, July to October. The average annual precipitation for the entire basin is about 18½ inches (470 millimetres), but its distribution is very uneven. In some years the bulk of the river volume comes from its tributaries. In the upriver areas the main source is snowfall in the mountains, with the high-water level (33 feet [10 metres]) occurring in the spring. The highest water levels in the middle and lower parts of the river (10 to 23 feet [three to seven metres]) occur in July and August. The maximum flow of water near Lan-chou is 7,000–8,000 cubic yards (5,350–6,115 cubic metres) per second; near Lung-men, 13,000; and in the lower parts of the river, 47,000 (as recorded in 1943).

The Huang Ho is the world's muddiest river. It carries along about 57 pounds of silt per cubic yard (34 kilograms per cubic metre) of water, as compared with two pounds for the Nile, seven for the Amu Darya, and 17 for the Colorado. Floodwaters may contain up to 1,200 pounds of silt per cubic yard (700 kilograms per cubic metre) of water (70 percent by volume). The river carries down to the sea about 1,520,000,000 tons of silt a year, partly because much of the basin is composed of loess, which is loose and easily moved. Other factors are the steepness of the slopes, the rapidity of the current, and a lack of forested areas and reservoirs to check erosion and allow the silt to settle out. The irrigation system in the plain is not enough to slow the river current.

The Huang Ho freezes over in parts of its middle section for several months a year. On the great plain near K'ai-feng there are 15 to 20 icebound days a year but none at all farther downstream. Ice blockage is broken up with the help of bombs dropped from airplanes or sometimes artillery shelling.

Canals and dams

The economy. The 4,000-year-old irrigation system of the Huang Ho, augmented by recently constructed irrigation and navigation canals, brings water to about 10,000 square miles of land. The Grand Canal, cutting across the lower river, runs for more than 1,100 miles from Peking in the north to Hangchow in the south. The People's Victory Canal, completed in 1953, runs for 30 miles parallel to the Peking–Hankow railroad and links the Huang Ho north of Cheng-chou with the Wei Ho at Hsin-hsiang. It raises the level of the Wei Ho, improving navigation between Hsin-hsiang and Tientsin and also providing a

network of irrigation channels between Hsin-hsiang and the Huang Ho.

A 1,000,000-kilowatt power station and dam is located at San-men-hsia (Three-Gate Gorge) on the Honan-Shansi border, 130 miles west of Chengchou; the reservoir is 120 miles long. Before the construction of the reservoir, navigation by boat was restricted to a stretch of 100 miles or so on the lower reaches of the river. Current plans envisage the construction of new dams and reservoirs, the exploitation of its hydroelectric potential (estimated at 30,000,000 kilowatts), and extension of the navigable length of the river and its tributaries.

Study and exploration. Chinese civilization arose in the lower valley of the Huang Ho, which is mentioned in ancient Chinese writings of the 3rd millennium BC. During the reign of one Yao, runs the record, catastrophic flooding occurred on both the Huang Ho and Yangtze River, inundating all of the North China Plain. Regular records have been kept since the 6th century BC of major floods and also of changes in the river's course. Water levels have been studied since 1736. The first European to explore the upper reaches of the Huang Ho was a Russian traveller, Nikolay Przhevalsky, in 1879 and 1884. Since the 1950s the river has been studied intensively by Soviet scientists as well as by the Chinese. (I.V.P.)

SEA OF JAPAN

The Sea of Japan is a marginal sea of the western Pacific Ocean, bounded by Japan and the Soviet island of Sakhalin to the east and by the Soviet Union and Korea (where it is known as the East Sea) on the Asian mainland to the west. Its area is 389,100 square miles (1,007,800 square kilometres). It has a mean depth of 4,429 feet (1,350 metres) and a maximum depth of 12,276 feet (3,742 metres).

Physical features. *Physiography.* The sea is almost elliptical, having its major axis from southwest to northeast; to the north it is approximately bounded by latitude 51°45′ N, while to the south it is bounded by a line drawn from the Japanese island of Kyūshū westward through the Gotō-rettō (Gotō Islands) of Japan to the Korean island of Cheju (also known as Quelpart Island) and then northward to the Korean peninsula.

The sea itself lies in a deep basin, separated from the East China Sea to the south by the Tsushima-kaikyō (Tsushima Strait) and Korea Strait and from the Sea of Okhotsk to the north by Sōya-kaikyō (Sōya Strait, or La Pérouse Strait) and Tatar Strait, all of which have sill depths of less than about 700 feet. To the east it is also connected with the Inland Sea of Japan by Kanmon-kaikyō.

Component basins

Underwater, the sea is separated into the Japan Basin (about 9,800 to 11,500 feet deep) to the north, the Yamato Basin (8,200 feet deep) to the southeast, and the Tsushima Basin (6,600 feet deep) to the southwest. While a narrow continental shelf with an average width of about 19 miles (30 kilometres) fringes Siberia and the Korean peninsula, on the Japanese side of the sea there are wider continental shelves with depths at the edge of between about 430 and 1,300 feet, as well as groups of banks, troughs, and basins lying offshore. The banks lying off the coasts of Japan are divided into groups, which include Okujiri Ridge, Sado Ridge, Hakusan Banks, Wakasa Ridge, and Oki Ridge.

Geology. Yamato Ridge, which has an average depth of about 1,475 feet (450 metres), consists of granite, rhyolite, andesite, and basalt, with boulders of volcanic rock scattered on the seabed. The top of the Korea Plateau is about 3,000 feet below the surface and has a depression in its central part. Geophysical investigation has revealed that, while Yamato Ridge is of continental origin, the Japan Basin and the Yamato Basin are of oceanic origin.

Bottom deposits in the Sea of Japan indicate that earthborn sediments, such as mud, sand, gravel, and fragments of rock, exist down to depths of 650 to 1,000 feet; hemipelagic sediments (*i.e.,* half of oceanic origin), mainly consisting of blue mud rich in organic matter, are found down to depths of about 1,000 to 2,600 feet; and deeper pelagic sediments, consisting of red mud, are found down to a depth of nearly 13,000 feet.

A number of submarine canyons are found on the continental slope, at depths of more than 6,500 feet on the west side of the basin, while those near the islands of Japan lie at depths of only about 2,600 feet.

Argument over the formation of the Sea of Japan has not yet ended, although there is agreement that the four straits that connect the sea either to the Pacific Ocean or to marginal seas were formed in very recent geological periods. The oldest of these straits are the Tsugaru-kaikyō and Tsushima-kaikyō, whose formation interrupted the migration of elephants into the Japanese islands at the end of the Tertiary Period (about 2,500,000 years ago); the most recent is Sōya-kaikyō, which was formed at the end of the Wisconsin Ice Age (60,000 to 11,000 BP) and which closed the route once used by the mammoths whose fossils have been found in Hokkaidō.

Climate. The Sea of Japan contributes greatly to the mild climate of Japan because of the effect exerted by its relatively warm waters; evaporation is especially noticeable in winter, when an estimated 5,000,000,000 tons of water vapour rise as steam fog near the Polar Front (*i.e.,* the frontier between the cold, dry polar air mass and the warm, moist tropical air mass). From December to March the prevailing northwest monsoon wind carries cold and dry continental polar air masses over the warmer waters of the sea, resulting in persistent precipitation in the form of snow along the mountainous coasts of Japan. In summer the southerly tropical monsoon blows from an area of higher atmospheric pressure over the North Pacific onto the Asian mainland, causing dense fog when its warm and moist winds blow over the cold currents that prevail over the northern part of the sea at that season. The winter monsoon brings rough seas and causes coastal erosion as a result of the heavy surf that breaks along the western coasts of Japan. In summer and fall typhoons occasionally occur.

The northern part of the sea, especially off the Siberian coast as well as in Tatar Strait, freezes in winter; as a result of convection, melted ice feeds the cold currents in that part of the sea in spring and summer.

Hydrography. The waters of the sea generally circulate in a counterclockwise pattern. A striking contrast occurs between the cooler and relatively fresher water in the western part and the warmer and relatively more saline water in the eastern part. A branch of the Kuroshio Current, the Tsushima Current, together with its northern branch, the East Korea Warm Current, flows north, bringing warmer and more saline water, before flowing into the Pacific through the Tsugaru-kaikyō as the Tsugaru Current, as well as into the Sea of Okhotsk through the Sōya-kaikyō as the Sōya Current. Along the coast of the Asian mainland, on the other hand, three cold currents—the Liman, North Korea, and Central (or Mid-) Japan Sea cold currents—bring cooler, relatively fresh, and turbid water southward.

Prevailing currents

The economy. Fisheries and mineral deposits form the main economic resources. Fisheries may be divided into pelagic (oceanic) and demersal (sea-bottom) categories. Pelagic fishes include mackerel, horse mackerel, sardines, anchovies, herring, fishes of the salmon and trout family, sea bream, and squid; the demersal category includes cod, Alaskan pollack (bluefish), and Atka mackerel. Seals and whales are also to be found, as well as such crustaceans as prawns and crabs. The fishing grounds are for the most part on the continental shelves and their adjacent waters, as well as in the Polar Front zone and on the submarine banks.

Herring, sardines, and bluefin tuna have traditionally been caught, but since 1946 the fisheries have been becoming depleted. Squid fishing is carried on in the central part of the sea, salmon fishing in the shoal areas of the north and southwest, and crustacean trapping in the deeper parts.

Mineral resources on or in the sea bottom include magnetite sands as well as natural gas and petroleum deposits off Japan and Sakhalin Island.

As trade increases between Asian countries, the Sea of Japan is being increasingly used as a commercial waterway. While the waters of the sea historically have served to protect Japan from foreign invasions, the southern straits have been the scene of some historical naval battles. On two occasions—in 1274 and 1281—the fleets of Kublai

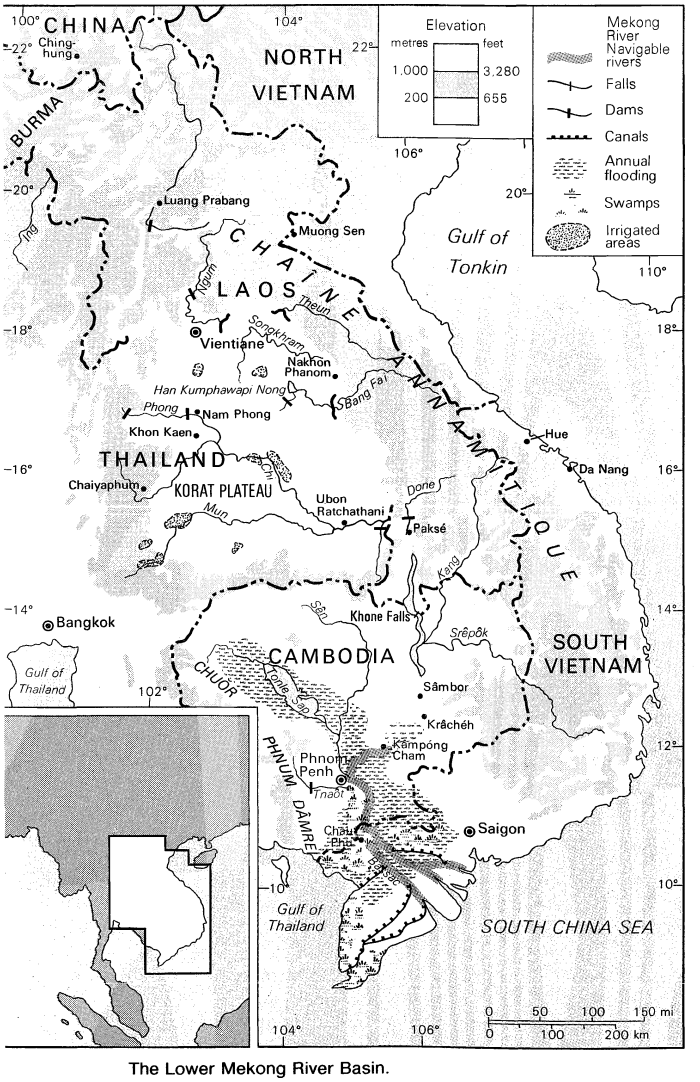
Khan, founder of the Yüan (Mongol) dynasty in China, attempted to cross the straits to conquer Japan, but the conqueror's forces were either destroyed by typhoons or beaten by the Japanese. In 1905, at the battle of Tsushima Strait, the Japanese navy almost completely destroyed the Baltic fleet of the Russian tsar. (M.U.)

MEKONG RIVER

The Mekong, 2,500 miles (4,000 kilometres) long, is the longest river in Southeast Asia and the seventh longest in all of Asia. Rising in Tsinghai Province, China, it flows through Yunnan Province, after which it forms part of the international border between Burma and Laos, as well as between Laos and Thailand, also flowing through Laos, Kampuchea, and Vietnam before draining into the South China Sea to the south of Ho Chi Minh City (Saigon). Vientiane, capital of Laos, and Phnom Penh, capital of Kampuchea, both stand on its banks. About 77 percent of the drainage area of the Mekong lies within the four countries traversed by its lower basin—Laos, Thailand, Kampuchea, and Vietnam.

Physical features. The Mekong River drains more than 307,000 square miles (795,000 square kilometres) of land, stretching from the Tibetan Plateau to the South China Sea. Among Asian rivers, only the Yangtze, Ganges, and Irrawaddy have larger minimum flows. The Mekong River's small width in the first 1,150 miles of its course, and the contrast between the physical conditions that prevail above and below the reach where it flows down off the Yunnan highlands, divide it into two major parts.

The Upper Mekong is a long, narrow valley comprising roughly 26 percent of the total area, cutting through the mountains and plateaus of China. The Lower Mekong, below the point where it forms the border between Burma and Laos, is a stream 1,454 miles in length that claims the drainage from the Korat Plateau of Thailand, from most of Kampuchea, and from the westward slopes of the Chaîne Annamitique (Annamite mountain chain) in Laos and Vietnam before reaching the sea through the distributary channels of its delta.



The Lower Mekong River Basin.

In its upper reaches the Mekong is one of the cluster of great streams rising in the plateau between the Salween and the Yangtze; the streambed has cut deeply into the rugged landscape through which it flows. Where it flows between Burma and Laos, it drains about 8,000 square miles of Burmese territory, all of which consists of rough and relatively inaccessible terrain. In its more gentle lower stretches, where for a considerable distance it constitutes the boundary between Laos and Thailand, it forms a subject of both friction and cooperation among the four countries of Kampuchea, Laos, Thailand, and Vietnam.

Origin of the river

Physiography. The upper sources, known locally as the Pam and the Dzi Chu, rise at elevations of more than 16,000 feet (4,875 metres) in the Tibetan highlands on the southern border of Tsinghai. They flow southeasterly through the Chamdo region of Tibet. The main stream, called the Lan Ts'ang Chiang, descends in a southerly direction across the highlands of Yunnan, which are cut by erosion into hills and valleys, to a point south of Yiinching Hung, where it becomes the border between Burma and China. The river then moves in a southwesterly direction; over a reach of more than 125 miles it forms the Burmese–Laotian border. Although two great roads cross it—the caravan route from the southeast to Lhasa, and the K'un-ming-to-Burma road—much of the river valley in the high plateau and in the Yunnan Mountains is extremely inaccessible and sparsely populated.

Below Burma, the river basin may be divided into six major sections—the northern mountains, Korat Plateau, eastern highlands, southern lowlands, southern highlands, and delta sections. All these sections have somewhat similar landforms, vegetation, and soils. Most of the vegetation in the lower basin is of the tropical broadleaf variety, although the occurrence of individual species varies with latitude and topography.

The northern mountains section has highly folded mountains reaching elevations of about 9,000 feet above sea level, many with slopes slanting at steep angles. As far south as the latitude of Vientiane, these dissected uplands (*i.e.*, cut by erosion into hills and valleys) are covered with dense deciduous forest that has deteriorated as a result of inroads made by shifting cultivation. To the south of the east–west course of the river above Vientiane lies the Korat Plateau, which embraces almost all of the Thai portion of the basin as well as the lower parts of the Mekong's Laotian tributaries. This is an area of flat or gently rolling plains traversed by relatively flat valley bottoms. Soils and deciduous vegetation on the uplands are thin, and much of the original forest has been replaced by grassland as a result of grazing and repeated burning.

The Korat Plateau

The eastern highlands form part of the Chaîne Annamitique of mountains, from which streams drain west into the Mekong. Throughout most of the distance between Muong Sen in Vietnam and Krâchéh in Kampuchea the watershed forms the border between Vietnam to the east and Laos and Kampuchea to the west. There is greater relief in the northern than the southern parts of the watershed, but the highlands in general are characterized by rapid streams that flow through narrow valleys before entering the Korat Plateau or other lowlands. Forest degradation, that has resulted from lumbering, the temporary use of land for cultivation, and grazing, is widespread.

The southern lowlands border both sides of the Mekong below Paksé in Laos. The part of the river that flows through Kampuchea has wide stretches of alluvium in its floodplain. Near Phnom Penh, the Kampuchean capital, a junction occurs between the Mekong and the Tonle Sap river, which connects it to the Tonle Sap lake, sometimes called the Great Lake. The direction of flow of the Tonle Sap river varies according to the season. In midsummer, when the Mekong is in flood, waters flow down the Tonle Sap to the lake, which at this time increases its area from

The Great Lake of Tonle Sap

about 1,000 square miles to about 3,000 square miles. In midwinter when the floods subside, the Tonle Sap reverses its flow to drain southeastward into the Mekong. Tonle Sap lake is one of the most productive fishing grounds in the world, producing as much as 26 tons of fish a year per square mile.

Along the southern border of the basin in Kampuchea, the Éléphant and Cardamon mountains form a string of southern highlands that drain northward into the southern lowlands.

The river divides into two streams—the Mekong and the Bassac—below Phnom Penh. From this point onward the delta spreads out to the sea. The delta, which has a total area of about 25,000 square miles (65,000 square kilometres), has three major sections. The upper section, above Chau Phu, has strong natural levees (embankments built on either side of the river by accumulated deposits of silt) behind which are low, wide depressions. The middle section has some areas that are well drained, others that are poorly drained and swampy. Along the lower section, formed by the river mouths and by the area to the southwest, sediment carried down from the upper river is in the process of being deposited, and the flooding is less extreme than in the upper sections of the delta.

Hydrology. The mean annual flow of the river at Krâchéh in Kampuchea is about 500,000 cubic feet per second (cusecs), making a total annual discharge for the year of about 378,000,000 acre-feet. The recorded minimum at Krâchéh is about one-twelfth of the mean, and the annual peak flow about four times the mean. Below Krâchéh the peak flows diminish as the water spreads out into the distributary channels and backswamps. Flow comes chiefly from rainfall in the lower basin and reflects the variation in seasonal rainfall caused by the monsoon winds; this variation generally forms a regular annual pattern: in April the flow is ordinarily at its lowest; in May or June the flow begins to increase, doing so most rapidly in the eastern highlands and northern mountains; the highest water levels are reached as early as August or September in the upper reaches and as late as October in the southern reaches. The northeast monsoon wind, beginning ordinarily in November in the southern areas, brings dry weather until May.

The annual sediment load is recorded as being highest at Paksé, where it amounts to 132,000,000 tons; it is about half that amount at the Burmese border and about two-thirds at Phnom Penh. The dominant hydrologic fact affecting agriculture is the long dry period in which rice cultivation is impossible without irrigation.

The people. The inhabitants of the lower basin amount to half the population of the four riparian countries. About 80 percent are engaged in agriculture, and rice is the major crop. Heaviest population concentrations are in the delta and on the Korat Plateau. The small urban population has been growing rapidly, chiefly through migration to the capital cities. There is no common ethnic tie among the basin populations. Ethnic groups range from Sino-Tibetan, including Karen and Meo mountain groups, in the Upper Mekong region, to Khmer, Cham, Tai, Mon, and Vietnamese lowland groups in the Lower Mekong Basin. The Vietnamese are heavily concentrated in the delta, and the Khmer and Tai are the most widely distributed in the lower basin.

Ethnic groups

The economy. *Irrigation and flood control.* In the lower basin the management of water offers major opportunities to increase the economic productivity of the tributary lands. Farmers practicing shifting cultivation on the uplands and the rice growers on the rain-fed lowlands are able, under normal conditions, to grow only one crop a year, taking advantage of wet-season precipitation. Half the cultivated land is dependent upon some form of inundation by flood flows. Control of water, however, makes it possible to store water during the dry season, and thus permits the harvesting of a second or third crop. If irrigation is combined with flood control, the losses and delays caused by floods pouring over the river's banks are reduced. Where storage facilities and the degree of downward slope are favourable, hydroelectric power can be generated. If navigational conditions on the Mekong's main stream and on some of its tributaries were to be improved, transport costs for landlocked areas could be reduced.

After the Bureau of Flood Control of the United Nations Economic Commission for Asia and the Far East (ECAFE) had recommended study of the Lower Mekong, surveys were carried out by ECAFE and by the U.S. Bureau of Reclamation. This led to the adoption by Cambodia, Laos, Thailand, and South Vietnam of a 1957 statute creating the Committee for Coordination of Investigations of the Lower Mekong; a 1978 agreement extended the work of the Committee but no longer included Kampuchea, as there was no competent authority to represent it. By the mid-1980s the three riparian countries had sponsored a series of preinvestment and general scientific investigations and had undertaken construction of multiple-purpose water projects. They had continued to cooperate despite the political stresses produced by the war in Vietnam and its aftermath and had enlisted the assistance of other countries.

Investigations undertaken included basic mapping, hydrologic observations, soil surveys, fisheries studies, health studies, engineering-feasibility studies, power-market surveys, agricultural research and pilot farms, and many other inquiries. The engineering studies had provided for reconnaissance appraisal of all of the tributary basins and for more detailed examination of selected projects. Flood forecasting was begun and river navigation was improved. Mineral surveyors explored possible opportunities for bauxite and iron development. Some irrigation and power projects have been initiated or completed.

The focus of the Mekong Development Project as a whole shifted after the Vietnam War to the planning of comprehensive programs for agricultural and community development in areas where water supply was available, with each country working out its individual financial arrangements. No formal negotiations with the two upstream riparian countries, Burma and the People's Republic of China, or with adjacent countries that might be markets for power, had as yet taken place.

Navigation. In the Vietnamese part of the delta there is an elaborate system of canals. Smaller seagoing vessels can sail upstream as far as Phnom Penh and vessls drawing almost 15 feet can reach Kompong Cham during high water. Continuous water transport is blocked chiefly by the barriers of the Khone Falls and other falls between Sâmbor and Paksé, and upstream uses of the river are limited to local traffic. (G.F.W.)

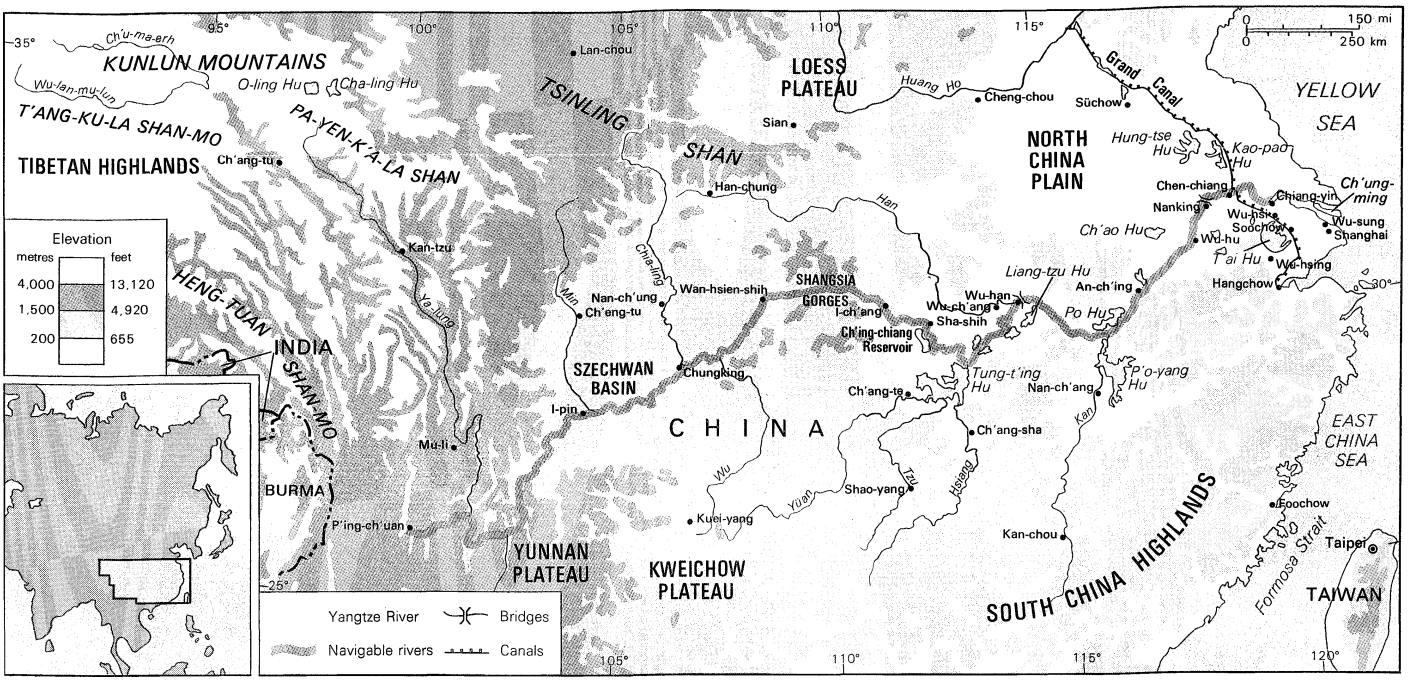
YANGTZE RIVER

The Yangtze River (in Chinese Ch'ang Chiang, in Pinyin Chang Jiang, also spelled Yangtze Kiang, English Blue River) is the longest river in both China and Asia, and is the fourth longest river in the world. Its length is 3,434 miles (5,525 kilometres); its basin extends for 2,000 miles from west to east and is over 600 miles wide. Rising from its source in the west of China, the river, either completely or partially, traverses 12 provinces or regions, including the autonomous region of Tibet. Population distribution in the Yangtze Basin is uneven; it is greatest in the plains that adjoin the banks of the river and its tributaries in central and eastern China and is most sparse in the highland areas to the west of the basin. More than three-quarters of the river's course runs through mountains.

The name Yangtze, which came to be applied to the entire river, is derived from the name of the ancient fiefdom of Yang; it came to be applied by Europeans to the whole river. Among the Chinese people this name is not commonly used; instead the generally accepted name in China for the river is Ch'ang Chiang (Long River); to a lesser extent it is also called Da Kiang (Great River). Yet another name, also given by Europeans, the Blue River, is a misnomer; in its middle and lower course the water of the Yangtze has a brown-yellow hue.

The principal tributaries

The Yangtze has eight principal tributaries. On its left bank, from source to mouth, these are: the Ya-lung Chiang (about 730 miles [1,170 kilometres] long), the Min Chiang (490 miles), the Chia-ling Chiang (700 miles [1,120 kilometres]), and the Han Shui (950 miles); while on the right



The Yangtze River Basin.

bank are the Wu Chiang (630 miles [1,010 kilometres]), the Yüan Chiang (600 miles), the Hsiang Chiang (500 miles), and the Kan Chiang (470 miles).

The Yangtze Basin is the granary of China and contributes almost half of the crops of the country, including up to 70 percent of the total gross volume of rice. In addition, many other crops are grown, among them cotton, wheat, barley, corn (maize), beans, and hemp. Such large cities as Shanghai, Nanking, Wu-han, Chung-king, and Ch'eng-tu are located in the basin, all of which have populations of more than 1,000,000.

Physical features. *The upper course.* The upper course of the Yangtze flows through the Tibetan highlands, where the population consists mainly of persons engaged in cattle breeding and in primitive agriculture. The summers are warm and the winters cold, and the growing season lasts for four or five months. In the most heavily populated points live a small number of Chinese, Dungan, Nepalese, and Indians. In the mountainous regions adjoining the highlands of the southwest, the population is predominately Chinese; like the other minorities, they are primarily engaged in agriculture.

Sources of the Yangtze

The Yangtze initially rises from two sources, which are located in the T'ang-ku-la Shan-mo (T'ang-ku-la Mountains) at an altitude of 18,000 feet (5,500 metres) above sea level. The main (southern) source is known locally by its Tibetan name, Ulan Muren (Wu-lan-mu-lun in Chinese). From the confluence of its source streams, the river flows through a shallow, spacious valley, the bottom of which is studded with lakes and small reservoirs. This part of its course lies in the higher regions of the Tibetan highlands. The river's character changes sharply upon reaching the eastern limits of the highlands. Here, in exiting from Tibet, the river descends from a great altitude, winding its way south of the high Pa-yen-k'a-la Shan and forming a narrow valley between one and two miles in depth. Individual mountain peaks exceed 16,000 feet above sea level and are crowned with glaciers and perpetual snow. The steep, rocky slopes are cut with gorges and deep valleys. For several hundred miles the Yangtze flows in southeasterly direction, before turning southward to flow downward in rushing rapids. For a considerable distance the river flows through passes that are so steep that no room is left even for a narrow path. Villages, which are rarely found, are located high above the river. In this region the Yangtze runs close and parallel to both the Mekong and Salween rivers; all three rivers are within 15 to 30 miles of one another and continue to flow in mutual proximity for a distance of more than 250 miles.

North of latitude 26° N the paths of the rivers diverge, with the Yangtze turning to the east. The Yangtze then flows through mountains to the city of I-pin, passing through a winding valley with steep slopes. Admitting the waters of many tributaries, among which the Ya-lung Chiang is the largest and contributes the most water, the Yangtze widens to between 1,000 and 1,300 feet, reaching depths not infrequently exceeding 30 feet. In the narrow gorges the water width decreases by almost half, but the depth increases sharply.

The Ya-lung Chiang

Toward the end of the upstream part of its course, the Yangtze descends to a level of 1,000 feet above sea level. Thus, the total fall upstream for about 1,600 miles is more than 17,000 feet, or an average of over 10 feet per mile of its course. In the mountains, however, there is a substantial stretch where the fall of the river is several score feet for each mile of its path.

The middle course. The middle course of the Yangtze is characterized by hot summers and not very cold winters, while the precipitation measures from 40 to 60 inches (1,016 to 1,524 millimetres) a year; the growing season lasts for more than half a year. A large part of the rainfall occurs in summer; this is favourable to the development of agriculture, which in this region dominates the economy. The main population is Chinese, with many other national minorities represented.

The river's course between I-pin and I-ch'ang

In its middle course, which stretches for about 630 miles between the cities of I-pin and I-ch'ang, the Yangtze crosses hilly Szechwan Province, where the mountainous region connects the highlands in the southwestern part of China with the mountainous region of Tsinling east of the Himalayas and south of Sian. The river's width here is from about 1,000 to 1,600 feet, and the depth in places exceeds 30 feet. The current is fast; the banks are often high and steep. The fall of the river in Szechwan province is more than 820 feet (about one foot per mile of flow). The people of Szechwan call this area the "Land of Plenty." The soil here is very fertile, and the climatic conditions are entirely favourable to agriculture, since the high mountains associated with the steppe (treeless plain) safely protect the province from the cold north and west winds. The mild climate also favours the development of sericulture (the production of raw silk by raising silkworms). Deposits of various useful minerals are concentrated in this section of Szechwan Province: coal, copper, phosphorus, gold, oil, and gas. The population is especially dense on the plain in the vicinity of the city of Ch'eng-tu. Also located in this area is Chungking, a large-scale industrial centre and river port.

Leaving Szechwan province behind, the Yangtze flows for a distance of 125 miles through a mountainous region where the three narrow gorges of Tsyuytan, U, and Silin are located, before exiting onto the plain. The gorges have steep, sheer slopes composed mainly of thick limestone rocks, rising from 1,300 to 2,000 feet above the river and presenting the appearance of fantastic towers, pillars, or spears. The first gorge—about five miles long—is the shortest, but it is considered the most dangerous for navigation; the river here is very narrow and contains many rapids and eddies. The following gorge, which stretches for about 30 miles, is a narrow, steep corridor with almost vertical walls of heights up to 1,600 or even 2,000 feet. The last gorge is located upstream of I-ch'ang and extends for a distance of 21 miles; in places limestone cliffs rise directly out of the water to heights of hundreds of feet. The gorges are rocky, and the walls speckled with cracks, niches, and indentations. The width of the Yangtze between the gorges is 1,600 to 2,000 feet and about 500 to 600 feet within the gorges; the depth in gorges, however, also increases to between 500 and 600 feet, thus making the Yangtze the deepest river in the world.

The lower course. The lower part of the Yangtze River Basin, situated at the edge of an extensive plain in the lowlands of eastern China, experiences a temperate, coastal climate featuring a warm spring, a hot summer, a cool fall, and a comparatively cold winter. Monsoons (rain-bearing winds) dominate the weather of the region, and in the summer and fall typhoons are experienced periodically. The natural conditions here are exceptionally favourable to agriculture; the growing period continues for eight to 11 months, and in some areas two or three successive crops can be harvested in a single year.

This region is the most industrially and agriculturally developed area in China. A majority of the population of the basin, mostly Chinese, are concentrated here, and the population density is very high. About 1,300 miles downstream, the Yangtze begins its exit from the Shangsia Gorges, flowing into an extensive plain. Here the river forms a complex system consisting of the main river channel, many tributaries, and lakes. The Yangtze flows to the city Wu-han across the flat Liang-tzu Hu Plain, contributing alluvial and lake deposits to the bed of the large Yün-meng Tse (Yun-men Lakes).

The lake region

On the wide terrace-like slopes are situated a number of lakes that begin as small and end as large reservoirs, such as Ch'ang Hu, Hung Hu, and Tung-t'ing Hu. The total area of the lakes exceeds 6,600 square miles (17,100 square kilometres). The lakes are of national economic significance, mainly as fisheries. At the edge of the Liang-tzu Hu Plain the Yangtze widens markedly, the course of its stream wandering in the form of a large loop. The width of the river is up to 2,600 feet, the depth is over 100 feet, and the water current flows at a rate of three to four feet a second. The banks are built up for protection from floods. In the southern part of the plain lies Tung-t'ing Hu, which is the largest lake in the region, having an area of from 1,400 to 1,800 square miles; it shares four tributaries and two canals with the Yangtze, whose flow it serves to regulate. The surrounding area, agricultural and studded with lakes, is called the rice bowl of China.

On the Yangtze, near the mouth of the Han Shui (Han River), are located three cities—Han-yang and Hankow on the left bank, and Wu-ch'ang on the right bank. In the course of time these cities have come to form one huge city—Wu-han, a large metallurgical base and river port. Further east on the Liang-tzu Hu Plain, the Yangtze flows into a narrowing, picturesque valley and then exits onto the plain, in the southern part of which is located another large lake—P'o-yang Hu, which has an area of about 1,000 square miles and which is linked to the Yangtze by a wide tributary. The river then turns to the northeast, flows through a wide, closed valley into which many lakes empty, and exits into the North China Plain. The width of the river increases at this point to between 3,000 and 6,000 feet, and the depth in places approaches 100 feet. In this region there are a number of large cities, An-ch'ing, Wu-hsi, Soochow, and Shanghai among them. The Grand Canal, which, with a length of 1,300 miles, is one of the longest canals in the world, leaves the Yangtze in the vicinity of the city Chen-chiang.

The Yangtze Delta. The Yangtze Delta, which begins beyond the city of Chen-chiang, consists of a large number of branches, tributaries, lakes, ancient riverbeds, and marshes that are connected with the main channel. During major floods the delta area is completely submerged. T'ai Hu, with an area of about 1,300 square miles, is notable as the largest of the many lakes in the delta. The width of the Yangtze in the delta, as far as the city of Chiang-yin, ranges from less than a mile to almost two miles; further downstream the channel gradually widens and becomes a large estuary, the width of which exceeds 50 miles near the mouth of the river.

T'ai Hu

Before emptying into the sea, the Yangtze divides into two arms that drain independently into the East China Sea. The left branch has a width of from three to six miles, the right branch of from six to 15 miles. Between the branches is situated Ch'ung-ming Tao (Ch'ung-ming Island), which has an area of 300 square miles and which was formed over 1,000 years ago as a result of the deposit of alluvium at the mouth of the Yangtze. The depth of the river in places approaches 100 to 130 feet but decreases to several feet near the sea at the mouth of the river due to the presence of sandbars.

The section of river from the mouth to 250 miles upstream is subject to the influence of tides. The maximum range of the tides near the mouth is 13 to 15 feet.

The present bed of the Yangtze in this area is several yards above the plain. Thus, for protection of the surrounding region from floodwaters, the banks of the main and other rivers are built up; the total length of banks on the Yangtze that have thus been diked is about 1,700 miles. Dams have also been built for flood protection on the shores of several lakes; the Ch'ing-chiang Reservoir, for example, built for this purpose near Tung-t'ing Hu, has a capacity of 194,000,000 cubic feet (5,500,000 cubic metres). The delta is protected from the sea by two gigantic parallel banks that are faced with stone in most parts. The extensive area of the delta is used for the planting of rice and cotton.

Hydrology. The Yangtze Basin is comparatively well irrigated; the average yearly rainfall amounts to about 43 inches (1,100 millimetres). Most of the rain is brought by the monsoon winds and falls primarily in the summer months. In the mountainous part of the basin most of the precipitation takes the form of snow. Floods, which result from the monsoon rains in the middle and lower parts of the basin, usually begin in March or April and last for about eight months. In May the water level decreases somewhat, but then sharply increases again, continuing to rise until August, when it reaches its highest level. After that the water level gradually falls to the premonsoon levels, the decrease continuing through the fall and most of the winter until February, when the lowest annual level is reached.

Spring floods

The annual range of water level fluctuations is very great—an average of about 65 feet—with 26 to 35 feet during years of low water. In the gorges, the range of water levels caused by flooding reaches huge proportions—from 130 to 150 feet. Downstream, the impact of the water level variation is lessened by the regulating effect of the lakes; here tides exert the greatest influence on the water level. Near the city of Wu-sung the daily tidal range is 15 feet, and the yearly range is 20 feet.

A breakdown of the water volume delivered to the mouth of the Yangtze shows that the highland part of the basin contributes 10 percent of the flow, while all of the rest of the water in the river is contributed by the middle and downstream parts of the basin, with the Tung-t'ing Hu and P'o-yang Hu being responsible for about 40 percent of the volume.

The Yangtze carries a tremendous volume of water. Even in the upstream areas the average volume of flow exceeds 70,000 cubic feet a second. For comparison, it may be noted that the second longest river in China, the Huang Ho, has a volume of flow at its mouth of 52,580 cubic feet per second, which is significantly less than that of the Yangtze. After the inflow from the first large tributary—

the Ya-lung Chiang—the volume in the Yangtze sharply increases, approaching an average of 194,000 cubic feet a second. Further downstream the Yangtze admits many tributaries, and the volume gradually increases, reaching 529,000 cubic feet per second at the end of the Shangsia Gorges near I-ch'ang, 847,000 cubic feet per second near Hankow, and 1,100,000 cubic feet per second near Nanking. The average volume at the mouth of the Yangtze equals 1,200,000 cubic feet per second, while the total volume entering the sea annually is 244 cubic miles (1,072 cubic kilometres), ranking it fourth in volume of flow among rivers of the world.

During the seasonal rains the Yangtze widely floods the lower areas, and the maximum volume of water entering the sea can amount to from 1,800,000 to 2,500,000 cubic feet per second. During the dry season the flow decreases, sometimes to 212,000 and 282,000 cubic feet per second. In spite of the fact that the current volume of the Yangtze exceeds that of the Huang Ho, the Yangtze is significantly less muddy than the Huang Ho. This is explained by the fact that in the Yangtze Basin the soil is secured by a greater amount of vegetation. In the mountainous part of the basin, particularly in the Tibetan highlands, the waters of the Yangtze are transparent and contain little silt.

The volume of silt

Downstream, however, the waters become muddy and acquire a coffee colour. It is estimated that the Yangtze carries between 280,000,000 and 300,000,000 tons of alluvium to its mouth, depositing an estimated 150,000,000 to 200,000,000 tons on the river bottom in addition. Thus, the total amount of particle substance carried or deposited is between 430,000,000 and 500,000,000 tons in a year. As a result of the depositing of alluvium at the river's mouth, the delta extends into the sea an average of one mile in 64 years.

During the period of monsoon rains the Yangtze and its tributaries spill over, creating extensive floods. If the floods in the main channel coincide with flooding in one or more of the major tributaries, powerful, destructive flood waves can result, an occurrence that has happened repeatedly in the history of China.

The flooding frequently results from the deposit of silt in the bed of the Yangtze. Upon leaving the mountains and entering the plain, the current in the Yangtze sharply decreases, and thus the flow cannot continue to carry the entire amount of silt. As a result, a significant portion is deposited in the bed, causing the bottom to rise. An analogous situation occurs in many of the Yangtze's tributaries. Flooding thus presents a great danger to the inhabitants of the adjacent plains.

A legendary flood

In ancient literature there is much information about a legendary flood that is thought to have occurred in 2297 BC. The flood occurred as a result of extremely intensive and extended heavy rains. The Huang, Wei, and Yangtze rivers overflowed their banks and submerged almost the entire North China Plain, turning it into a huge sea; in the lowest places the water remained for many years. Since that catastrophe many other major floods have occurred. Historical records indicate that, during the period from 206 BC to AD 1960, China has experienced more than 1,030 major floods. Especially extensive flooding has occurred on the Yangtze more than 50 times and on its tributary, the Han Shui, 32 times; on the average, the Yangtze Basin is the scene of catastrophic flooding every 50 to 55 years. Extensive flooding may also take place in shorter periods of time. This has been the case in the last 100 years, during which time the Yangtze Basin was flooded in 1870, 1896, 1931, 1949, and 1954. Of these, the 1931 and 1954 floods have been general, national disasters. The 1931 flood resulted from heavy, continuous monsoon rains that covered most of the middle and lower parts of the basin. During May and June, six huge flood waves swept down the river, destroying the protecting dams and dikes in 23 places and flooding more than 35,000 square miles of land; 40,000,000 persons were rendered homeless or otherwise suffered. Many population centres, including Nanking, Wu-han, and others, were underwater. In Wu-han, the water remained for more than four months, the depth exceeding six feet, and in places ranging up to more than 20 feet. In the summer of 1954 a new, powerful flood occurred, resulting from continued monsoon rains. The water level sharply increased and at times exceeded the 1931 flood levels by almost five feet. As a result of effective flood-control measures, however, many of the potential consequences of the flood were averted.

The economy. *Navigation.* The Yangtze is the principal river used for transportation in China. Along the river for 1,700 miles there is intensive cargo and passenger traffic. The river serves as a continuation of the sea routes, binding the inland and sea ports together with other major cities into a transportation network in which Nanking, Wu-han, and Chungking play the leading roles. Junks are widely used for transporting cargo, and a small number of sail craft with displacements of 50 to 100 tons are also used. Large ships of up to 10,000-tons displacement can travel as far upriver as Wu-han, which is 700 miles (1,120 kilometres) from the coast, and craft of up to 2,000 tons can reach I-ch'ang, but only smaller craft can reach P'ing-ch'uan. Water routes in the Yangtze Basin total about 35,000 miles (56,000 kilometres). The Yangtze is joined to navigable branches of the Huang Ho and Wei Ho by the Grand Canal, which is further connected with the seaports of Hangchow and Tientsin.

Rivercraft

Two large railroad bridges have been built across the Yangtze at Nanking and Wu-han. Another bridge crosses the river at Chungking.

Resources. The Yangtze River and its associated tributaries and lakes, including the large Tung-t'ing Hu, P'o-yang Hu, and T'ai Hu, abound with fish. The fishing trade is widely developed and is a major livelihood for much of the population of the region. Up to 500 kinds of fish can be found in Chinese rivers, the majority of which inhabit the Yangtze and its tributaries. About 30 kinds of riverine fauna have economic significance, especially carp, bream, Chinese perch, gapers (a species of large burrowing clam), and lamprey; the most valuable economically are white and black amur (large members of the carp family), flatfish, and spotted flatfish. Sturgeon are also important; the gorges are a favourite spawning area. Further downstream great amounts of roe can be found, and these are collected and distributed throughout the country for artificial cultivation. Several billion roe are collected annually. The artificial cultivation of fish for trade involves mainly white and black amur, flatfish, and carp.

In the Yangtze Basin the extensive territory under cultivation—especially for rice, the most important crop—requires man-made irrigation facilities. Even in the areas of highest precipitation severe droughts are experienced, resulting in the loss of crops. This is explained by the extremely irregular distribution of precipitation over the course of the year, with 60 to 80 percent falling in the summertime. Rainless periods sometimes last for six to eight weeks. Irrigation has existed in the Yangtze Basin since ancient times, but many modern irrigation projects have been undertaken.

Hydroelectric power. The resources for the production of energy from the Yangtze are extremely great, although they have not been developed to a large extent. The total potential power is estimated to be 217,200,000 kilowatts, representing 40 percent of the total energy potential of all of the rivers of China. The power resources in the vicinity of the Tsyuytan, U, and Silin gorges are especially great, amounting to about 40,000,000 kilowatts. Many tributaries of the Yangtze that have significant fall and volume, such as the Ya-lung Chiang, Min Chiang, Chialing Chiang, and other rivers that are tributaries of Tung-t'ing and P'o-yang lakes, also have a potential for contributing large amounts of hydroelectric energy. It is anticipated that the prevention of destructive floods, the development of new agricultural territory, and the improvement of conditions for agriculture can be achieved simultaneously by the development of the hydroelectric-energy potential of the Yangtze Basin. (A.P.M.)

YELLOW SEA

The Yellow Sea (in Chinese Huang Hai) is a large inlet of the western Pacific lying between the People's Republic of China and the Korean peninsula. It is situated to the north of the East China Sea, which it bounds on a line running

from the mouth of the Yangtze River, near Shanghai, to Cheju-do (Cheju Island) off South Korea. It measures about 600 miles (960 kilometres) from north to south and about 435 miles from east to west. In the northwest part of the sea, northwest of a line between the Liaotung Peninsula to the north and the Shantung Peninsula to the south, is the Gulf of Chihli (Po Hai). The area of the Yellow Sea proper (excluding the Gulf of Chihli) is 156,000 square miles (404,000 square kilometres), its mean depth is 144 feet (44 metres), and its maximum depth is 338 feet. The Yellow Sea derives its name from the colour of silt-laden water discharged from major Chinese rivers.

Origin of the name

Physical features. *Physiography and geology.* The Yellow Sea, including the Gulf of Chihli and the Korea Bay, forms a partly enclosed, flat, and shallow marine embayment. Most of the sea, which is deeper than the Gulf of Chihli, consists of an oval-shaped basin with depths of about 200 to 260 feet (60 to 80 metres). The sand deposits of the Great Yangtze Bank lie off the Yangtze River to the south, near the East China Sea.

The Yellow Sea as it now exists was formed in the postglacial epoch (*i.e.,* not more than 11,000 years ago). Generally, the east and west coastal regions have a sandy bottom because of the strong tidal currents; the central basin has muddy silt; there is also a preponderance of mud on the side nearer the China mainland because of the vast amount of sediment discharged by the Huang Ho (Yellow River) and other major rivers.

Climate. Generally, the climate is characterized by very cold, dry winters and wet, warm summers. From late November to March, a strong northerly monsoon prevails, which in the Gulf of Chihli is sometimes accompanied by severe blizzards. Typhoons occur in summer, and in the colder season there are occasional storms. Air temperatures range from 82° to 50° F (28° to 10° C) and precipitation from about 20 inches (500 millimetres) in the north to 79 inches in the south. Sea fog is frequent along the coasts, especially in the upwelling cold-water areas.

Hydrography. The warm current of the Yellow Sea is a part of the Japanese Tsushima Current, which diverges near the western part of Kyūshū Island and flows at less than 0.5 knots (nautical miles per hour) northward into the middle of the sea. This warm current is modified by the low salinity of the coastal water but still maintains its relatively higher salinity. On the other hand, along the continental coasts southward-flowing currents prevail, which strengthen remarkably in the winter monsoon period when the water is cold, turbid, and of low salinity.

The tidal range is considerable (13 to 26 feet) along the shallower west coast of the Korean peninsula, with a maximum spring tide of almost 27 feet. Along the coasts of China it amounts to about three to 10 feet (except the Gulf of Chihli, which has more than 10 feet). In the Yellow Sea, the tide rises twice daily. The tidal system rotates in a counterclockwise direction. The speed of the tidal current is less than one knot in the middle of the sea, but near the coasts and in the straits and channels stronger currents of more than two or three knots are recorded.

The innermost coastal sections of the Gulf of Chihli freeze in winter, and drift ice and ice fields hinder navigation in parts of the Yellow Sea. Surface temperature ranges from freezing level in winter in the Gulf of Chihli to summer temperatures of from 72° to 82° F (22° to 28° C) in the shallower parts. Thus, the annual range is very large—from 40° to 50° F (4° to 10° C), a good indication of continentality. In winter, the temperature and salinity in the sea are homogeneous from surface to bottom. In spring and summer, the upper layer is warmed and diluted by the fresh water from rivers, while the deeper water remains cold and saline. This deep layer of cold water stagnates and moves slowly south in summer. Around this mass of water, especially at its southern tip, concentrations of commercial bottom fishes are found. The dominant salinity in the Gulf of Chihli is 30 to 31 parts per thousand (‰), and that in the Yellow Sea proper is 31 to 33 parts per thousand. In the southwest monsoon season (June to August), the increased rainfall and runoff cause a very low salinity in the upper layer in summer.

Temperature and salinity

The economy. The Yellow Sea, like the East China Sea, is famous for its fishing grounds. Since the early 20th century, rich demersal (bottom-dwelling) fishing resources have been exploited by Japanese, Chinese, and Korean trawlers. The annual catch amounts to several hundred thousand tons and consists of sea bream, croakers, lizard fish, prawns, cutlass fish, horse mackerel, squids, flounders, and other species. Overfishing and the consequent decline of fisheries had, by the 1970s, been brought under control. Along the coastal waters, spawning areas and fish nurseries are established. (M.U.)

LENA RIVER

The Lena is, in terms of water volume, the second largest river in the Soviet Union; in terms of length it is one of the largest rivers in the world, flowing 2,730 miles (4,400 kilometres) from its source in a small lake in the mountains west of Lake Baikal, in Central Asia, to the mouth of its delta on the Arctic Laptev Sea. The area of the river basin is about 961,000 square miles (2,490,000 square kilometres).

Physical features. *Physiography.* The Lena (which the Yakuts call Big River) has three main sections, each of about 900 miles (1,450 kilometres): the upper section from the source to the Vitim tributary; the middle course from the Vitim to the mouth of the Aldan; and the lower section from the Aldan to the Laptev Sea.

In the section from the source to the Vitim, the Lena flows in a deep-cut valley, the rocky and steep slopes of which are raised above the river, often up to 1,000 feet (300 metres). These slopes are formed on the right bank by the ledges of the Northern-Baikal Upland. The width of the river valley varies from one to six miles, but here and there it narrows in ravines to only 700 feet. The best known ravine, named Pyany Byk (Drunken Bull), is situated 147 miles below Kirensk.

In the first 110 miles from its source the Lena has a great number of shallow, rocky shoals, which occur as far as the Kirenga tributary. Below the mouth of the Kirenga the depth of the riverbed in the pools increases to 30 feet, and the rate of the flow decreases with a decrease in gradient. In the middle course, from the mouth of the Vitim to the Aldan, the Lena becomes a large, deep river. The water supply increases, especially after the junction with the Olyokma, and the width of the river reaches one mile. From the mouth of the Vitim to the Olyokma, the river skirts the Patom Plateau, on the right bank, forming an enormous bend; the width of the valley increases in places to 20 miles. Its slopes are gentle and green with forests, and along them run well-marked terraces formed by rivers. In this section of the valley there occurs an extensive water meadow, in which small lakes are scattered.

Below the Olyokma, the character of the valley changes sharply. In a stretch of about 400 miles, from Olyokma to Pokrovsk (60 miles above Yakutsk), the Lena flows along the bottom of a narrow valley with sheer, broken slopes. The enormous rocks sometimes resemble the ruins of a castle, or columns, or the figures of people and animals; and the area is a favourite place for tourists and rock climbers. In this section the Lena receives its largest tributaries. In addition to the Vitim aand the Aldan, these include the Great Patom and the Olyokma, flowing in on the right, and the Nyuya, on the left. Below the mouth of the Aldan the width of the valley of the Lena extends 12 to 16 miles, and the width of the water meadow reaches four to nine miles. Here the Lena enters into the borders of the Central Yakut Plain. The water meadow abounds with lakes, often marsh-ridden, and the riverbed divides, forming many islands and branches. The depth is from 50 to 70 feet, but there are many shallow sections with sandbanks.

Scenic attractions

In the lower section between the island of Zholdongo and the beginning of the delta the Lena Valley becomes narrow, its width being about one mile. The islands of the delta are low-lying, and covered with peat bogs; some fossilized ice may be found.

Climate. The climatic features of the Lena Basin are determined by its location well inside the Asian continent. In winter a powerful anticyclone is formed, the spur of which occupies all of Eastern Siberia. Because of the anti-

cyclone the winter is notable for its clear skies and prevailing calm. The frosts reach −76° to −94° F (−60° to −70° C), the average monthly air temperature in January being −22° to −40° F. In July averages range between 50° and 68° F (10° to 20° C). Due to its remoteness from the sea, the amount of precipitation in the basin is slight. Only in the southern mountains does the yearly total reach 24–28 inches (600–700 millimetres); in most of the basin it is 8–16 inches and in the delta four inches. Between 70 and 80 percent of the precipitation falls in the summer in the form of steady rain. In winter, on the average, not more than two inches of snow fall, so that the snow cover is very slight.

The very cold temperatures result in large accumulations of ice in the form of icy knolls. These are formed of groundwaters that accumulate between the layers of soil frozen permanently over many years and layers of seasonally frozen soil. Sometimes the icy knolls disintegrate with considerable force, scattering ice blocks. The riverbeds and floodlands are also covered with permanent ice in some places.

Plant and animal life. The vegetation of the Lena Valley reflects the features of the natural zones across which the river flows. The main part of the river basin is covered with taiga (swampy northern coniferous forest); in the lower reaches are found tundra and scattered forest. Spruce, cedar, birch, and poplar predominate in the moister regions. In the central valley occur some expanses of steppe, a rare occurrence above latitude 60° N. Characteristic of the water-meadow area are peat bogs and swamps and thickets of willow, alder, and dwarfish birches. In the lower valley course appear tundra plants such as mosses and lichens, partridge grasses, whitlow grass, arctic poppy, and cotton grass.

The plankton of the Lena is rather poor and restricted in variety. More than 100 species of animals are found. Commercially important fish include sturgeon, salmon, roach, dace, and perch. The main concentration of these varieties is in the mouth region, where in the summer there is comparatively warm water.

Hydrography. More than 95 percent of the Lena's water is obtained from melting snow and from rain; 1 to 2 percent of the yearly flow comes from subsoil waters. Typical of the Lena Basin are high floods, especially flash floods, in summer, and very little flow in winter. Complete cessation of flow may occur with the freezing of the river through to the bottom. The average volume of the Lena at the mouth is 21,300 cubic yards (16,300 cubic metres) per second; the greatest volume is 260,000 cubic yards per second, and the minimum is 478 cubic yards per second. The total yearly volume approaches 100 cubic miles. During the high-water period the water level rises by an average of 30 to 50 feet, and in the lower course reaches 60 feet.

The river's volume

The highest temperature of the water in the upper course of the river reaches 66° F (19° C) and in the lower course about 57° F (14° C). The river is free of ice in the south for five to six months and in the north four to five months. The breakup of ice in the spring causes significant damage to the shores: the ice floes grind the rocks, pull trees out by their roots, and carry away large sections of the banks.

The Lena carries out into the sea about 12,000,000 tons of suspended alluvium and 41,000,000 tons of dissolved matter a year. The proportion of suspended alluvium in the water is small: even in floodwaters it does not exceed 50 to 60 grams per cubic metre. The mineralization of the water in the lower Lena during low water is from 80 to 100 grams per cubic metre, and in floodwaters 160 to 500 grams per cubic metre.

The people. Among the peoples living in the Lena Basin, mainly on the banks of the river and its tributaries, are Russians, Evenks, Yakuts, and Yukaghirs. There are many industrial and cultural centres, collective farms, and state farms. In the territory of the basin is found the Yakut Autonomous Soviet Socialist Republic, with the capital of Yakutsk.

The economy. The Lena is navigable by small boats in sections below Kachug and from Ust-Kut by larger vessels. The largest wharves are Bulun, Zhigansk, Yakutsk, Vitim, Kirensk, Osetrovo, Zhigalovo, and Kachug; Osetrovo is an important river port, equipped with modern machinery. The largest navigable tributaries are the Kirenga, Vitim, Olyokma, Aldan, and Vilyuy. Wood products, furs, gold, mica, industrial products, and food are the main cargoes.

Within the borders of the Central Yakut Plain various agricultural crops—barley, oats, wheat, potatoes, cucumbers, and others—are grown. The large meadows and pastures present successful livestock raising. Future prospects for the economic development of the Lena Basin are extremely good. The basin is rich in useful fossil minerals, and gold and coal occur in exploitable quantities. In 1955, in western Yakutia, rich deposits of diamonds were discovered in the Vilyuy Basin. To the east of the diamond-bearing region are found vast deposits of natural gas; near Olyokminsk are found salt beds, and to the south of Yakutsk, iron ore and coking coal.

The hydroenergy potential of the Lena and its tributaries is about 40,000,000 kilowatts.

Study and exploration. The first exploration of the Lena was conducted by Russians at the beginning of the 17th century. In 1631 a fortress and a settlement was founded at Ust-Kut. The first scientific research was conducted by the Great Northern Expedition in 1733–43. Cartography was begun in 1910. In 1912, in the icebreakers "Taymir" and "Vaygach," the delta was surveyed and mapped. Further surveying was conducted in the interwar period, and is currently being conducted by the Yakut branch of the Academy of Science of the U.S.S.R. and by other government bodies. (I.V.P.)

OB RIVER

One of the greatest rivers of Asia, the Ob flows across Western Siberia in a twisting diagonal from its southeastern sources in the Altai Mountains to its northwestern outlet through the Gulf of Ob (Obskaya Guba) into the Kara Sea of the Arctic Ocean. It is a major communications artery, crossing territory at the heart of the Soviet Union that is extraordinarily varied in terms of physical environment and the character of its peoples: even allowing for the barrenness of much of the region surrounding the ice-threatened lower course of the river and the inhospitable waters into which it discharges, it drains a region of great economic potential, much of which is being realized under long-term Soviet development plans.

Physical features. The Ob proper is formed by the junction of the Biya and Katun rivers, in the foothills of the Soviet sector of the Altai, from which it has a course of 2,268 miles (3,650 kilometres); but, if the Irtysh River is regarded as part of the main course rather than as the Ob's major tributary, then the maximum length, from the source of the Black (Cherny) Irtysh in China's sector of the Altai, is 3,362 miles (5,410 kilometres) or, if the Gulf of Ob be included, 3,959 miles (6,370 kilometres). The catchment area is approximately 1,150,000 square miles (2,975,000 square kilometres) down to the delta's end or 1,343,000 square miles (3,479,000 square kilometres) if the gulf be included—the aggregate 172,000 square miles (445,000 square kilometres) of undrained land in the south of the basin being included in both totals. The Ob's catchment area constitutes about half of the whole Kara Sea; it is the largest Soviet catchment area and the sixth largest in the world.

The basin. The West Siberian Plain covers about 85 percent of the Ob Basin, the rest of which is occupied in the south by the terraced plains of Turgay (Kazakhstan) and the small hills of northernmost Kazakhstan, and in the southeast by the Kuznetsky Alatau, by the Salair Range (Salairskoye Kryazh), by the Mountains of Shoria, and, behind them, by the Altai Mountains.

There are more than 1,900 rivers within the basin, their aggregate length being about 112,000 miles. The Irtysh, a left-bank tributary 2,639 miles long, itself drains 634,362 square miles (a somewhat larger area than that drained by the Upper and Middle Ob before the Irtysh confluence); and some 70 percent of the whole basin is drained by left-bank tributaries.

The huge basin of the Ob could be taken, geographically,

as a representative cross section of Soviet territory: in the far south, around Lake (Ozero) Zaysan (recipient of the Black Irtysh and source of the Irtysh proper), there is arid semidesert; in the central regions of the West Siberian Plain—that is to say, over more than half of the basin—there is the swampy coniferous forest known as taiga, with very large expanses of marshland; and in the north there are vast stretches of the icy, treeless plains known as tundra.

The mainstream and its tributaries. The Upper Ob runs from the Biya–Katun junction to the confluence of the Tom, the Middle Ob from the Tom confluence to the Irtysh confluence, the Lower Ob from the Irtysh confluence to the Gulf of Ob.

The Biya and the Katun both rise in the Altai Mountains: the former in Lake (Ozero) Teletskoye, the latter to the south, among the glaciers of Mount (Gora) Belukha.

The Upper Ob

From their junction the Upper Ob at first flows westward, receiving the Peschanaya, the Anuy, and the Charysh tributaries from the left; for this reach, the river has low banks of alluvium, a bed studded with islands and shoals, and an average gradient of one foot per mile. From the Charysh confluence the Upper Ob flows northward on its way to Barnaul, receiving another left-bank tributary, the Aley, and widening its floodplain as the valley widens. Turning westward again at Barnaul, the river receives a right-bank tributary, the Chumysh, from the Salair Range: the valley hereabouts is three to six miles wide, with steeper ground on the left than on the right; the floodplain is extensive and characterized by diversionary branches of the river and by lakes; the bed is still full of shoals; and the gradient is reduced, but the depth increases markedly. At Kamen-na-Obi, however, where the river begins to loop northeastward, the width of the valley shrinks to two to three miles and that of the floodplain to less than one mile, and reefs of rock emerge in places from the bed. Just above Novosibirsk another right-bank tributary, the Inya, joins the Upper Ob; and a dam at Novosibirsk forms the great reservoir known as the Ob Sea. Below Novosibirsk, where the river leaves the region of forest steppe to enter a zone of aspen and birch forest, both valley and floodplain broaden notably, till at the Tom confluence they are respectively 12 and three or more miles wide. The depth of the Upper Ob (at lower water) varies between six and a half and 20 feet (two and six metres).

The Middle Ob

The Middle Ob begins where the Tom flows into the mainstream, from the right. Taking at first a northwesterly course, the river henceforth becomes much deeper and wider, especially after receiving its mightiest right-bank tributary, the Chulym, shortly below the confluence of the Shegarka from the left. Successive tributaries of the northwesterly course, after the Chulym, include the Chaya and the Parabel (both left), the Ket (right), the Vasyugan (left), the Tym (right), and the Vakh (right). Down to the Vasyugan confluence the river passes through the southern belt of the taiga (marshy forest country); thereafter it enters the middle belt. Below the Vakh confluence the Middle Ob changes its course from northwesterly to westerly and receives more tributaries: the Tromyegan (right), the Great (Bolshoy) Yugan (left), the Lyamin (right), the Great (Bolshoy) Salym (left), the Nazym (right), and finally, at Khanty-Mansiysk, the Irtysh (left, as has been said). In its course through the taiga the Middle Ob has a minimal gradient, a broadening valley (18–30 miles wide), and a correspondingly broadening floodplain (12–18 miles), through which it flows in a complex network of channels, with the main bed widening from less than one mile on the higher reaches to nearly two miles at the end and becoming progressively free of shoals. Low-water depths vary between 13 and 26 feet (four and eight metres); and at high water there are great floods every year, sometimes spreading from 15 or even 50 miles across the valley and lasting from two to three months.

The Lower Ob

From its start at the Irtysh confluence the Lower Ob flows northwestward as far as Peregrebnoye and thereafter northward, crossing the northern belt of the taiga till it enters the zone of forest tundra in the vicinity of its delta. The valley is wide, with slopes steeper on the right than on the left, and the vast floodplain is much intersected by channels of the river and dotted with lakes. Below Peregrebnoye the river divides itself into two main branches: the Great (Bolshaya) Ob, which receives the Kazym tributary and the Kunovat from the right; and the Little (Malaya) Ob, which receives the Northern (Severnaya) Sosva, the Vogulka, and the Synya from the left. These main branches are reunited below Shuryshkary into a single stream nearly 12 miles wide; but after the confluence of the Poluy (from the right) the river divides itself again to form a delta, the two principal arms of which are the Khamanelskaya Ob, which receives the Shchuchya from the left, and the Nadymskaya Ob, which is the more considerable of the pair. At the base of the delta lies the Gulf of Ob, which represents a 500-mile extension of the river's valley invaded by the sea, with a width reaching 50 miles at certain points and its own catchment area (forest tundra and tundra proper) of more than 40,000 square miles.

Climate and hydrology. The Ob Basin has short, warm summers and long, cold winters. The average temperature for the year ranges from 14° F (−10° C) on the shores of the Kara Sea to 34° F (1° C) and 36° F (2° C) in the forest zone and to 38.8° F (3.8° C) on Lake (Ozero) Zaysan. The absolute maximum temperature, in the arid south, is 104° F (40° C); the minimum, in the Altai Mountains, is −76° F (−60° C). Rainfall, which occurs mainly in the summer, varies between averages of 12 inches (300 millimetres) a year in the north to 20 inches in the taiga zone and 100 inches on the steppes, but the western slopes of the Altai may receive as much as 62 inches a year. Snow cover, which lasts for 240–270 days in the north and for 160–170 in the south, is deepest in the forest zone (24–36 inches) and in the mountains (80 inches), much shallower on the tundra (12–20 inches), and very thin on the steppe (8–16 inches).

Annual flooding

On the Upper Ob the spring floods begin very early in April, when the snow on the plains is melting; and they have a second phase, ensuing from the melting of snow on the Altai Mountains. The Middle Ob, scarcely affected by the Upper Ob's phases, has one continuous spring–summer period of high water, which begins in mid-April. For the Lower Ob, high water begins later in April or early in May. Levels in fact begin to rise when the watercourse is still obstructed by ice; and maximum levels, which will have been attained by May on the Upper Ob, may not be attained till June, July, or even August on the lower reaches. For the Upper Ob, the spring floods are over in July, but autumnal rains bring high water again in September–October; for the Middle and for the Lower Ob, the spring–summer floodwaters gradually recede until freezing sets in (on the lower reaches, flooding may last four or five months). The rising of levels on the Ob proper and on the Irtysh obstructs the drainage of the minor tributaries' individual catchment areas.

The autumnal ice drift on the Ob lasts from about October 31 to November 10; then the lower reaches begin to freeze solid; by November 25 the whole river is frozen; and the upper reaches remain so for some 150 days, the lower for 220. The thawing of the ice, which takes longer than the freezing, lasts from the end of April (upstream) to the end of May; and the spring drift (about five days in duration) produces considerable ice jams.

The difference of level between high water and low is approximately 10 feet on the Upper Ob; 36 feet on the Middle Ob down to Aleksandrovskoye, but only about 28 feet between Surgut and the Irtysh confluence; and at most 39 to 40 feet on stretches of the Lower Ob, but less than 20 feet at the mouth of the river.

The water is warmest in July, reaching a maximum of 82° F (28° C) in the vicinity of Barnaul.

In terms of drainage, the Ob is the third greatest river of the U.S.S.R.—after the Yenisey and the Lena. Every year it pours about 515,000,000,000 cubic yards (394 cubic kilometres) of water into the Arctic Ocean—about 12 percent of that ocean's total intake from drainage.

The volume of flow at Salekhard, just above the delta, is nearly 56,000 cubic yards (42,800 cubic metres) per second at its maximum, 2,600 (2,000) at its minimum, while for Barnaul, on the Upper Ob, the corresponding figures are 12,673 cubic yards (9,690 cubic metres) per

second and 211 (161). Most of the water comes from the melting of seasonal snow and from rainfall; much less of it comes from ground drainage, from mountain snow, and from glaciers.

The waters of the Ob are only slightly mineralized: dissolved substances account for an annual outpouring of 30,200,000 tons of ions into the Kara Sea. The total amount of solid matter brought down by the Ob every year amounts only to 50,000,000 tons.

Plant and animal life. Rich meadows extend in bands one to two miles wide for great distances along the banks of the Ob and cover many of the numerous islands. Pine, cedar, silver fir, aspen, and birch also grow on the banks and occasionally constitute isolated forests on the higher ground of the floodplain; and large areas near the river are covered with willow, snowball trees (*Viburnum*), bird cherry (*Prunus padus*), buckthorn (*Hippophaë*), currant bushes, and wild roses.

Of some 50 species of fish to be found in the river or in the gulf, the most valuable economically are sturgeon, sterlet, and such "whitefish" as nelma (*Stenodus leucichthys nelma*), musksun (*Coregonus muksun*), chir (*C. nasus*), and pelyad (*C. pelea*); pike, burbot, Siberian dace, carp, and perch are also caught. For lack of oxygen in the water, however, many fish die every winter in the reaches between the Tym confluence and the delta.

Fur-bearing mammals of the Ob Valley include European and Siberian mole, Siberian and American mink, ermine, fox, wolf (in the taiga), elk, white hare, water rat, muskrat, otter, and beaver. Among more than 170 species of birds breeding in the floodplain are grouse, partridge, goose, and duck.

The people. Politically, most of the Ob Basin belongs to the Russian S.F.S.R., but the south of it forms the northernmost part of Kazakhstan. Russians, Ukrainians, and Belorussians constitute the majority of the population, but there are, of course, numerous non-Slavic peoples also. These include the Kazakhs in the south, the Altai and Shor peoples of the mountains, the Tatars of the Irtysh Basin, the Khants (Ostyaks) and the Mansi (Voguls), whose *natsionalny okrug* (national area) occupies part of the taiga; and the Nenets, Nganasan, Enets, and Selkup peoples of the north. The valleys of the river are more densely populated than other parts of the basin.

The economy. The earliest known sailing directions for the Lower Ob are those appended to Pyotr Ivanovich Godunov's *Chertyozh* ("Chart") of Siberia (1667); further details were provided by Semyon Ulyanovich Remezov in his *Chertyozhnaya Kniga Sibiri* ("Chart-Book of Siberia"; completed 1701); and the Russian scientists of the Great Northern Expedition (1733–43) investigated the Lower Ob as well as other Siberian rivers. For the next 150 years the river system was investigated chiefly for purposes of communication within the basin. Hydrological studies, however, were inaugurated by the end of the 19th century and were pursued intensively in the 20th; and during the Soviet period the hydroelectric potential of the rivers was not only studied but also developed.

Exploration and development

The Ob's total hydroelectric potential is estimated at 250,000,000,000 kilowatt-hours. There are three main stations: one on the Ob proper, at Novosibirsk, the other two on the mountainous reaches of the Irtysh, at Bukhtarminsk and at Ust-Kamenogorsk.

The Ob, one of Western Siberia's principal means of communication, is navigable for 190 days of the year on its upper reaches, for 150 on its lower. It serves the basin both for importing and for exporting. The Trans-Siberian Railway crosses the Irtysh at Omsk and the Upper Ob at Novosibirsk. Railways going into Kazakhstan from Novosibirsk and from the foothills of the mountains cross the Upper Ob at Barnaul. (L.K.M.)

YENISEY RIVER

One of the world's great rivers, the Yenisey (from the Evenki name Ioanesi, Great River) runs from south to north across the Asiatic heart of the Soviet Union. It traverses a vast region of strikingly varied but generally old landscapes in which are found ancient people and customs as well as enormous programs of economic development.

General course of the river

Physical features. The river begins at the Siberian city of Kyzyl at the confluence of two of its tributaries—the Great Yenisey (Bolshoy Yenisey), or By-Khem, which rises on the Eastern Sayan Mountains of the Tuva A.S.S.R. (Tuvinskaya A.S.S.R.), and the Little Yenisey (Maly Yenisey), or Ka-Khem, which rises in the Darkhat Bowl of the Mongolian People's Republic. From the confluence the Yenisey River runs for 2,167 miles (3,487 kilometres), mainly along the border between Eastern and Western Siberia, before emptying into the icy Kara Sea. If the Great Yenisey is reckoned as the source, then the river as a whole is 2,543 miles (4,092 kilometres) long; if the Little Yenisey, then 2,549 miles (4,102 kilometres). The headwaters of the Selenga River, which rise in eastern Mongolia and flow through Lake Baikal (Ozero Baikal), the world's largest freshwater lake, into the Angara tributary of the Yenisey, may be reckoned as the river's ultimate source, in which case its length totals 3,442 miles (5,540 kilometres) draining a basin that at 996,000 square miles (2,580,000 square kilometres) is larger than all but six of the nations of the world. The system within Soviet boundaries comprises some 20,000 tributary or subtributary streams, with an aggregate length of some 550,000 miles. All the major tributaries of the Yenisey join its right-bank zone, a region constituting 80 percent of the basin area.

Physiography. Extending for some 2,175 miles from south to north and for some 1,060 miles from east to west, the Yenisey Basin exhibits a considerable diversity of features. Lowlands constitute only 6 percent or 7 percent of the total area: a narrow strip on the edge of the West Siberian Plain and part of the North Siberian Lowland (Severo Sibirskaya Nizmennost). In the south, the Western and Eastern Sayan, the Tuva, the Baikal, and the Khentey mountains constitute a larger proportion of the basin's area, with elevations mostly between 2,300 and 7,200 feet (700–2,200 metres), steep valleys, and vast bowls between ranges. In southern Tuva and in the Sayans there are some magnificent higher peaks, culminating in Mount (Gora) Munku-Sardyk (11,450 feet [3,491 metres]). Most of the basin stretches over the western sector of the Central Siberian Plateau, with elevations between 1,640 and 2,300 feet, bordered in the northwest by the Putorana Mountains (Gory Putorana; rising to 5,580 feet), in the west by the Yenisey Ridge (Yeniseysky Kryazh; 3,622 feet), and in the southeast by the Angara Ridge (3,353 feet).

The Yenisey River proper is divisible into three principal sections: the 295 miles from Kyzyl to Oznachennoye on the southern edge of the Minusinsk Basin; the 544 miles from Oznachennoye to the Angara confluence; and the 1,328 miles from the Angara confluence to the sea.

For the first 115 miles from Kyzyl, the Yenisey flows westward through the Tuva Bowl (Tuvinskaya Kotlovina), often divided into branches by gravelly shoals and varying in width between 100 and 700 yards (90 and 640 metres). After receiving the Khemchik River from the left it turns northward to wind through the Western Sayan for some 180 miles down a spectacular narrow canyon (sometimes 100 yards wide) with many rapids—including the cascading Great Falls.

On flowing out of the mountains at Oznachennoye, the Yenisey broadens its valley in the Minusinsk Basin: just below the Abakan confluence the valley is more than three miles wide; the bed, about 500 yards from bank to bank, is studded with islands; the flow velocity is reduced to about two yards per second; and the huge 240-mile stretch of the narrow Krasnoyarsk Reservoir (Krasnoyarskoye Vodokhranilishche), contained on the east by northwestern spurs of the Eastern Sayan, begins. A little way upstream from Krasnoyarsk the valley becomes still broader, but there are submerged ridges of rock in the bed, representing extensions of the Yenisey Ridge; one such ridge causes the rapids near Kazachinskoye.

The Lower Yenisey

Below the Angara confluence, the right bank of the Yenisey remains mountainous and is often precipitous, but the left bank assumes the character of a floodplain. The bed, having been only 870 yards wide above the confluence, is at least 2,000 yards wide below it; depth increases to between 32 and 56 feet, and the flow velocity drops by 50 percent or more. Farther downstream, however,

just above the Podkamennaya Tunguska confluence, the Yenisey cuts through spurs of the Yenisey Ridge; rapids occur at Osinovo, and below them the river plunges down a steep and scenic gorge in which its bed is reduced to a width of 800 yards. The widening and deepening process is then resumed: the valley is about 25 miles wide around the Lower Tunguska (Nizhnyaya Tunguska) confluence and about 93 miles wide around Dudinka and Ust-Port; the bed's width increases to 2,700 and then to 5,500 yards; and depths, from a minimum of 16 feet for the whole of the Yenisey's lower course, reach a maximum of more than 80 feet.

The estuary of the Yenisey begins as far upstream as the confluence of the Kureyka (the next considerable tributary north of the Nizhnyana). Below Dudinka the bed is in places divided by islands, some of them 10 or 12 miles long; and a true delta begins north of Ust-Port, where the numerous Brekhov Islands (Brekhovskiye Ostrova) divide the river into channels, with the westernmost bank about 47 miles from the easternmost. The several channels are reunited in a single wide "throat" down which the river flows into the Yenisey Gulf (Yeniseysky Zaliv) of the Kara Sea.

The largest tributaries of the Upper and Middle Yenisey are the Khemchik and the Abakan from the left and the Tuva from the right. Fed chiefly by rainwater in the mountains and by snow in the foothills and on the plains, they begin their spring high water in May or June and are swollen by summer rain floods. The Angara, on the other hand, being influenced for its whole length by the fact of its rising in Lake Baikal, practically never experiences low water. With a length of 1,105 miles, and with its own basin of more than 407,700 square miles—twice the size of the Yenisey's prior to their confluence—and with a volume approaching 5,900 cubic yards (4,500 cubic metres) per second at its mouth (as against the Yenisey's annual average of about 4,500 before the confluence), the Angara might better have been recognized as the upper course of the main river than as a tributary. The Podkamennaya Tunguska and the Lower Tunguska, with an aggregate volume of more than 5,600 cubic yards per second, also play an important role in the formation of the Yenisey runoff.

Hydrology. About 50 percent of the Yenisey's water comes from snow, between 36 percent and 38 percent from rainwater, and the remainder from groundwater. For the greater part of the system the East Siberian hydrological regime prevails: violent spring floods are followed first by a rapid fall of levels, then by a slower fall, with summer and autumn rain floods punctuating the sequence; in winter the runoff is reduced sharply, but levels remain high as ice jams are formed. In terms of runoff, the Yenisey is the greatest river in the U.S.S.R., with about 150 cubic miles (600 cubic kilometres) annually. It carries about 10,500,000 tons of alluvium into the Kara Sea every year, in addition to nearly 30,000,000 tons of dissolved mineral substances. In midsummer the water temperature varies from 57° F (14° C) to 66° F (19° C), but freezing begins on the Lower Yenisey early in October and affects the whole river by early December; ice jams and underwater ice are characteristic. Thawing occurs toward the end of April on the upper reaches, in May on the middle, and from May to mid-June on the lower. The water of the Middle Yenisey is very turbid in spring and summer and contrasts sharply with the limpid water of the Angara; and in summer the two streams flow in the same bed without mingling for nine miles or so from their confluence.

Ice and alluvium

Climate. The climate of the Yenisey Basin is distinctly continental, affected by cold Arctic Ocean air masses, and its cold season prevails from late September to mid-June in the north and from mid-October to late April in the south. Even summer is cold in the northern basin, with temperatures between 46° and 54° F (8° and 12° C) only in July, when frosts may still occur; but summer is hot in the south, with July temperatures between 64° and 68° F (18° and 20° C). The average temperature for January is below 4° F (−20° C): though it rises to 3° F (−16° C) in the foothills of the Sayan Mountains, it falls to −18° F (−28° C) and in places to −33° F (−36° C) in the Central Siberian Plateau. There is an annual rainfall of about 16 or 20 inches (400 or 500 millimetres) in the west and of 12 or 16 in the east of the Central Siberian Plateau; but in the mountainous south and southeast rainfall generally ranges between 20 and 47 inches a year, while the enclosed bowls of the southwest experience much less (12 inches around Minusinsk and less than eight in Tuva). Most of the rain (from 80 to 90 percent) falls in the warmer season and chiefly in its latter period. Since the snow cover in most of the basin is slight (16 inches in the south, 24 in the north, 35 on the Yenisey Ridge alone), the soil and subsoil are frozen to a considerable depth for long periods over most of the basin and all the year round north of the Lower Tunguska.

Plant and animal life. Most of the basin is covered with taiga (dense, marshy forest), with pine predominating in the south and larch farther north. Siberian cedar grows on the southern mountains. In Mongolia, Transbaikalia, and Tuva there are also steppes, bordered in the extreme south of the Selenga Basin by semi-desert. In the far north of the basin, taiga is superseded by lonely stretches of tundra (marshy, moss-covered plain).

The Yenisey and its tributaries are rich in fish: the mountain streams of the headwaters support grayling, trout, goldilocks (*Brachymystax lenok*), roach, and dace; the middle course has sterlet, trout, goldilocks, several species of whitefish (*Coregonus*), and grayling; the lower course has Siberian lamprey, Siberian sturgeon, sterlet, alpine char, trout, gold and silver carp, pike, and many others. The estuary has fewer species of fish but is rich in the economically valuable sturgeon. The lower reaches of the Yenisey are also much favoured in summer by migrant waterfowl from the south; the small lakes and islands support ducks, geese, and swans; and the muskrat has adapted itself to the channels of the delta.

The people. The peoples of the Yenisey Valley are various. Around the western headwaters (Great and Little Yenisey) Tuvinians predominate, with an admixture of Russians promoting the economy. To the north of Tuva the Krasnoyarksy *kray* (territory) of the Russian S.F.S.R. extends down the whole valley northward to the Kara Sea; its population comprises Russians, Ukrainians, Tatars, and numerous other indigenous peoples. The Khakass people occupy an "autonomous region" southwest of Krasnoyarsk. The vast Evenk "national district," extending from south of the Podkamennaya Tunguska to north of the Kureyka, is inhabited both by the Evenk people and by Russians from the west and Yakuts from the east. In the far north the Taymyr "national district" has a majority of Russians and some Evenks, but also the Yakuts (Sakha) and Dolgan, the Nenets, and the Nganasan peoples.

The economy. Hunting, fishing, the breeding of reindeer, and fur farming are the native occupations of the more northerly peoples, and mining is also important, especially for graphite and coal. Processing and manufacturing industries are pursued in the south. The hydroelectric potential of the Yenisey is estimated at 18,000,000 kilowatts. To exploit it, the Krasnoyarsk Hydroelectric Station was built, with a reservoir area of 810 square miles (2,100 square kilometres), and the Sayan Hydroelectric Station was built farther upstream. On the Angara, the Bratsk and the Irkutsk stations provide power; and on the Khantayka River, the Khantaysk station (operative from 1970) provides power for the Norilsk industrial complex (southeast of Dudinka). By the end of the 20th century, it is predicted, the Yenisey Basin will have 18 hydroelectric stations, with a total reservoir area of more than 17,000 square miles (44,000 square kilometres).

Power and navigation

The Yenisey is regularly navigated between Oznachennoye and the sea. A great elevator capable of lifting ships along an inclined railroad between the upper and lower waters of the Krasnoyarsk Hydroelectric Station was being constructed in 1971 to permit through traffic. The chief ports are Krasnoyarsk, Strelka (at the Angara confluence), Yeniseysk, Igarka, Dudinka, and Ust-Port; seagoing vessels sail up to Igarka. Lumber is the main cargo; some of it going upstream to Krasnoyarsk, but the downstream traffic carries bread, coal, petroleum products, and machinery, as well as lumber. Tourist cruises are the latest innovation on an increasingly important river.

Study and exploration. Russians first settled on the Yenisey in 1607, when a winter station was established on the Turukhan River (a left-bank tributary joining the Yenisey just below the Lower Tunguska confluence). Novgorod merchants, however, may have been trading with peoples of the valley as early as the 11th century. In 1619 a fort was built at Yeniseysk; and in 1628 Krasny Yar (now Krasnoyarsk) was founded. From there roads went eastward into the Buryat country and southward into the fertile Minusinsk Basin. The Russian hold on the line of the Yenisey was definitively secured early in the 18th century. Exploration of the rivers was then initiated, with a detachment of the Great Northern Expedition (1733–43) operating on the Yenisey. Later, the Lower Yenisey was explored by an expedition of 1894–96; and from 1907 to 1912 a party made a more thorough investigation of the whole river. Studies for development plans or for scientific purposes continue. (C.G.T.)

Founding of Krasnoyarsk

BIBLIOGRAPHY. The works cited below represent only a fraction of the extant literature on Asia, the purpose being to provide a highly selective reading list that bears on some of the topics dealt with in the article itself. Additional references to further literature on a large number of special topics may be found in these works. For a detailed description of the land and people of Asia, see GEORGE B. CRESSEY, *Asia's Lands and Peoples: A Geography of One-Third of the Earth and Two-thirds of Its People,* 3rd ed. (1963); FRANK M. LEBAR, GERALD C. HICKEY, and JOHN K. MUSGRAVE, *Ethnic Groups of Mainland Southeast Asia* (1964); борис федорович добрынин *et al., Зарубежная Азия; Физическая география* (1956); and *Общии обзор* in *Советскии Союз* (1972). Works treating the influence of geographical conditions on political, social, cultural, and economic conditions include: J.E. SPENCER and WILLIAM L. THOMAS, *Asia, East by South: A Cultural Geography,* 2nd ed. (1971); C.A. FISHER, *South-east Asia: A Social, Economic and Political Geography,* 2nd ed. (1966); and JOSEPH E. SCHWARTZBERG (ed.), *A Historical Atlas of South Asia* (1978), which illustrates the geopolitical development of the region. OWEN LATTIMORE (ed.), *Silks, Spices and Empire: Asia Seen Through the Eyes of Its Discoverers* (1968), is a readable survey of early travel accounts. TIMOTHY SEVERIN, *The Sindbad Voyage* (1983), a book by a traveller and author, contains thorough accounts of Asiatic ports and maritime adventure; E.L. JONES, *The European Miracle: Environments, Economies, and Geopolitics in the History of Europe and Asia* (1981), is a comparative history; STANLEY WOLPERT, *Roots of Confrontation in South Asia: Afghanistan, Pakistan, India, and the Superpowers* (1982), is a history of culture and politics of the region, from ancient times to the third quarter of the 20th century.

For more specialized treatment of specific aspects of some Asian countries, the following works may be consulted: FRANK M. LEBAR (ed.), *Ethnic Groups of Insular Southeast Asia,* 2 vol. (1972–75); GEORGE B. CRESSEY, *Land of the 500 Million* (1955); KENNETH S. LATOURETTE, *The Chinese: Their History and Culture,* 3rd ed. rev. (1946); ARTHUR L. BASHAM, *The Wonder That Was India: A Study of the History and Culture of the Indian Sub-Continent Before the Coming of the Muslims,* 3rd rev. ed. (1967); JAWAHARLAL NEHRU, *The Discovery of India,* 4th ed. (1956); RUTH BENEDICT, *The Chrysanthemum and the Sword: Patterns of Japanese Culture* (1946, reprinted 1967); and GEORGE B. SANSOM, *Japan: A Short Cultural History,* 2nd rev. ed. (1952, reprinted 1978). WALTER A. FAIRSERVIS, *Asia: Traditions and Treasures* (1981), is an illustrated anthropological study. For a study of history and culture through literature, an invaluable source is G.L. ANDERSON, *Asian Literature in English: A Guide to Information Sources* (1981).

For Eastern religion and philosophy, see KENNETH P. LANDON, *Southeast Asia: Crossroad of Religions* (1949, reprinted 1969); SUKUMAR DUTT, *Buddhism in East Asia: An Outline of Buddhism in the History and Culture of the Peoples of East Asia* (1966); S. RADHAKRISHNAN, *Eastern Religions and Western Thought,* 2nd ed. (1975), and *The Hindu View of Life* (various editions); and JOSEPH M. KITAGAWA, *Religions of the East,* enl. ed. (1968). Works dealing with the varied impact of major European nations upon the peoples of Asia include: K.M. PANIKKAR, *Asia and Western Dominance: A Survey of the Vasco da Gama Epoch of Asian History, 1498–1945* (1953); EDWIN O. REISCHAUER, JOHN K. FAIRBANK, and ALBERT M. CRAIG, *East Asia: The Modern Transformation* (1965); GEORGE B. SANSOM, *The Western World and Japan* (1950); and NORMAN JACOBS, *The Origin of Modern Capitalism in Eastern Asia* (1958, reprinted 1980). The rise and development of nationalist movements in Asia is dealt with in JAN M. ROMEIN, *Das Jahrhundert asiens: Geschichte des modern asiatischen Nationalismus* (1958; Eng. trans., *The Asian Century: A History of Modern Nationalism in Asia,* 1962); PHILIP WARREN THAYER and WILLIAM T. PHILLIPS (eds.), *Nationalism and Progress in Free Asia* (1956); WILLIAM MACMAHON BALL, *Nationalism and Communism in East Asia,* 2nd ed. rev. (1956); and WILLIAM L. HOLLAND (ed.), *Asian Nationalism and the West* (1953, reprinted 1973).

Geology: JOHN W. GREGORY, *The Structure of Asia* (1929); KURT LEUCHS, *Geologie von Asien,* vol. 1, 2 pt. (1935–37); SSU-KUANG LI, *The Geology of China* (1939); JAPAN, GEOLOGICAL SURVEY, *Geology and Mineral Resources of Japan,* 3rd ed. (1977); MASAO MINATO, MASAO GORAI, and MITSUO HUNAHASHI (eds.), *The Geologic Developments of the Japanese Islands* (1965); and RAYMOND FURON, *Introduction à la géologie et à l'hydrogeologie de la Turquie* (1953). See also H.K. GUPTA and F.M. DELANY (eds.), *Zagros, Hindu Kush, Himalaya: Geodynamic Evolution* (1981); JOVAN STOCKLIN, "Structural History and Tectonics of Iran: A Review," *Bull. Am. Assoc. Petrol. Geol.,* 52:1229–1258 (1968); AUGUSTO GANSSER, *Geology of Himalayas* (1964); P.V. RAO, "Geology and Mineral Resources of India," *Int. Geol. Congr.,* 22nd session, New Delhi (1964). Other sources are REINOUT VAN BEMMELEN, *The Geology of Indonesia,* 2 vol. (1949); WARREN HAMILTON, *Tectonics of the Indonesian Region* (1979); F.A. VENINGMEINESZ, "Indonesian Archipelago: A Geophysical Study," *Bull. Geol. Soc. Am.,* 65:143–164 (1954); B.M. GOZON, "Geology of the Philippine Islands," *Petrol. Engr.,* 33:B64–66, 69 (1961).

Physical geography: General works include: PIERRE GOUROU, *L'Asie* (1953); R.R. RAWSON and W.G. EAST, *Asia* (1966); RAOUL BLANCHARD, *Asie occidentale,* and FERNAND GRENARD, *Haute Asie* (1929); J. SION, *Asie de moussouns,* 2 vol. (1928–29); R.R. RAWSON, *The Monsoon Lands of Asia* (1963); and CHIA LIN SIEN *et al., South-East Asia: A Systematic Geography* (1979). Representative photographs of the Asian landscape are presented in MARTIN HURLIMANN, *Asien: Bilder seiner Landschaften* (1956; Eng. trans., *Asia: 289 Pictures in Photogravure,* 1957). The geomorphology of Asia is treated in FRITZ MACHATSCHEK, *Das Relief der Erde,* 2 vol. (1955).

Noteworthy studies of insular areas of the region include: ERNST LÖFFLER, *Geomorphology of Papua New Guinea* (1977); TORAO YOSHIKAWA *et al., The Landforms of Japan* (1981); S.G. BURRARD and H.H. HAYDEN, *A Sketch of the Geography and Geology of the Himalaya Mountains and Tibet,* 4 pt. (1907–08; rev. ed., 1933–34), the first official publication of the government of India on the subject, profusely illustrated with charts, diagrams, cross sections, sketches, and maps; S.P. CHATTERJEE, *Physiography of India* (1968), a systematic study dividing the country into physiographic regions; MILDRED CABLE, *The Gobi Desert* (1942); and SVEN A. HEDIN, *Ater till Asien* (1928; Eng. trans., *Across the Gobi Desert,* 1931, reprinted 1968). Additional sources include ALONZO W. POND, *Climate and Weather in the Central Gobi of Mongolia* (1954); N.K. PANIKKAR and R. JAYARAMAN, "Biological and Oceanographic Differences Between the Arabian Sea and the Bay of Bengal as Observed from the Indian Region," *Proc. Indian Acad. Sci.,* 64:231-240 (1966); M.A. BEEK, *Atlas of Mesopotamia* (1962), a text dealing with salinization and delta formation, including primarily historical or archaeological maps; W.C. BRICE, *A Systematic Regional Geography,* vol. 8 of the "South West Asia Series" (1966), a reliable contemporary work; 21st INTERNATIONAL GEOGRAPHICAL CONGRESS, DELHI, *Mountains and Rivers of India* (1968), a useful guide on the origin, courses, and many special features of the rivers of the subcontinent; HERBERT CHATLEY, *The Yellow River As a Factor in the Development of China* (1939); CH'UAN-CHUN WU *et. al., Economic Geography of the Western Region of the Middle Yellow River* (1958); GEORGE BROWN BARBOUR, *Physiographic History of the Yangtze* (1935).

Contemporary descriptions and current data may be found in the *Annual Reports of the Yangtze River Commission* and in any geography or economic history of China. In Russian, an excellent account of the Yangtze is A. П. Мурaнов, *Река Хуанхэ (Желтая Река)* (1959). The basic maps of resources, with descriptive text, on the Mekong are compiled in the UNITED NATIONS, *Atlas of Physical, Economic and Social Resources of the Lower Mekong Basin* (1968). For oceanographic data on the Yellow Sea, see the *Oceanographic Handbook of the Neighbouring Seas of Korea* (1964), issued by the Fisheries Research and Development Agency of the Republic of Korea.

Plant and animal life: L.S. BERG, *Natural Regions of the U.S.S.R.* (1950; Eng. trans. from the 2nd Russian ed., 1938); H.G. CHAMPION, *A Preliminary Survey of the Forest Types of India and Burma* (1936); PHILIP J. DARLINGTON, JR., *Zoogeography* (1957, reprinted 1980); *Flore générale de l'Indochine,* 7 vol. (1905–52); GUSTAV FOCHLER-HAUKE, "Das Waldkleid un die Pflanzenbezirke Süd-Chinas," *Mitt. Geogr. Ges. Wien,* 78:158–178 (1935); *Forestry in Japan,* issued by the Tokyo Forestry Agency (1964); BUNZO HAYATA, "General Aspects of the Flora of Japan . . . ," in *Scientific Japan, Past and Present,* pp. 77–104 (1926); P. LEGRIS, *La Végétation de l'Inde* (1963); LIOU HO, *Lauracées de Chine et d'Indochine* (1934); PAUL MAURAND, *L'Indochine forestière* (1943); F.J. ORMELING, *The Timor*

Problem: A Geographical Interpretation of an Underdeveloped Island (1955); JULES VIDAL, *La Végétation du Laos* (1956); E.H. WALKER, "The Plants of China and Their Usefulness to Man," *A. Rep. Smithson. Instn.*, pp. 325–361 (1943); WANG CHI-WU, *The Forests of China* (1961); R.O. WHYTE, "The Phytogeographical Zones of Palestine," *Geogrl. Rev.*, 40:600–614 (1950).

Natural resources: E.A. ACKERMAN, *Japan's Natural Resources and Their Relation to Japan's Economic Future* (1953); VIOLET CONOLLY, *Beyond the Urals: Economic Developments in Soviet Asia* (1967); J.A. HODGKINS, *Soviet Power:* (1961, reprinted 1976); M.S. KRISHNAN, *Geology of India and Burma,* 5th ed. (1968); P.E. LYDOLPH and THEODORE SHABAD, "The Oil and Gas Industries in the U.S.S.R.," *Ann. Assoc. Am. Geogr.*, 50:461–486 (1960); A.A. MINC, "Geographische Probleme der Ausnutzung der natürlichen Ressourcen in der UdSSR," *Petermanns Geog. Mitt.*, 114:21–28 (1970); "Natural Resources in Malaysia and Singapore," in B.C. STONE (ed.), *Proceedings of the 2nd Symposium* (Kuala Lumpur, 1969); D.B. SHIMKIN, *Minerals: A Key to Soviet Power* (1953), and *The Soviet Mineral-Fuels Industries, 1928–1958: A Statistical Survey* (1963); WANG KUNG-PING, "Mineral Resources of China, with Special Reference to the Nonferrous Metals," *Geogrl. Rev.*, 34:621–635 (1944); R.O. WHYTE, *Grasslands of the Monsoon* (1968), and *Land, Livestock, and Human Nutrition in India* (1968). See also ROBERT G. JENSEN, THEODORE SHABAD, and ARTHUR W. WRIGHT (eds.), *Soviet Natural Resources in the World Economy* (1983).

Human resources and political geography: A number of Asian countries have published national or development atlases; of particular interest are those of China (1979), India (serially, 1968–), Israel (1970), Japan (1977), Malaysia (1977), Thailand (1969), and Turkey (1977). The *Tübinger Atlas des Vorderen Orients* (1977–) is an authoritative atlas of historical, cultural, and social aspects of the development of the Near East. W.G. EAST, O.H.K. SPATE, and C.A. FISHER (eds.), *The Changing Map of Asia,* 5th ed. (1971), is a general political geography. ALISTAIR LAMB, *Asian Frontiers: Studies in a Continuing Problem* (1968); and JOHN ROBERT VICTOR PRESCOTT, *Map of Mainland Asia by Treaty* (1975), concentrate on the problem of political frontiers. GUY WINT (ed.), *Asia Handbook* (1969), is a current events summary and detailed fact book dealing with the continent. C.G.F. SIMKIN, *The Traditional Trade of Asia* (1968), discusses the history of the growth of trade between the Occidental and the Asian regional zones, often with considerable geographic materials included. W.B. FISHER, *The Middle East: A Physical, Social, and Regional Geography,* 7th ed. (1978), provides complete coverage of southwestern Asia; and GEORGE B. CRESSEY, *Crossroads: Land and Life in Southwest Asia* (1960), is somewhat more thematic in its coverage. J.E. SPENCER and W.L. THOMAS, *Asia, East by South: A Cultural Geography,* 2nd ed. (1971), is a full-scale geography dealing with the region from Pakistan through Japan; whereas C.A. FISHER, *South-East Asia: A Social, Economic, and Political Geography,* 2nd ed. (1966); and DONALD W. FRYER, *Emerging Southeast Asia: A Study in Growth and Stagnation,* 2nd ed. (1979), deal in more detail with the mainland of Southeast Asia and Indonesia–Philippines. RICHARD A. BUTWELL, *Southeast Asia Today—and Tomorrow: Problems of Political Development,* 2nd ed. (1969), concentrates on the problems of political nationalism in the former colonial countries. P.E. LYDOLPH, *Geography of the U.S.S.R.*, 3rd ed. (1977), is a full-scale geography of the Soviet lands, including Central Asia and Siberia. W.H. PARKER, *An Historical Geography of Russia* (1968); and CHANAKYA SEN, *Soviet-Asian Relations in the 1970s and Beyond: An Interpretational Study* (1976), concentrate on the historical geography and political expansion of the Soviet state into Asia. EDWARD ALLWORTH (ed.), *Central Asia: A Century of Russian Rule* (1967), is a political-historical study of Russian control of Central Asia; and LAWRENCE KRADER, *Peoples of Central Asia,* 3rd ed. (1971), is primarily concerned with the historic ethnic groupings of Central Asia. Southeast Asia's political prospects are the subject of several works, including RICHARD BUTWELL, *Southeast Asia: A Political Introduction* (1975); CHAWLA SUDERSHAM (ed.), *Southeast Asia under the New Balance of Power* (1974); and JAY TAYLOR, *China and Southeast Asia: Peking's Relations with Revolutionary Movements* (1976). The political and economic developments of the Middle East are the subject of the annual *Middle East Contemporary Survey.* Economic conditions in Southeast Asia and the Far East and their implications for the Western world are discussed in ROY HOFHEINZ and KENT E. CALDER, *The Eastasia Edge* (1982).

Resource development: For the continent as a whole, the basic, most authoritative, and up-to-date sources of economic data and information are the publications of the United Nations Economic and Social Commission for Asia and the Pacific (ESCAP), formerly the United Nations Economic Commission for Asia and the Far East (ECAFE), and certain other intergovernmental organizations. The *Economic Survey of Asia and the Pacific,* prepared and published annually by ESCAP, is the most comprehensive account of the economic situation in the continent as a whole. The U.S. CONGRESS, JOINT ECONOMIC COMMITTEE, has published a number of studies on the Chinese economy, which include: *China, a Reassessment of the Economy* (1975); *China and the Chinese* (1976); and *Chinese Economy Post-Mao: A Compendium of Papers* (1980). SHIN-YONG CHUN (ed.), *Economic Life in Korea* (1978), addresses Korea's striking economic growth.

Also valuable is *The Political Economy of the Middle East* (1980), published by the U.S. CONGRESS, JOINT ECONOMIC COMMITTEE. A multi-disciplinary analysis of the problems of underdevelopment, development, and planning for development in Asian countries is found in GUNNAR MYRDAL, *Asian Drama: An Inquiry into the Poverty of Nations,* 3 vol. (1968); the countries covered include Burma, Sri Lanka, Cambodia (Kampuchea), Laos, Indonesia, Malaysia, Thailand, and Pakistan, but most of the work is devoted to India. The following studies in ESCAP's "Mineral Resources Development Series" may be particularly useful in respect to mining and industries: *Mining Developments in Asia and the Far East* (annual); *Mining Developments in Asia and the Far East: A Twenty-Year Review, 1945–1965* (1967); *Lignite Resources of Asia and the Far East: Their Exploration, Exploitation and Utilization* (1956); *Copper, Lead and Zinc Ore Resources of Asia and the Far East* (1960); *Bauxite Ore Resources and Aluminum Industry of Asia and the Far East* (1962); *Tin Ore Resources of Asia and Australia* (1964); *Mineral Raw Material Resources for the Fertilizer Industry in Asia and the Far East* (1967); *Mineral Resources of Lower Mekong Basin and Adjacent Areas of Khmer Republic, Laos, Thailand and Republic of Viet-Nam* (1972); and *Proceedings of the Third Session of the Committee on Natural Resources* (1977). For data on mineral resources located in China see K.P. WANG, "The Mineral Resource Base of Communist China," in U.S. CONGRESS, JOINT ECONOMIC COMMITTEE, *An Economic Profile of Mainland China,* vol. 1 (1967).

The numerous studies ECAFE and ESCAP have made relating to industries include *Asian Industrial Development News* (annual); *Industrial Development in Asia and the Far East* (1965); *Development Prospects of Basic Chemical and Allied Industries in Asia and the Far East* (1963); and *Industrial Developments in Asia and the Far East: Selected Documents Presented to the Asian Conference in Industrialization,* 4 vol. (1966). Helpful works dealing with industries in China include CHOH-MING LI (ed.), *Industrial Development in Communist China* (1964); FREDERICK M. CONE, *Chinese Industrial Growth: Brief Studies of Selected Investment Areas* (1968); YUAN-LI WU and KUNG-CHIA YEH, *Growth, Distribution, and Social Change: Essays on the Economy of the Republic of China* (1978); BARRY M. RICHMAN, *Industrial Society in Communist China* (1969); and YUAN-LI WU, *The Steel Industry in Communist China* (1965). The role of small industries in the development effort is discussed in "Modernization of Small Industries in Asia," *Economic Bulletin for Asia and the Far East,* 11:24–40 (1960); and CARL RISKIN, "Small Industry and the Chinese Model of Development," *China Quarterly,* 46:245–273 (1971). AMNUAY TAPINGKAE (ed.), *Higher Education and Economic Growth in Southeast Asia* (1976), examines the role that higher education plays in the economic growth of the countries of this region. Later overviews are presented in NOBUTOSHI AKAO (ed.), *Japan's Economic Security* (1983); YUJIRO HAYAMI and MASAO KIKUCHI, *Asian Village Economy at the Crossroads: An Economic Approach to Institutional Change* (1982); and K.V. SUNDARAM, *Geography of Underdevelopment: The Spatial Dynamics of Underdevelopment* (1983).

ECAFE and ESCAP publications on the development of power resources include: *Electric Power in Asia and the Far East* (annual); *Proceedings of the Regional Seminar on Energy Resources and Electric Power Development* (1962); and *The Role and Application of Electric Power in the Industrialization of Asia and the Far East* (1965). Projections of future energy requirements are made in the development plans of many countries, but the NATIONAL COUNCIL OF APPLIED ECONOMIC RESEARCH, *Demand for Energy in India, 1960–1975* (1960), is an early example of nonofficial estimates. Material on the power industry in China includes: YUAN-LI WU, *Economic Development and the Use of Energy Resources in Communist China* (1963); JOHN ASHTON, "Development of Electric Energy Resources in Communist China," in U.S. CONGRESS, JOINT ECONOMIC COMMITTEE, *An Economic Profile of Mainland China,* vol. 1 (1967); ROBERT CARRIN, *Power Industry in Communist China* (1969); and VACLAV SMIL, *China's Energy: Achievements, Problems, Prospects* (1976). Possibilities and the potential of nuclear power development are discussed in the *Reports* (1959–60) of the Preliminary Assistance Missions sent to several Asian countries by the International Atomic Energy Agency. A list of civilian power reactors in operation or under construction with all pertinent technical data may be found in the INTERNATIONAL ATOMIC ENERGY AGENCY, *Power and Research*

Reactors in Member States (published twice a year). The vast energy reserves to be found in the Soviet Union are the subject of IAIAN F. ELLIOT, *The Soviet Energy Balance: Natural Gas, Other Fossil Fuels, and Alternative Power Sources* (1974). See also TAISHIRO SHIRAI (ed.), *Contemporary Industrial Relations in Japan* (1983).

Agriculture: Among the several publications of the Food and Agriculture Organization of the United Nations (FAO) are the following: *The State of Food and Agriculture* and *FAO Rice Report* (both are published annually). The ASIAN DEVELOPMENT BANK, *Asian Agricultural Survey,* 2 vol. (1968); and YUJIRO HAYAMI, VERNON W. RUTTAN, and HERMAN M. SOUTHWORTH (eds.), *Agricultural Growth in Japan, Taiwan, Korea, and the Philippines* (1978), are comprehensive surveys covering most of Asia. A succinct account of Asian agriculture may be found in COLIN CLARK, "Agriculture in Asia," *Pacific Community,* 4:283–294 (1970); and the likely ways to improve its contribution to the economy are discussed in "Strategies for Agricultural Growth," *Economic Survey of Asia and the Far East, 1969,* pt. 1A (1970). The agricultural situation in China is described in J.L. BUCK, O.L. DAWSON, and YUAN-LI WU, *Food and Agriculture in Communist China* (1966); and JOLO BUCK, *Three essays on Chinese Farm Economy* (1980). The growth in cereals output and its economic and social impact may be found in LESTER R. BROWN, "The Agricultural Revolution in India," *Foreign Affairs,* 46:688–698 (1968); SAM-CHUNG HSIEH, "New Outlook for Asian Agriculture," *International Development Review,* 10:6–9 (1968); UNITED STATES DEPARTMENT OF AGRICULTURE, *Taiwan's Agricultural Development: Its Relevance for Developing Countries Today* (1968); as well as NORMAN E. BORLAUG, "The Green Revolution: For Bread or Peace?" *Bull. Atom. Scient.* (1971). An informative collection of data is JOHN W. LONGWORTH, *Beef in Japan: Politics, Production, Marketing, and Trade* (1983).

Commerce: (*Transport*): The status and problems of transport in Asian countries and its role in their economic development is exhaustively dealt with in "Transport Development," *Economic Bulletin for Asia and the Far East,* vol. 11, no. 3 (1960). The Asian Highway Project, designed to connect the capitals and seaports of Asian countries, is described by M.S. AHMAD, "The Asian Highway," *ibid.,* 19:45–48 (1968). Despite the role of animals as a major mode of transport in the rural areas of most Asian countries, not much literature is available on the subject; the INDIAN PLANNING COMMISSION, *Role of Bullock Carts and Trucks in Rural Transport: Case Studies* (1963), is most valuable. WILFRED OWEN, *Distance and Development: Transport and Communications in India* (1968), is an excellent treatment of the subject for India; as are VICTOR D. LIPPIT, "Development of Transportation in Communist China," *China Quarterly,* 27:101–119 (1966); and YUAN-LI WU, *The Spatial Economy of Communist China: A Study on Industrial Location and Transportation* (1967). (*Trade*): A historical perspective of Asian trade from its remote beginnings may be found in C.G.F. SIMKIN, *The Traditional Trade of Asia* (1968). J.C. VAN LEUR, *Indonesian Trade and Society: Essays in Asian Social and Economic History* (1955), includes a sociological interpretation of early Asian trade. B.G. GHATE, *Asia's Trade* (1948), is a descriptive account of the situation up to the early 1940s; and ALFRED K. HO, *The Far East in World Trade: Developments and Growth Since 1945* (1967), brings the general survey up to 1960. Among several studies prepared by ECAFE, the following are particularly useful: "Asia's Trade with Western Europe, with Special Reference to the Common Market," *Economic Survey of Asia and the Far East, 1962,* pt. 1 (1963); "Foreign Trade of ECAFE Primary Producing Countries," *Economic Survey of Asia and the Far East, 1959,* pt. 2 (1960); "Trade Between Developing ECAFE Countries and Centrally-Planned Economies," *Economic Bulletin for Asia and the Far East,* 15:16–51 (1964); and "Intra-Regional Trade As Growth Strategy," *Economic Survey of Asia and the Far East, 1969,* pt. 1B (1970). SEIJI NAYA, "The Commodity Pattern and Export Performance of Developing Asian Countries to the Developed Areas," *Economic Development and Cultural Change,* 15:420–437 (1967), is comprehensive and analytical. Material relating to the foreign trade of China includes: PAULINE LEWIN, *The Foreign Trade of Communist China: Its Impact on the Free World* (1964); and ALEXANDER ECKSTEIN, *Communist China's Economic Growth and Foreign Trade: Implications for U.S. Policy* (1966). Later monographs include DENNIS T. YASUTOMO, *Japan and the Asian Development Bank* (1983); and RAE WESTON, *Gold: A World Survey* (1983).

Demographic patterns: GEORG BORGSTROM, *The Hungry Planet: The Modern World at the Edge of Famine,* 2nd rev. ed. (1972); S. CHANDRASEKHAR, *India's Population: Facts, Problem and Policy* (1967), *Hungry People and Empty Lands,* 3rd ed. (1954), *Population and Planned Parenthood in India,* 2nd ed. (1961), *Communist China Today,* 4th rev. ed. (1964), *Asia's Population Problems* (1967, reprinted 1977), and *Infant Mortality, Population Growth and Family Planning in India* (1972); C.D. COWAN (ed.), *The Economic Development of Southeast Asia* (1964); KINGSLEY DAVIS, *The Population of India and Pakistan* (1951); PHILIP M. HAUSER (ed.), *Urbanization in Asia and the Far East* (1957); ALFRED DE SOUZA (ed.), *The Indian City: Poverty, Ecology and Urban Development* (1978); LEA JELLINEK, CHRIS MANNING, and GAVIN JONES, *The Life of the Poor in Indonesian Cities* (1978); PING-TI HO, *Studies on the Population of China, 1368–1953* (1959); JACQUES M. MAY, *The Ecology of Malnutrition in the Far and Near East* (1961); POLITICAL AND ECONOMIC PLANNING (PEP), *World Population and Resources* (1955); JOHN ROBBINS, *Too Many Asians* (1959); EDGAR SNOW, *Red China Today: The Other Side of the River,* rev. ed. (1971); JOSEPH E. SPENCER, *Asia, East by South: A Cultural Geography* (1954); MORRIS B. ULLMAN, *Cities of Mainland China: 1953 and 1958* (1960); United Nations, ECONOMIC AND SOCIAL COUNCIL, *Report on the World Social Situation* (1957–); United Nations, *Department of Economic and Social Affairs, The Future Growth of World Population* (1958); and C.K. YANG, *A Chinese Village in Early Communist Transition* (1959). ESCAP has published a number of country monographs that analyze the demographic trends of individual Asian countries. Transformations in Asian nations are analyzed in BENEDICT ANDERSON, *Imagined Communities: Reflections on the Origin and Spread of Nationalism* (1983).

Contemporary studies of international issues involving Asia include EDWARD INGRAM, *The Beginning of the Great Game in Asia, 1828–1834* (1979), and his *Commitment to Empire: Prophesies of the Great Game in Asia, 1797–1800* (1981); JOHN P. FOX, *Germany and the Far Eastern Crisis, 1931–1938: A Study in Diplomacy and Ideology* (1982); KEITH ST. CARTMAIL, *Exodus Indochina* (1983); LINDA MASON and ROGER BROWN, *Rice, Rivalry, and Politics: Managing Cambodian Relief* (1983); RICHARD H. SOLOMON (ed.), *Asian Security in the 1980s: Problems and Policies for a Time of Transition: Conference on East Asian Security in the 1980s* (1980); RUSSELL D. BUHITE, *Soviet-American Relations in Asia, 1945–1954* (1981); DONALD S. ZAGORIA (ed.), *Soviet Policy in East Asia* (1982); RAJENDRA K. JAIN (ed.), *U.S.–South Asian Relations, 1947–1982,* 3 vol. (1983); and WILBUR SCHRAMM and ERWIN ATWOOD, *Circulation of News in the Third World: A Study of Asia* (1981).

ENAYETUR RAHIM, *Scholars' Guide to Washington, D.C. for South Asian Studies* (1982), is a reference source listing resources of research collections.

Atatürk

Mustafa Kemal, known as Atatürk (Father of Turks), a distinguished Turkish soldier, reformer, and statesman, was the founder of the Republic of Turkey and its first president. His successful struggle for the liberation of Turkey against the powers of the Entente (an alliance of Britain, France, and Russia) after Turkey's defeat in World War I has inspired many embryonic states in Asia and Africa to fight for their independence.

Early life and career. Mustafa was born in 1881 in Salonika, Greece, Ottoman Empire, of a Turkish family of humble origin. His mother was Zübeyde Hanım, his father Ali Riza, a minor government employee. When Mustafa was still in the primary school he lost his father, and his mother took the boy to live in the country with her brother. Returning to Salonika later, Mustafa finished primary school and entered the military secondary school in that town in order to become an officer in the Ottoman army. It was at this school that one of his teachers, who admired the boy for his skill in mathematics and who was also called Mustafa, suggested he should call himself Mustafa Kemal (maturity and perfection). After finishing secondary school, Mustafa Kemal went on to the military high school in Manastir, where, observing with hatred the continuous attacks of the Christian Macedonian anarchists on the Turkish population, he became, like most of his fellow cadets, an ardent nationalist.

Education

In 1899 Mustafa Kemal entered the Military Academy in Istanbul. There he came to take a close interest in politics; he and his fellow students read secret pamphlets attacking the despotic rule of Sultan Abdülhamid II. He was especially influenced by the patriotic and liberal thinking of the poet Namık Kemal Bey. He also read books on the French Revolution and developed an admiration for Napoleon.

In 1902 Mustafa Kemal was graduated from the Military Academy and entered the General Staff College, where his interest in politics continued. After finishing the college with the rank of captain, he was appointed to the cavalry regiment in Damascus. There, together with some of his friends, he founded a secret society called "Fatherland and Freedom," which, however, failed to make much progress. On his return to Salonika, Mustafa Kemal, like many of his fellow officers, joined the secret "Committee of Union and Progress," which spread its revolutionary activities throughout the armed forces and caused the proclamation of the Constitution of 1908.

Mustafa Kemal devoted all his time and energy to his profession in the following years. In 1911, when the Italians attacked Tripoli, an Ottoman province at that time, he hurried there together with some of his officer friends and, forming troops of the natives, launched successful guerrilla raids on the enemy. In the same year Mustafa Kemal was promoted to major. In the Balkan War of 1912 he was charged with the defense of the Gallipoli Peninsula—a task that gave him an excellent opportunity to study the strategic position of this important area. In 1913 he was sent to Sofia as military attaché, and during his stay there he acquired a good knowledge of the Western standards in taste, the arts, and the relations between men and women in polite society. He made good use of this knowledge later when he set about reforming social life in his country. While still in Sofia, Mustafa Kemal was promoted to lieutenant colonel.

Activities during World War I

When World War I broke out, Mustafa Kemal was appointed to the command of the 19th Division at Çanakkale. He defeated the British at Gallipoli twice and gained for himself in the Turkish press the title of "the Saviour of Istanbul." And he was promoted to colonel. In 1916, serving on the eastern front, he stopped the advance of the Russian forces to the south and was promoted to brigadier general.

In 1917 Mustafa Kemal accompanied the crown prince, Vahideddin, on a state visit to Germany. During a tour of the German Western Front he did not hesitate to express openly his view about the vulnerability of the front and Germany's position in the war. On his return to Istanbul Mustafa Kemal fell ill, and for treatment he went to Vienna and Carlsbad (now Karlovy Vary, Czechoslovakia) where he had a further opportunity to observe European civilization.

In 1918 Mustafa Kemal was appointed to command the 7th Army in Palestine; when he took up his duties, however, the fight with the British had all but ended, and the enemy was advancing northward without meeting any resistance. The Arab guerrillas too were launching attacks on the Turkish army. To avoid the capture of the whole 7th Army, Mustafa Kemal withdrew his forces to the north of Aleppo. When, after the Armistice of Mudros (Moúdhros), the German officers and commanders serving in Turkey returned to their country, Mustafa Kemal assumed the command of all the forces of the southeastern front. Disagreeing with the British over the enforcement of the terms of the Armistice, however, he was appointed to a post in the Ministry of War. On his arrival in Istanbul he found the fleet of the Entente anchored in the harbour. The terms of the Armistice were hard enough, but information was now received about a secret agreement reached by the states of the Entente for the partition of the Ottoman territories. Moreover, the minorities in Istanbul and elsewhere had seized the opportunity to organize themselves against the Turks. The Turkish people looked for a means of redress; in some parts of the country they formed organizations called Associations for the Defense of Rights to fight against them.

Activities after World War I. In Istanbul there were two main ideas about Turkey's future: the Sultan and his supporters were thinking of placing the country under English protection, while some well-known Turkish journalists and intellectuals were spreading propaganda for placing Turkey under an American mandate. In both cases the aim was to maintain the Ottoman Empire in its cosmopolitan structure. Mustafa Kemal, however, persisted in the idea of an independent Turkish nation living within its national boundaries and believed that this could be achieved if the nation was prepared for a new struggle. Before deciding on a course of action, he had talks with many Turkish and foreign notables, including the Sultan and his ministers. Then he discussed it with his friends, all commanders who were bitterly disillusioned over the abolition of the Ottoman army by the terms of the Armistice, and saw the solution in starting a war of independence in Anatolia.

An excellent opportunity for this presented itself soon: the powers of the Entente were putting pressure on the Turkish government to take measures against riots likely to break out in the eastern provinces. The Sultan appointed Mustafa Kemal as Inspector of the Third Army in Erzurum, endowing him with power over military and civilian authorities. On May 15, 1919, immediately before Mustafa Kemal's departure for Erzurum, the Greeks occupied Izmir.

After a secret interview with the Sultan, Mustafa Kemal left Istanbul with a large suite of staff officers and set foot in Samsun on May 19. In Amasya, with the approval of local corps commanders, he issued a secret circular dated June 22 in which he described the dangers that the country faced: how the government in Istanbul had yielded weakly to the forces of occupation and how the only hope of salvation lay in the nation's own struggle for its liberation. Such a struggle, he added, had already begun, and to make it the decision of the nation itself a national congress would be convoked in Sivas with the participa-

Mustafa Kemal's secret circular

tion of three delegates from each province. He ordered all unit commanders to strengthen their forces, disregarding the terms of the Armistice about the demobilization of the Turkish army. Finally, he warned both the military and civilian authorities that henceforth they would take their orders from him alone.

Mustafa Kemal's demands were fervently complied with by the military because his demands meant saving the army from extinction. The army took under its control all postal and telegraphic communications in Anatolia and forced into obedience those civil administrators who tried to resist Mustafa Kemal's orders.

In all the towns and cities he called at on his way to Sivas, Mustafa Kemal met the leading citizens and explained to them his views on a national struggle for independence. He arrived in Sivas amid warm demonstrations of support by the people, and after important talks with the notables of the city, he proceeded to Erzurum, ignoring all orders given by the Sultan's government, under pressure from the states of the Entente for his immediate return to Istanbul. In Erzurum a congress was to be convened on July 23 by the Association for the Defense of the Rights of Eastern Anatolia. In the meantime the military and civilian authorities of Erzurum received an order for Mustafa Kemal's arrest and transport to Istanbul. Although this order went unheeded, Mustafa Kemal resigned his commission in the army, deeming it necessary to have more freedom as the leader of the national struggle he had started. Thus he entered the congress as a mere delegate and was elected its president. At his suggestion the congress accepted the National Covenant, which was in the nature of an oath requiring the indivisibility of the fatherland and the successful completion of the national movement. In addition, a Standing Committee of nine members was elected of which Mustafa Kemal was chosen president. On September 4 he opened the National Congress in Sivas, and he was again elected president of the congress. The National Congress adopted all the decisions taken by the Congress of Erzurum and rejected decisively the idea of placing Turkey under American mandate. Resisting a motion for establishing a new state in Anatolia, Mustafa Kemal proposed the joining of all local Associations for the Defense of Rights into one society to be called the Association for the Defense of the Rights of Anatolia and Rumelia. The proposal was accepted, and so the prototype of a political party was formed. The Ottoman cabinet of Ferid Paşa and succeeding Ottoman governments continued to view the national movement as an act of rebellion and Mustafa Kemal's activities as illegitimate.

Participation in nationalist congresses

Role in the founding and reform of modern Turkey. On December 27 Mustafa Kemal transferred the seat of the national struggle to Ankara, thinking it a more convenient location for his purposes. In the meantime, at the general elections of members for the Ottoman Chamber of Deputies in Istanbul, Mustafa Kemal's supporters won an overwhelming majority and succeeded in getting the chamber to proclaim as its own decision the principles of the National Covenant. They also secured the cancellation of the government's former decree about Mustafa Kemal's dismissal from the army. Alarmed at these indications of change in the Ottoman policy, the British occupied Istanbul officially on March 16, 1920, and dissolved the Chamber of Deputies. Mustafa Kemal vehemently protested the British government's action; but, in fact, the occupation of the Ottoman capital and especially the dissolution of the chamber were extremely useful to his aims because they removed the legal obstacle that Istanbul presented to his plan of forming a national government in Anatolia. So, after a new election of deputies, he opened on April 23 in Ankara the first Grand National Assembly of Turkey, and he was elected its president. At Mustafa Kemal's proposal, a constitutional law was passed changing the name of the state to Turkey and stipulating that sovereignty and executive powers would be used on its behalf by the Grand National Assembly. Accordingly, as president of the assembly Mustafa Kemal took upon himself the offices of the prime minister and of the president of the state. Thus ended the Islāmic form of government that had existed in Turkey since the Middle Ages; and, as in the French Revolution, the Turkish people passed suddenly from rule of absolutism and the caliphate to a regime based on national sovereignty. This important change caused serious uprisings in some regions, but they were quickly suppressed by the national forces.

Mustafa Kemal now busied himself with the work of gaining control of such parts of the country as were then under occupation. First, in the east, the Armenians and the Georgians were defeated, and through the mediation of Soviet Russia, a treaty was signed with them that regained for Turkey even the territories it had lost in 1878. After extensive guerrilla warfare, the French in the south evacuated Turkish territories and withdrew to Syria and recognized the legitimacy of the National Government in Ankara. Ignoring Ankara altogether, the British got the Ottoman government to sign the Treaty of Sèvres. The National Government proclaimed that it did not recognize as legitimate a treaty of such terms, whereupon the Greek army extended its area of occupation, advancing within 50 kilometres of Ankara. At this time of great anxiety the National Assembly appointed Mustafa Kemal the commander in chief with extraordinary powers. On August 26, 1922, after an all-out offensive planned and directed personally by the Commander in Chief, the Greek army was defeated and forced within two weeks to leave Anatolia completely. Upon this decisive victory and with the mediation of the states of the Entente, an armistice was signed with the Greeks according to which they evacuated all Turkish territories. The British ceded Çanakkale and Istanbul to the National Government. Vahideddin, the last Ottoman sultan, fled abroad, and upon a motion by Mustafa Kemal the National Assembly terminated the 600 years of Ottoman rule in Turkey. The Treaty of Lausanne, signed on July 24, 1923, established the integrity of Turkey's national frontiers and its complete independence. All privileges granted to the European countries by the Ottomans were cancelled. Thus, Mustafa Kemal realized his dream of founding a completely independent and national Turkish state in place of the Ottoman Empire, that "sick man of Europe" that had been for a long time a subject of strife among the great powers of Europe.

Struggle to consolidate the country

In 1922 Mustafa Kemal married Latife Hanım, the well-educated daughter of a wealthy family in Izmir. The marriage was contracted in the modern manner, not in the tradition of Islām. In order to show the Turkish people that the place of women in society was by the side of their men, he took his wife with him on his trips around the country. His marriage did not last long, however. During his long years of single life he had developed an independent habit of living that he found difficult to give up and that his wife could not tolerate.

Marriage

On one occasion as early as 1917, Mustafa Kemal had remarked that, had he the power and the authority, he would change social life in Turkey at one blow. This opportunity had now presented itself, and he launched on a program of reforms. In place of the Association for the Defense of the Rights of Anatolia and Rumelia, he founded the People's Party (later renamed the Republican People's Party) and became its leader. With the general elections held immediately after the signing of the Treaty of Lausanne, this party, as Turkey's only political party, took complete control of the government. On October 29, 1923, Mustafa Kemal proclaimed the Republic and was elected its first president. In 1924 he abolished the caliphate. In the meantime, a group of his friends who were against his drastic methods of reform and who believed in more gradual progress over a longer period of time founded the Progressive Republican Party. Mustafa Kemal continued his reform program: he closed down all institutions based on the Muslim canon law, monasteries, and religious orders. "Science is the most reliable guide in life," he remarked, and abolishing the traditional system of education, which was mainly religious, he established secular schools of the modern type. The whole Ottoman legal system was modernized, and a new civil and penal code was adopted. His reforms penetrated the daily life of every citizen. The Oriental forms of dress that carried a religious significance were discarded in favour of European dress. Dances, balls, and other forms of entertainment in-

Reforms

volving both men and women were encouraged, and the enlightened classes adopted the European way of life.

Mustafa Kemal's reforms did not go unchallenged. In Eastern Anatolia a man called Şeyh Said stirred up a rebellion to restore the Muslim canon law; in Izmir preparations for a plot to assassinate Mustafa Kemal were reportedly discovered; and there were said also to be some local attempts at rebellion against the use of hats. Mustafa Kemal punished severely all the leaders of these movements, closed down the Progressive Republican Party, and, reverting to the former authoritarian regime, pursued his program of reform. Setting aside all the old laws and traditions that held women inferior to men, he established complete equality between the sexes, including the right of electing and being elected. In 1928 he substituted Roman characters for the Arabic that had been used in Turkey for centuries. He endeavoured to popularize Western classical music and the theatre in Turkey. In 1930 he made a second attempt at introducing a multiparty regime by allowing the creation of the Free Republican Party; but, as this party soon became a centre for antireformist ideas and activities, it met the same fate as the Progressive Republican Party. Mustafa Kemal also launched a large-scale program of research in the fields of Turkish language and history. By this means he wanted to strengthen in society the ties of national feeling in place of the old ties of religion. In 1933 a law was passed to make the use of family names compulsory, and the National Assembly gave Mustafa Kemal the name Atatürk, which soon became so popular as to supersede his previous name and titles.

Atatürk's foreign policy can be summed up by his motto: "Peace at home, and peace in the world." In economy, he followed a policy of national economy, nationalizing all foreign firms and companies. On the question of Turkey's industrialization, he placed his hope on private domestic capital for a while, but discovering its insufficiency, he decided to encourage etatism (state socialism). In neither case, however, did he achieve any important success. If one or two items of foreign policy are excepted, there was a gradual slowing down in the last five years of Atatürk's life, and his final year passed in serious illness. He died on November 10, 1938, in Istanbul, where he had gone to rest.

Atatürk made major reforms in Turkey in the field of politics, law, and culture that only affected, however, bureaucrats and a minority of well-to-do people in the cities. The poorer part of the population, and especially the peasants who still subsisted in an agricultural order of the medieval type, continued to live much the same as before. Nevertheless, the Western view of life had gained enough power among the educated classes to make a return to the old way of life impossible.

BIBLIOGRAPHY. For a complete bibliography, see J.P.D. KINROSS, *Atatürk: The Rebirth of a Nation* (1964); B. LEWIS, *The Emergence of Modern Turkey* (1961); and UNESCO, *Atatürk* (Eng. trans. 1963). Additional information may be found in the SOCIETE POUR L'ETUDE D'HISTOIRE TURQUE, *Histoire de la république Turque* (1935), which was written under the patronage of Atatürk.

(M.Ak.)

Athens

The capital of the republic of Greece and generally considered the nursery of Western civilization, Athens (Greek Athínai) lies five miles (eight kilometres) from the Bay of Phaleron, an inlet of the Aegean (Aigaíon) Sea where Piraeus (Piraiévs), the port of Athens, is situated, in a mountain-girt arid basin divided north-south by a line of hills. Greater Athens has an area of 165 square miles (427 square kilometres). The Kifisós River, only a trickle in summer, flows through the western half; and the Ilisós River, often dry, traverses the eastern half. The surrounding mountains—Párnis, 4,636 feet (1,413 metres); Pentelicus (Pendéli), 3,631 feet; Hymettos (Imittós), 3,365 feet; and Aigáleon, 1,535 feet—add to the impression of barrenness. Yet such considerations are superficial when compared with the fecundity of Athens' bequests to the world, such as its philosophy, its architecture, its literature, and its political ideals.

For treatment of the city in its regional setting, see GREECE; historical and cultural aspects are treated further in the article GRECO-ROMAN CIVILIZATION: *Ancient Greek Civilization.*

This article is divided into the following sections:

Physical and human geography

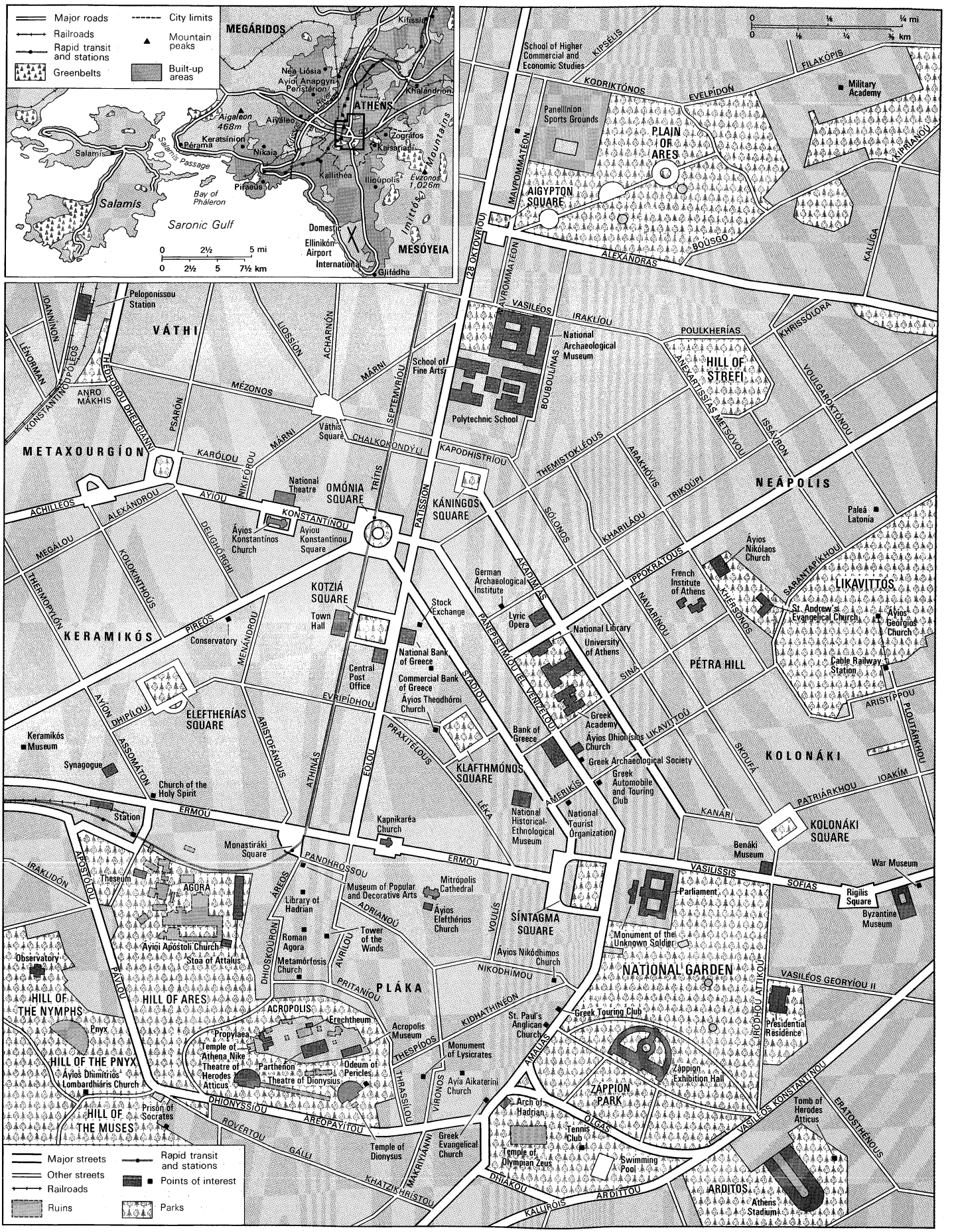
CHARACTER OF THE CITY

Athens, when approached from the Middle East, is the first European city, with tall buildings, newspaper kiosks, modern shops, and modishly dressed citizens. Approached from Europe, it seems, if not exactly the first Oriental city, at any rate not quite European, in its ill-fitting, locally tailored modernity. The European notes a medley of characterless concrete and out-of-style dress, with the smell of spitted meat and spices in narrow, unpaved streets as clamorous as bazaars and a few streets away from the centre.

Nevertheless, it is wrong to say that Athens is a mixture of East and West: it is Greek and, more particularly, Athenian. The Athenians, after all, nurtured Western civilization. Yet, some three centuries after the death of Pericles (429 BC), they entered upon a period of bondage that lasted almost 2,000 years. Athens was freed in 1833, and in the following 135 years it was the scene of 14 revolutions, another brutal foreign occupation, and a civil war of especial savagery. This long history of passion and suffering has had considerable effect on the Athenian character. The core of that character is an implacable will to survive, buttressed by a profound sense of loyalty (especially to the family) and patriotism. The Greek Orthodox Church, which is directed by a synod sitting in Athens, was a main force in keeping alive the Greek language, tradition, and literature when such things were forbidden, and most people still support it.

The 2,000-year bondage

The millennia of oppression, instead of driving the Athe-



Central Athens and (inset) its metropolitan area.

nians into obtuse moroseness, have honed their wit and rendered them tough but supple, while centuries of privation have only preserved their warmth and generosity. The long oral tradition, alive even under the invader, has reflected and stimulated a taste for rich talk. Of course, the poetic impulse to make a good story better leads to considerable exaggeration in daily conversation, suiting a vanity that goes with a sharp-edged sense of personal and family honour and the spoiling of children. The ancient heroes, too, were vain about both themselves and honour, boasting as much about outwitting the enemy as about outfighting him. Cunning, as in the *Odyssey,* is still a virtue there.

THE LANDSCAPE

Climate. The climate of Athens is benign: frost is rare (the minimum temperature is 32° F, or 0° C) and snow seldom lies, while the summers, though hot (maximum temperature is 99° F, or 37° C), are dry, and a fresh northeasterly wind often blows by day. The nights are cool. All of this permits outdoor activity the year around and has had an important effect on both the style of architecture and the life and political institutions of the city.

The city plan. In 1833 there was almost no Athens at all. During the fight for independence, it had been entirely evacuated in 1827, and six years later held perhaps 4,000 people in the straggle of little houses on the north slope below the Acropolis. The newly imported king of the Hellenes, Otho, the 18-year-old son of Ludwig I of Bavaria, was installed in the only two-story stone house, while his German architects hurried ahead with plans for a palace and a new Athens far out in the fields.

Below the well-sited but very plain palace, a large garden square, Síntagma (Constitution) Square, was laid out. Today it is garnished in the tourist season with some of Europe's most luxurious cafe chairs, and at all seasons it is hemmed in by tall new buildings and elderly luxury hotels. Broad avenues were created and are still the city centre's principal thoroughfares (Stadíou and Panepistimíou [Elefthérios Venizélou]), between which an orderly grid of narrow side streets was laid out. The housing that developed was generally the sort of architecture familiar in Victorian London: solid, porched, rather imposing, the later imitations graceless and monotonous. In Athens it is called the Othonian style, but there is little of it left as the centre encroaches on old residential areas.

Growth of the new capital

Once the new capital was established, the city grew at a regular rate of about 7 percent a year, soon reaching 50,000 inhabitants, a figure not much exceeded in the days of Athens' greatest power and glory. By 1907 the municipality had a population of 167,479. Omónia (Harmony) Square had been built at the western end of the two main streets, with other broad avenues radiating from it, but it did not develop as the hoped-for balance to Síntagma.

By then the railway to Piraeus had been built, its station near the antique Agora. Indeed, the city plan projected a logical growth southward along this axis, but a real-estate developer beckoned northward—the National Archaeological Museum is out this way—and the newly rich followed. The palace garden almost touched the Arch of Hadrian and the 15 mammoth columns (some of them seven feet 10 inches in diameter) of the temple of Olympian Zeus, last of the Classical buildings built in Athens, and beyond lay empty fields. The slopes of Mt. Likavittós, outside the town limits, were still pine-clad.

Since then, the garden has become one of the painfully rare public parks in Athens. Likavittós now rears up in the middle of the city (as if Hyde Park or Central Park were a 1,112-foot mountain), its lower slopes built upon, and many of the trees felled for a road leading to a cog railway and restaurant.

Along Panepistimíou Street rose the Academy of Athens, in marble from Mt. Pentelicus, its pediments and colonnades gilded. Its new neighbours were the University of Athens (refounded in 1837), the colonnade adorned with paintings, and the National Library. All were done in Greek Revival style by the court's German architects. A new Royal Palace (now the Presidential Residence) was built during 1891–97, a little southeast of the old (which is now a Parliament house) on Herodes Atticus Street. This leads to the 70,000-seat Panathenaic (Athens) Stadium, reconstructed by an expatriate Greek millionaire in time for the revival of the Olympic Games in 1896.

In 1921 the orderly progress of Athens was overturned and haphazard development began, for ethnic minorities were exchanged between Greece and Turkey, and approximately 1,500,000 Greeks, most of them penniless, came home from Asia Minor. Despite government efforts to resettle them elsewhere, many swarmed into shantytowns around the fringes of Athens and Piraeus, and the area's population soared from 473,000 to 718,000. After that, the city began to spread in two directions, south toward Piraeus and north toward the village of Kifisiá, which first became a smart suburb when Herodes Atticus built his villa there in the 1st century BC.

In the 1940s hideous things happened in Athens. During the German occupation many people died from starvation, and the city began to fall apart from lack of maintenance. When the Germans left, part of the Allied-equipped resistance refused to lay down its arms, and the civil war began. For a while the government held only the Parliament building, neighbouring embassies, and a part of Síntagma Square, while the palace garden was used as a common grave.

Housing. A construction boom began in the 1950s. New apartment houses pushing up everywhere erased old social boundaries (though the Kolonáki district on the southeast slope of Likavittós remained an enclave of respectable fortunes), and villages that had been attached to the city in the previous expansion lost their physical and political identities. A network of major highways was thrown up. The west side of the historic olive grove by the Kifisós River was shorn, and hillside greenery began to disappear under housing, either unauthorized or made legal through political skulduggery. Open space vanished, without provision for parks, playgrounds, or even schools, and Athens spread down to the sea by Glifádha, joining up with Piraeus. Piraeus itself was transformed from one of the world's celebrated honky-tonk ports into a clean, newly built, flower-decorated city.

Enlargement of the city's master plan

The Athens master plan was enlarged several times to keep pace with spread, which by 1964 had already attained 75 square miles, with a built-up area of 17 square miles outside the plan altogether. Land values in the centre quadrupled, then octupled, and rose proportionately elsewhere. Traffic increased almost to the saturation point at rush hours, and the city continued to sprawl beyond its planned limits. As international tourism increased, Ellinikón Airport, south of the city, was expanded and modernized.

The city water supply from an artificial lake at Marathon was insufficient to supply the new building construction, and the Mórnos River 110 miles to the northwest was dammed and tapped. Installation of a modern sewer system was undertaken, together with controls to check the floods that roar into Athens when heavy rains pour off the denuded mountains.

Old Athens

Traditional features. The older Athens has not entirely disappeared in all this hubbub. Older men may have given up smoking hookahs in shadowy cafes but not their 33-bead *kombouloi* ("worry beads"), which were acquired from the Turks.

Old Athens occupies the six streets sidling off Monastiráki Square, by the excavated Agora. In this area are tiny open-fronted shops hung with tinselled folk costumes and all of the monuments of Athens reproduced in copper, plaster, plastic, and paint. There is an alley of antique dealers, a street of smithies, one of hardware merchants, and another of wildly assorted miscellany.

Close to this lively quarter is the Pláka, on the north slope of the Acropolis. Small, one-story houses, dating from about the time of independence, are clustered together up the hillside in peasant simplicity. There are appropriately tiny squares with tavernas, once celebrated for their folk music, dancing, and simple fare. There are vine-covered pergolas and some unpaved streets too narrow for cars. The baths built by the Turks still function morning and

afternoon, but the bouzouki, a local relative of the lute, is giving way to the electric guitar. The taverna signs are multilingual, and the ubiquitous kitchen chair is being replaced by the plastic-ribbed restaurant seat. Progress laps at the Pláka like a vengeful sea, but the Acropolis is just up above, just under the stars.

The Acropolis. Many of Athens' bequests (all, if the theatre of Herodes Atticus may be regarded as an embodiment of the city's literature) to the world are expressed in and around the Acropolis, the natural centre of Athens. Rising some 500 feet above sea level, with springs near the base and a single approach, the Acropolis was an obvious choice of citadel and sanctuary from earliest times. That it could be something more is evidenced in the Parthenon, one of the brightest jewels in mankind's, let alone Athens', treasury. As deceptively simple as Socrates' conversation, this columned, oblong temple is the expression—without a trace of strain or conflict—of a human ideal of clarity and unity. The architectural genius is concentrated in the exterior, for within was a shelter for the goddess Athena—the patroness who lent her name to the city—not a place for mass worship. Its spiritual quality, the sensation of being almost afloat, is enhanced by the lack of a single, straight, vertical line in the peristyle (the surrounding colonnade); each vertical is almost imperceptibly bowed, theoretically meeting some 11,500 feet in the sky. The columns, of diminishing thickness toward the centre of the colonnade, with diminishing space between them, lean toward the centre, too; all these differences are virtually invisible to the beholder. Even the 20 flutings of each column diminish in width as they rise, and the humblest details of craftsmanship are perfect.

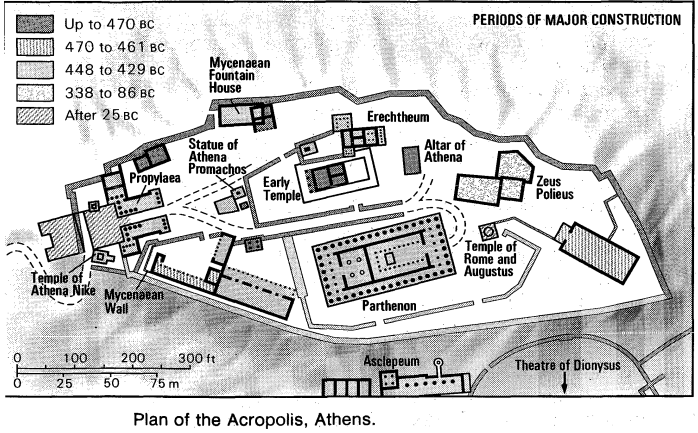
The Parthenon

On the northeast corner of the interior are faint traces of Christian wall paintings dating from the temple's service as the Church of St. Mary, and in the southeast corner of the porch is the stair leading to the minaret that emerged through the roof when the building was a Turkish mosque. The Parthenon was also used as a powder magazine, when, on September 26, 1687, Venetian artillery, attacking the Turks from the Hill of the Muses, scored a direct hit. A member of the party with the field commander, General Königsmark, wrote, "How it dismayed His Excellency to destroy the beautiful temple which had existed three thousand years!" Conversely, Francesco Morosini, the commander in chief, when reporting to the Venetian government, called it "a fortunate shot." Wishing to bring home more than just good news, he also tried to lower Athena's horses in the centre of the west pediment, but his men's dexterity was not as highly developed as their marksmanship and the masterpieces smashed to bits on the rock below.

The Turks regained possession of the Acropolis the following year and later began selling souvenirs to Europeans. The Duc de Choiseul, formerly French ambassador in Constantinople, picked up a piece of the frieze and two metopes. In 1801 the British ambassador, Lord Elgin, arrived with an imperial decree permitting him to pull down Turkish houses on the Acropolis to seek fragments of sculpture. Among the 50 pieces he took home (the shipping charges were £75,000, a huge sum for those days) was most of the remaining Parthenon sculpture, which he later sold to the British Museum for £35,000. The Greeks have forgiven the clumsiness of the Venetian engineers, the accuracy of Venetian cannoneers, and the vandalism of the Turks, but they still nurture rancour against Elgin. He also removed one of the caryatids from the Erechtheum, a temple of Athena called after a shrine dedicated to the legendary king Erechtheus or to Poseidon Erechtheus, but replaced her with a plaster cast. From London he sent a town clock for Athens, duly erected in the Agora and lost in the fire of 1885.

Lord Elgin

Adapted from *Westermann Grosser Atlas zur Weltgeschichte*: Georg Westermann Verlag, Braunschweig



Plan of the Acropolis, Athens.

The Erechtheum (5th century BC), which later became a church under the Byzantines and subsequently the Turkish commander's harem, was originally dedicated to both Athena and Poseidon and was the most venerated of the Acropolis temples.

The Propylaea, the matchless entryway into the Acropolis, was the only opening in the surrounding wall. Just in front of it and to the left is the 27-foot-high pedestal for the thank offering to Agrippa, the victor of the Battle of Actium, who interceded for Athens, which had supported the loser, Mark Antony. To the right was the temple of Athena Nike (Giver of Victory), 27 feet long and 18½ feet wide, which stood untouched until the Turks demolished it in 1686 to use the stones as defenses against the Venetians. In 1836 it was badly restored, and 100 years later its foundations began to slip into previously undiscovered Turkish cisterns, revealing the 6th-century foundations of Peisistratus' earlier shrine to Artemis Epipyrgidea (Artemis on the Tower). The temple was then more accurately reconstructed. The northern wing of the Propylaea, the Pinakotheke, was used by the Frankish dukes, who reconstructed the interior to make a two-story building. In the 12th century, Greek Orthodox bishops lived in the Pinakotheke, and in the 14th century, the dynasty of Athenian dukes from Florence turned the Propylaea into a fortified castle with a Tuscan tower, which Heinrich Schliemann, the German excavator who discovered Troy, paid to have dismantled in 1875.

The Pinakotheke

When the Turks, who had occupied Athens since 1456, departed, they left the monuments in a state of ruin, the ground covered with garden plots, and several hundred small huts. After Greece won its independence, Otho, the first king of the Hellenes, had everything that postdated the classical period swept away, set scholars to work identifying the remains, and encouraged some reconstruction.

According to Pausanias, the Greek traveller and geographer of the 2nd century AD, the colossal 30-foot-high bronze seated statue of Athena Promachos (Athena Who Fights in the Foremost Ranks), by the 5th-century BC Athenian sculptor Phidias, was set up in the open behind the Propylaea, her gleaming helmet and spear visible to mariners off Cape Sunium (Soúnion) 30 miles away. The 6th-century Byzantine emperor Justinian carried the statue off to Constantinople (now Istanbul), just as Phidias' ivory and gold statue of Athena had been taken from the Parthenon. Both of these masterpieces were lost to other looters in the crusaders' sack of Constantinople in 1204. Other statues stood in profusion amid small temples, such as the sculptor Myron's group of Marsyas and Athena, his Perseus, and his heifer; Phidias' Lemnian Athena and his Pericles; and a gigantic bronze effigy of the Trojan horse. There was an altar to Athena Hygeia (the Health Giver), a precinct sacred to the goddess Artemis Brauronia (named after a statue of her, brought from the town of Brauron), the Pandroseum (a building named after Pandrosos, a girl associated with Athena in legend), where the sacred olive tree of Athena grew, and beyond the Parthenon the great altar of Athena.

Other notable buildings. Below the Acropolis sanctuary, on the southwest slope of the hill, Herodes Atticus, a rich Roman, built a 5,000-seat odeum as a memorial to his wife in AD 161. A conventional Roman theatre, except that the semicircular auditorium was hollowed out of the rock, it was roofed in cedar and had a three-story facade of arches. Repaired but roofless, it is now used for the Athens summer festival of music and drama. A 300-yard-long portico stretching toward the Theatre of Diony-

The theatre of Herodes Atticus

sus had been built some 300 years earlier. The Dionysiac theatre itself, scooped out of the south slope early in the 5th century, replaced the Agora stage as the drama centre. It also replaced the Pnyx as the meeting place for the popular assembly. Rebuilt many times, the ruined theatre now visible is largely Roman, the last construction work on the stage probably dating from the early 3rd century AD. The Dionysia, the spring festival, which drew crowds from many parts of Greece and colonies in Asia Minor and Italy, was held in this theatre, which had 13,000 seats in 67 rows. The jury had larger front seats and the ecclesiastical dignitaries small stone thrones, on which their titles can still be read. Three tragic and four comic plays were presented in competition for the prize. Production costs were met by private sponsors who, when their choruses won the prize tripod, displayed it in an elaborate memorial in the Street of Tripods to the east of the theatre. The only one of these monuments still standing is that of Lysicrates, erected 334 BC, a small circular temple 21½ feet high, its six columns an early example of the Corinthian order. The monument was preserved through incorporation into a convent (in which the English poet Lord Byron had a study) and influenced British Georgian and Regency architecture through the engravings of the Edinburgh artist "Athenian" Stuart. Farther east lay the Odeum of Pericles, and to the west are traces (420 BC) of the precinct of Asclepius, the god of healing, which took the form of a hospital portico for patients and temples decorated with votive reliefs.

On the Hill of Ares, the god of war, to the right of the descent from the Propylaea, a legendary jury of gods spared Ares from execution for the murder of the sea god Poseidon's son. Trials for homicide continued to be heard on this hill through the ages, and the Supreme Court of Greece still bears the name.

Other hills in Athens

Across Apostólou Pávlou (Apostle Paul Avenue) are the Hill of the Nymphs, where an Austro-Greek, Baron Sina, built an observatory in 1842; the Hill of the Muses, crowned with the remains of the marble monument to Philopappus, a Syrian who was Roman consul in the 2nd century AD; and the middle hill, the Pnyx (Tightly Crowded Together), the meeting place of the Ecclesia, the assembly of 18,000 citizens who heard the great Athenian orators. (In fact, attendance of more than 5,000 persons was rare at any gathering, but the Pnyx would still have been crowded.)

The Agora. The avenue leads down to the Agora, which the American School of Classical Studies started restoring in 1931, paying $2,500,000 compensation to the several hundred families living there. Financed by, among others, the Rockefeller Foundation, the Marshall Plan, and the Greek government, the work went on until 1960. It includes what has been called "the pitiless replica of a 180-columned portico of the 2nd century BC," which serves as a museum.

The Theseum

At the approaches to the Agora is the best preserved of all Greek temples, the Theseum (5th century BC). Although virtually intact and absolutely genuine, it has all the deadness of a latter-day reproduction. The beauty, the mystery, and the genius that render the Parthenon incandescent eluded the architects and builders of the Theseum.

The Horologium and the Orthodox cathedrals. Another monument is the octagonal, 42-foot-high marble Horologium of Andronicus of Cyrrhus, usually called the Tower of the Winds because each side bears a weather-beaten figure of the wind from that particular compass point. It used to have a sundial, a water clock for telling the hour on cloudy days, and a weather vane. The Turks left it unchanged, believing it to be the tomb of two local prophets, Sakhratis and Aflatun (Socrates and Plato).

In the shadow of the 19th-century, neo-Byzantine Greek Orthodox Cathedral (Mitrópolis) nestles the 12th-century Mitrópolis, Áyios Elefthérios, one of three genuine Byzantine churches still surviving. It is red brick, like the others, and tiny. Its Pentelic marble is ruddied with age, and its outer walls are artfully, if promiscuously, decorated with classical Greek tidbits: panels, votive tablets, and morsels of frieze. Like its sisters, this retired cathedral is charming, unassuming, and comforting.

THE PEOPLE

The population of Greater Athens increased considerably after the war of independence in the early 1830s. The rapid growth was largely due to the great influx of refugees from Asia Minor in the early 1920s and the migration of rural inhabitants from the provinces during World War II and the Communist rebellion (1946–49). By the 1960s Athens had become a bustling cosmopolitan city. Almost all Greeks adhere to the Greek Orthodox faith.

THE ECONOMY

Industry and trade. Since World War I Athens has become the hub of all mercantile business, export and import. With Piraeus, it is the most important manufacturing city in Greece. Athens accounts for half of the jobs in industry and handicrafts, and earnings are much higher than the national average. There are cloth and cotton mills, distilleries, breweries, potteries, flour mills, soap factories, tanneries, chemical works, and carpet factories. Exports include olive oil, tomato products, wine, cement, bauxite, and textile manufactures. Publishing enterprises are important.

The brilliant Attic light, however, is now dimmed by the pall of pollution hovering over the city. To discourage new factories from adding to the problem and to stimulate the economic growth of other regions, an industrial wage tax has been imposed in the Athens area, and tax incentives have been offered to new factories set up in other areas.

Transportation and shipping. Athens accounts for more than half of the nation's cars, trucks, and buses. Furthermore, the number of merchant ships registered in Greece (mostly at Piraeus, the country's largest port) has increased as Greek shipowners, since the late 1960s, have answered the government's call to bring their foreign-registered ships home (though many Greek ships remain under other flags). Scores of shipping offices have opened in refurbished Piraeus, while on weekends shipping magnates sail to the nearby islands of Hydra and Spetse in chrome-fitted luxury yachts, flying Panamanian and Liberian flags.

The metropolitan transit system includes an electrified rail line, buses, and trolleys. The electric railway, which connects Piraeus in the south with the suburb of Kifisiá in the north, runs underground within Athens proper. Larissa, the main railway station, links the city with the rest of the country and the continent. (B.E./Ed.)

History

THE EARLY PERIOD

Factors inducing settlement. The site of Athens has been inhabited since the Neolithic Period (before 3000 BC). Evidence for this has come from pottery finds on and around the Acropolis but particularly from a group of about 20 shallow wells, or pits, on the northwest slope of the Acropolis, just below the Klepsydra spring. These wells contained burnished pots of excellent quality, which show that even at this remote period Athens had a settled population, with high technical and artistic standards. There are similar indications of occupation in the Early and Middle Bronze ages (3000–1500 BC).

Evidence of early inhabitation

The earliest buildings date from the Late Bronze Age, particularly about 1200 BC when the Acropolis was the citadel. Around its top was built a massive wall of cyclopean masonry (a type of construction using huge blocks without mortar). The construction of this wall probably marks the union of the 12 towns of Attica (the department in which Athens lies) under the leadership of Athens, an event traditionally ascribed to Theseus. The palace of the king was in the area of the later Erechtheum, but almost no traces of it have been identified. The town, insofar as it was outside the Acropolis, lay to the south, where wells and slight remains of houses have been found. The principal cemetery lay to the northwest, and several richly furnished chamber tombs and many smaller ones have been discovered in the area that later became the Agora.

Whether through the strength of its walls, the valour of its citizens, or its geographical position away from the main route to the Peloponnesus, Athens seems to have weathered the Late Bronze and Early Iron ages, troubled

times, better than other, more important centres. There is no evidence of complete or widespread destruction, as at Mycenae and Pylos; in fact, the pottery styles show an unbroken development through sub-Mycenaean (later than the Mycenaean but not yet Greek) to Protogeometric (the earliest phase of Geometric) and Geometric Period (1000 BC to about 750 BC). Furthermore, there is positive evidence that from about 1000 BC the city began to expand in a northwesterly direction, into the area that had previously been confined to cemeteries. Wells appear, indicating occupation by the living, and any graves in the area are increasingly confined to restricted plots or placed along the roads outside the town limits. The Agora and some of the public buildings seem, to judge from scattered notices in later writers, to have been located west and northwest of the Acropolis. Though there are few remains of buildings, the wealth and prosperity of the city can be appreciated from late Geometric graves found in the area of the later Dipylon and Erian gates. These graves were adorned with large vases, sometimes more than five feet high, decorated with geometric patterns and with scenes of battles, processions, and funeral ceremonies.

Changes in the 6th century BC

Athens' expansion. The 6th century BC was a period of phenomenal growth, particularly during the tyranny of Peisistratus and his sons (*c.* 560–510 BC). On the Acropolis, the old primitive shrines began to be replaced with large stone temples. About 580 BC a temple to Athena, known as the Hecatompedon (Hundred-Footer), was erected on the site later to be occupied by the Parthenon. The pediments (triangular spaces forming the gable) of this temple were decorated with large-scale sculpture in gaily coloured, porous limestone, representing groups of lions bringing down bulls, and with snaky-tailed monsters in the angles. These sculptures are now displayed in the Acropolis Museum. In 566 BC Peisistratus reorganized the Panathenaic Games in honour of Athena on a four yearly basis. About 530 BC a large peripteral temple (one having a row of columns on all sides) to Athena Polias (Guardian of the City) was erected near the centre of the Acropolis, on the site of the old Bronze Age palace. It had marble pedimental sculpture representing the battle of the gods and giants. Besides these two major temples there were five smaller buildings, treasuries and the like, and a wealth of votive offerings in marble, bronze, and terra-cotta. The Acropolis thus became a full-fledged sanctuary.

This change from citadel to sanctuary is also reflected in the arrangement of the entrance at the west. Instead of a winding path suitable for defense, there was, from about the middle of the 6th century BC, a broad ramp, designed as a ceremonial approach, leading up to the gate. This basic change of attitude toward the Acropolis must mean that the whole lower town was surrounded by a fortification wall and the Acropolis was no longer needed for defense. The ancient historians Herodotus and Thucydides tell of such a wall, but no trace of it has been found, and its course and date are uncertain.

Building of the new Agora

In the lower town, too, the 6th century was a period of growth and change. The old Agora, below the western approach to the Acropolis, was now inadequate, and a new one was therefore laid out in the low ground to the northwest. This was accomplished by demolishing houses and filling in wells and gullies, to create a broad, open square, which was used for gatherings of all sorts: political, judicial, religious, and commercial. Dramatic contests were held there, too, before the construction of a separate theatre. Various public buildings and shrines were erected around the borders of the square, including the Basileios (Royal) Stoa, where the archon Basileus, one of the chief magistrates of the city, had his headquarters; the Old Bouleuterion (or Council House); and a large enclosure (100 square feet) that probably housed the Heliaia, the largest of the popular lawcourts. At the southeast corner of the square a fountain house received water from outside the city through a conduit of terra-cotta pipes.

In 480 BC this flourishing city was captured and destroyed by the Persians. The Acropolis buildings were burned and the houses in the lower town mostly destroyed, except for a few that had been spared to house the Persian leaders.

Athens at its zenith. When the Athenians returned, in 479 BC, they immediately rebuilt their fortification wall larger than before. About 20 years later the famous Long Walls were built, connecting the city with its port, Piraeus, four miles away. They were parallel over most of their course, forming a corridor 550 feet wide. These walls played a vital part in the history of Athens during the Classical period, for they allowed it to carry the supplies brought in by its powerful fleet in safety to the city, even when enemy forces roamed the Attic countryside.

For 30 years after the Persian destruction, the Athenians built only fortifications and some secular buildings in the Agora, notably the Stoa Poikile, or Painted Colonnade, with its famous paintings by Polygnotus and Micon, one of which represented the Battle of Marathon. The Tholos, the round building that served as the headquarters of the executive committee of the council, was also built at this time. Lack of attention to the Acropolis was partly the result of the oath, sworn before the Battle of Plataea in 479 BC, that sanctuaries destroyed by the barbarians would

E. Streichan—Shostal Assoc.

Modern Athens and the Acropolis.

not be rebuilt but left as memorials of their impiety. In 449 BC, however, peace with Persia was at last officially established, and the oath was annulled. Athens, moreover, had ample funds, for the silver mines in the Laurium (Lavrion) Hills of southern Attica were in full production. These mines had always been exploited, but in 483 BC a big strike was made, the proceeds of which were used to build the ships that won the Battle of Salamis in 480 BC. Thereafter, the mines remained productive throughout the 5th and 4th centuries, providing Athens with the sinews of its strength in the great Classical age. Another source of revenue was the tribute that the allies had been paying, as members of the Delian League, to prosecute the war against Persia. Athens had been collecting and administering this money and, even though the war was officially over, continued to collect it in spite of the protests of the allies, who degenerated into subjects of Athens. Pericles deemed it proper, over the protests of his opponents, to use this money on beautifying the city; in this way he could keep the money in circulation and provide jobs for the whole population. Thus began one of the largest and most enduring works programs in history.

Peace with Persia

In a period of 40 years the Acropolis was entirely rebuilt in gleaming white marble quarried from Mt. Pentelicus, 10 miles north of the city. The first great work was the Parthenon, begun in 447 BC and finished, except for some details, in 438 BC. The architects were Ictinus and Callicrates, and Phidias was in charge of the whole artistic program. The building was considerably larger than was usual, having eight columns across the ends and 17 on the long sides, against six by 13 for the average temple. It was richly decorated with sculpture, having a running frieze all around the top of the cella (the walled-in chamber within the colonnade) wall outside, and sculptured metopes and sculptured pediments. Inside the cella stood the cult statue, the great gold and ivory figure of Athena, the work of Phidias. No sooner was the main work on the Parthenon completed than the Propylaea was begun. This was the monumental gateway with five doors at the head of the approach, designed by the architect Mnesicles. Its large outer vestibule was covered by a marble ceiling, supported by marble beams with a free span of 18 feet, about which Pausanias wrote, "The Propylaea has a ceiling of white marble which in the beauty and size of the stones remains supreme even to my time." Work on the Propylaea was nearly finished when it was stopped by the outbreak of the Peloponnesian War in 432 BC, but, as things began to go well for Athens, the little temple of Athena Nike was erected on the bastion in front of the Propylaea, perhaps in 425 BC. Around the time of the Peace of Nicias (421 BC) the Erechtheum was begun. This was a small Ionic temple of highly irregular plan, which housed various early cults and sacred tokens. When the building was about half-finished, work was suddenly interrupted, probably because of the disastrous Athenian expedition to Sicily (415–413 BC), but it was resumed in 409, and the building was completed in 406. The final defeat of Athens two years later put an end to all building, but the Acropolis had been completed, and in later centuries only secondary buildings and monuments were added.

The building of the Erechtheum

In the second half of the 5th century there was also some building activity in the lower town. Even before the Parthenon, work was begun on the temple of Hephaestus (the god of fire), the Theseum, which still stands on a low hill. In the Agora itself, a new Bouleuterion was built, and two colonnades, the Stoa of Zeus and the South Stoa, were constructed. On the south slope of the Acropolis, next to the theatre, Pericles built an odeum, a large enclosed concert hall, its roof supported by a forest of columns. Of the theatre itself there are no identifiable remains, but the arrangements were no doubt quite simple, and it is known that a theatre existed on this spot from the late 6th century BC because of the old temple of Dionysus (the god of wine) nearby, which dates from the same period. A sanctuary of Asclepius was founded on the south slope of the Acropolis in 420 BC.

Athens was slow in recovering from its defeat in the Peloponnesian War, but in 394 BC its admiral, Conon, won a decisive naval victory over Sparta off Cnidus, on the west coast of Asia Minor. As a result he rebuilt the Long Walls, which the Spartans had demolished to the music of flutes 10 years before, believing they were inaugurating the freedom of Greece. The walls of Piraeus were also rebuilt, and those of the city were repeatedly strengthened in the course of the 4th century, notably by the addition of a ditch, or moat, as protection against siege machinery.

Apart from military works, there was little building in 4th-century Athens until the years 338–322 BC, when the orator Lycurgus was in control of the state finances and there was great activity. On the Pnyx, the broad-backed hill west of the Acropolis where the Athenian popular assembly had met since the reforms of Cleisthenes in the 6th century, a large auditorium was constructed. At the same time, two large stoas were started on the terrace above. The Theatre of Dionysus was rebuilt and greatly enlarged, with stone seats to accommodate the crowds. (Lycurgus did another service to the theatre by having definitive copies made of the old plays.) The Panathenaic stadium was also built about then, partly with state funds and partly by private contributions; the land was donated by a certain Deinias, and one Eudemus of Plataea provided 1,000 yoke of draft animals to level the ground. The period was one of lavish private expenditure in other fields as well. The tripods won in choral contests were displayed on elaborate monuments, sometimes even resembling small temples; the best preserved of these is that of Lysicrates (334 BC), a small round building with six Corinthian columns. Tombs also became increasingly elaborate, often portraying the whole family in high relief. In 315 BC a stop was put to all this extravagance by the sumptuary laws of Demetrius of Phalerum.

Building in the late 4th century

Meanwhile, the philosophy schools flourished. Plato (*c.* 428–348/347 BC) established himself in the Academy, a gymnasium that had existed since at least the 6th century BC in the great olive grove about a mile west of the city. Plato himself had a house and garden nearby. Aristotle and his Peripatetics occupied the Lyceum, another gymnasium, just outside the city to the east, and his successor Theophrastus lived nearby. Antisthenes and the Cynics used the Cynosarges gymnasium to the southeast of the city. Zeno held forth in the heart of the city, in the Stoa Poikile, in the Agora, and his followers were therefore known as Stoics. Epicurus and his followers had a house and garden in town.

Apart from its temples and public buildings and its great avenues, however, Athens seems to have made a poor impression. A 3rd-century-BC visitor complained that the city was dry and ill-supplied with water, that it was badly laid out because of its great antiquity, and that most of the houses were mean. The streets were in fact narrow and winding, and the houses, it is true, presented a blank wall to the street except for the entrance door, but then they were built around a central courtyard, off which the various rooms opened. There was often an upper story, and the court had a well. Water brought in by the aqueducts was not considered good because it was hard (containing salts of magnesium or calcium) and caused rheumatism. Waste water was carried off in an elaborate system of underground drains beneath the streets.

Hellenistic and Roman times. Athens in Hellenistic and Roman times depended for its embellishment less on its own resources than on the generosity of foreign princes. One of the Ptolemies (rulers of Egypt) gave a gymnasium, erected near the sanctuary of Theseus, and the Ptolemies were probably also instrumental in the founding of the sanctuary of the Egyptian gods Isis and Sarapis. More important were the donations of the Attalids of Pergamum (a dynasty of Asia Minor); Eumenes II (197–159 BC) gave a large, two-story colonnade on the south slope of the Acropolis near the theatre. His brother Attalus II (159–138 BC), who had studied at Athens under the philosopher Carneades, head of the New Academy, likewise gave a colonnade. This was a large, elaborate, two-story building more than 350 feet long with a row of shops at the rear. It was located on the eastern side of the Agora and has been reconstructed in modern times (1953–56) to serve as the museum of the Agora excavations. The Stoa of Attalus was the first element in a large-scale reconstruction of the

Donations of the Attalids

Agora. It was followed in quick succession by three buildings, the Middle Stoa, the East Building, and the South Stoa, which together formed a separate South Square.

The capture of Athens by the Roman general Sulla in 86 BC was accompanied by great slaughter and much destruction of private houses, but the only public building to be destroyed was the Odeum of Pericles, burned by the defenders lest its timbers be used by the enemy. The odeum was rebuilt a few years later, through the generosity of King Ariobarzanes of Cappadocia.

Under the Roman Empire, Athens enjoyed imperial favour. A spacious market for the sale of oil and other commodities was laid out east of the old Agora with funds originally provided by Julius Caesar and supplemented by the emperor Augustus. In the old Agora itself, a new odeum, or concert hall, was built in the middle of the square by Marcus Agrippa, the emperor's son-in-law and one of his chief lieutenants. A large building, perhaps a lawcourt, was also erected at the northeast corner. At the southeast corner of the Agora a handsome library was erected about AD 100, the gift of one T. Flavius Pantainus and his family. It was decorated with a group of marble sculpture representing Homer flanked by the *Iliad* and the *Odyssey*. On the Acropolis a small round temple was erected to the goddess Roma and the emperor Augustus.

The emperor Hadrian (AD 117–138) completed the great temple of Olympian Zeus, started more than 600 years earlier by the Peisistratids. This temple formed the chief ornament of the new eastern suburb of Athens, and Hadrian gave the area a monumental entrance through a gateway, the inscriptions on which proclaimed, on one side, "This is the Athens of Theseus, the old city," and, on the other, "This is the city of Hadrian, not of Theseus." Hadrian also built a library, a gymnasium, and a pantheon (a sanctuary of all the gods). His aqueduct, which brought water from the mountains to the north, has been reconditioned and still serves the modern city.

Hadrian's devotion to Athens

In the reign of Valerian (AD 253–260), the walls of Athens, which had been neglected since Sulla's capture of the city in 86 BC and had fallen into ruin, were rebuilt, and the circuit was extended to include the new suburb northeast of the Olympieion. This was done because of the threat of a barbarian invasion, but when that invasion came, in AD 267, the walls were of no avail. The Heruli, a Germanic people from northern Europe, easily captured Athens, and though the historian P. Herennius Dexippus rallied 2,000 men on the city outskirts, they could only resort to guerrilla tactics. The lower town was sacked, and all the buildings of the Agora were burned and destroyed. The Acropolis, however, may have held out; at least there is no evidence of extensive damage at this time.

This sack of Athens is comparable only to that by the Persians in 480 BC, but now the reaction was quite different. The Athenians abandoned the outer circuit and established a new and much smaller line north of the Acropolis, leaving even the Agora area outside the walls. This new wall, which, on the evidence of coins, was built in the reign of Probus (276–282), consisted of material taken from ruined buildings in the lower town.

Athens remained confined within this narrow circuit for several generations, but in the 4th and 5th centuries it experienced a revival. The old outer circuit of the walls was restored, and many new buildings were erected. Athens at this time was still the cultural capital of the Greek world and a stronghold of paganism. Its schools of philosophy, which retained their ancient names, however different their outlooks may have been, flourished, attracting students from all parts. These included the emperor Julian the Apostate and two Fathers of the Church, Basil and Gregory of Nazianzus. While the schools existed, Athens remained a place of consequence, but when they were closed by the emperor Justinian in AD 529, Athens sank to the level of a small provincial town. Power and wealth had long since moved to Constantinople, the new centre of the Greek world.

THE BYZANTINE AND TURKISH PERIODS

Christianity started early in Athens, with the visit of the Apostle Paul in AD 51 and the conversion of Dionysius the Areopagite, a former archon and member of the Court of the Areopagus that had heard Paul's defense of his teachings. The little Christian community did not flourish, however, and Athens remained a stronghold of older ways. In the 5th and 6th centuries, however, after the formal establishment of Christianity and the abolition of pagan worship, churches began to be built. These were sometimes ancient temples converted to Christian worship; for example, the Parthenon, the Erechtheum, and the temple of Hephaestus (the Theseum). Newly built churches had a basilica plan and a wooden roof, but these now survive only in foundations. In all, some 22 churches of this period are known.

Effects of Christianity on Athens

The 7th to 10th centuries were dark times for Athens. The city is almost never mentioned in the history of the period, and archaeological remains are few. In the 11th and 12th centuries a measure of prosperity returned, and the taste of Athenians then can be gauged by the number of small stone and brick churches surviving, built on the Byzantine cross-in-square plan, such as the Kapnikaréa, and those of St. Theodore and the Holy Apostles.

Athens fell to the crusaders in 1204, remaining in Latin hands for 250 years. The town's outward appearance changed little, except that the Parthenon, now a Roman Catholic not an Orthodox cathedral, received a bell tower.

When the Turks captured Athens in 1456, the Parthenon became a mosque, and its bell tower was turned into a minaret. Other mosques were built in the lower town, but in general the age of gunpowder was to prove disastrous for Athenian architecture, especially on the Acropolis, which was still virtually intact as late as the mid-17th century.

ATHENS AFTER GREEK INDEPENDENCE

Greek insurgents surprised the city in 1821 and captured the Acropolis in 1822; but in 1826 Athens again fell into the hands of the Turks, who bombarded and took the Acropolis in the following year (the Erechtheum suffered greatly, and the monument of Thrasyllus was destroyed). The Turks remained in possession of the Acropolis until 1833, when Athens was chosen as the capital of the new kingdom of Greece. Its subsequent history is that of the kingdom.

In World War I Athens was the scene of the incidents of 1916–17 that led to the deposition of King Constantine by the Allies. It was occupied by German troops during World War II, but the city was spared aerial bombardment.

(E.V./Ed.)

BIBLIOGRAPHY

History and antiquities: PAUSANIAS, *Description of Greece*, Book I, a description of Athens by the traveller Pausanias (2nd century AD) that contains much of interest; among the best English translations of this work are the volume, with brief commentary, by PETER LEVI, *Guide to Greece*, 2 vol. (1971); and the classic translation, with commentary, by J.G. FRAZER, *Pausanias' Description of Greece*, 6 vol. (1898). Other works include T.B.L. WEBSTER, *Everyday Life in Classical Athens* (1969), the Athenian at home and in public, *Art and Literature in Fourth Century Athens* (1956), the cultural life of the city when it was the intellectual capital of the world, and *Athenian Culture and Society* (1973), an overview for the general reader; ERIKA SIMON, *Festivals of Attica: An Archaeological Commentary* (1983), an important study of origins; RICHARD E. WYCHERLEY, *The Stones of Athens* (1978), a survey of the architecture; ANGELO PROCOPIOU, *Athens, City of the Gods: From Prehistory to 338 B.C.* (1964), richly illustrated; GERHART RODENWALDT, *Acropolis* (1957; 5th German ed., 1956); and HOMER A. THOMPSON and R.E. WYCHERLEY, *The Agora of Athens* (1972), two detailed and learned expositions of classical Athens' important sites, with many illustrations; SUSAN I. ROTROFF, *Hellenistic Pottery: Athenian and Imported Moldmade Bowls* (1982); MARTIN HÜRLIMANN, *Athens* (1956; German ed., 1956), chiefly photographic, with introductory text by REX WARNER and detailed historical notes accompanying the pictures.

For the specialist: A.W. PICKARD-CAMBRIDGE, *Theatre of Dionysus in Athens* (1946, reissued 1973), a thorough study of the theatre and its various uses; HUMFRY PAYNE and GERARD YOUNG, *Archaic Marble Sculpture from the Acropolis* (1950), a scholarly photographic catalog chiefly for the archaeologist and art historian; JON D. MIKALSON, *Athenian Popular Religion* (1983), an argument and theory based solely on forensic evidence.

(B.E./E.V./Ed.)

Atmosphere

The atmosphere that surrounds the Earth consists of a mixture of gases, primarily nitrogen and oxygen. This gaseous envelope, commonly called the air, also contains numerous less abundant gases, water vapour, and minute solid and liquid particles in suspension. Rocket probes and especially the drag encountered by artificial satellites at altitudes of several thousand kilometres have demonstrated that the terrestrial atmosphere extends to a very great distance.

The composition of the atmosphere encodes a great deal of information bearing on its origin. Furthermore, the nature and variations of the minor components reveal extensive interactions between the atmosphere, terrestrial environment, and biota. The development of the atmosphere and such interactions are discussed in the first major section of this article, with particular attention given to the rise of biologically produced molecular oxygen, O_2, as a major component of air.

The atmosphere is considered in terms of layers, or regions, arranged like spherical shells above the surface of the Earth. The chemical and physical properties of these various regions are treated in considerable detail. Such upper atmospheric phenomena as airglow, auroras, and the Van Allen radiation belts are included in the coverage.

For coverage of related topics in the *Macropædia* and *Micropædia,* see the *Propædia,* sections 221 and 241.

This article is divided into the following sections:

Development of the Earth's atmosphere

A complete reconstruction of the origin and development of the atmosphere would include details of its size and composition at all times during the 4,500,000,000 years since the formation of the Earth. This goal could not be achieved without knowledge of the pathways and rates of supply and consumption of all atmospheric constituents at all times. Information regarding these processes, however, is incomplete even for the present atmosphere, and there is almost no direct evidence regarding atmospheric constituents and their rates of supply and consumption in the past.

The contrast with related fields of the Earth's history is notable. Fossils and other structural and chemical details of ancient rocks provide information useful to evolutionary biologists and historical geologists, but ancient atmospheres, "mere vapours," have not left such substantial remnants. These vapours are, however, the stuff of stars and the moving force of storms and erosion. Although historians of the atmosphere must rely heavily on inference, the object of their interest has played so important a role in the Earth's history that evidence related to its development, though indirect, is abundant.

CONCEPTS RELATED TO ATMOSPHERIC DEVELOPMENT

The atmosphere as part of the crust. To the Earth scientist, the crust includes not only the top layer of solid material (soil and rocks to a depth of 6–70 kilometres, separated from the underlying mantle by differences in density and by susceptibility to surficial geologic processes) but also the hydrosphere (oceans, surface waters on land, and groundwater beneath the land surface) and the atmosphere. Interactions among these solid, liquid, and gaseous portions of the crust are so frequent and thorough that considering them separately introduces more complexities than it eliminates. As a result, a description of the history of the atmosphere must concern itself with all volatile components of the crust.

Important volatile components

Materials. Volatile compounds as well as elements important in present and past atmospheres or in interactions between the atmosphere, biosphere, and other portions of the crust include the following:

1. Present major components: molecular nitrogen (N_2) and molecular oxygen (O_2)
2. Noble gases: helium (He), neon (Ne), argon (Ar), krypton (Kr), and xenon (Xe)
3. Abundant variable components: water vapour (H_2O) and carbon dioxide (CO_2)
4. Other components: molecular hydrogen (H_2), methane (CH_4), carbon monoxide (CO), ammonia (NH_3), nitrous oxide (N_2O), nitrogen dioxide (NO_2), hydrogen sulfide (H_2S), dimethyl sulfide [$(CH_3)_2S$], sulfur dioxide (SO_2), and hydrogen chloride (HCl).

Some elements appear in multiple form—for example, carbon as carbon dioxide, methane, or dimethyl sulfide. It is useful to consider the occurrence of the elements before focusing on the more specific aspects of atmospheric chemistry (the forms in which the elements are present). One can speak of the Earth's "inventory of volatiles," recognizing that the components of the inventory may be reorganized from time to time, but that it is always composed primarily of the compounds of hydrogen, carbon, nitrogen, and oxygen, along with the noble gases.

Differentiation of sources

Processes. A process that delivers a gas to the atmosphere is termed a source for the gas. Depending on the question under consideration, it can make sense to speak in terms of either an ultimate source, the process that delivered a component of the volatile inventory to the Earth, or an immediate source, the process that sustains the abundance of a component of the present atmosphere. Any process that removes gas either chemically, as in the consumption of oxygen during the process of combustion, or physically, as in the loss of hydrogen to space at the top of the atmosphere, is called a sink.

Throughout the history of the atmosphere, sources and sinks have often been simultaneously present. While one process consumes a particular component, another produces it, and the concentration of that component in the atmosphere will rise or fall depending on the relative strengths of the sources and sinks. If those strengths are balanced (or nearly so), the composition of the atmosphere will not change (or will change only very slowly, perhaps imperceptibly); however, the molecules of the gas in question are passing through the atmosphere and are not permanently resident. The rate of the resulting turnover of molecules in the atmosphere is expressed in terms of the residence time, the average time spent by a molecule in the atmosphere after it leaves a source and before it encounters a sink.

PROCESSES AFFECTING THE COMPOSITION OF THE EARLY ATMOSPHERE

Formation of the solar system from the primordial gas cloud

Ultimate sources. The material from which the solar system formed is often described as a gas cloud or, at a later stage, solar nebula. The cloud was rich in volatiles (termed primordial gases) and must have been the ultimate source of the atoms in the present atmosphere. What is of primary concern, however, is the sequence of events and processes by which the volatiles present in the initial gas cloud were transferred to the Earth's inventory and the efficiency with which this was accomplished.

The formation of the solar system began when one portion of the gas cloud became dense enough due to compression by some external force—a shock wave from the explosion of a nearby supernova, perhaps—to attract gravitationally the material around it. This material "fell" into the region of higher density, making it even denser and attracting other material from still farther away. As gravitational collapse continued, the centre of the cloud became very dense and hot, because the kinetic energy of the incoming material was released as heat. Thermonuclear reactions began at the core of the central object, the Sun.

Capture and retention of primordial gases. Far from the central point, the material in the gas cloud tended to settle to an extensive equatorial plane around the Sun. As the material in this disk cooled, chunks of rock grew and accreted to form the planets. The planets are much less massive than the Sun, but if they grew large enough and if the gases around them were cool enough, they could accumulate an atmosphere from the volatile components of the gas cloud. A partial inventory of that cosmo-chemical stockpile, the starting point for atmospheric development, is shown in the column for the solar system in Table 1. This direct capture is the first of three source mechanisms that can be described.

A planetary atmosphere accumulated in this way would consist of primordial gases, but the relative abundances of the individual components would differ from those in the gas cloud if the gravitational field of the new planet were strong enough to hold some, but not all, of the gases around it. It is convenient to express the strength of a gravitational field in terms of escape velocity, the speed at which any particle (a molecule or spacecraft) must be traveling in order to overcome the force of gravity. For the Earth this velocity is 11.3 kilometres per second (seven miles per second), and it follows that, once the solid material of the Earth had accumulated, gas molecules passing the planet at lower speeds would have been captured and accumulated to form an atmosphere.

The speed at which a gas molecule moves is proportional to $(T/M)^{1/2}$, where T is absolute temperature and M is molecular mass. As will be shown later, the uppermost layers of the present atmosphere are still very hot and might have been much hotter early in the Earth's history. At temperatures below 2,000 kelvins (K), however, molecules of any compound with a molecular weight greater than about 10 will have an average velocity of less than 11.3 kilometres per second. On this basis, it was long thought that the earliest atmosphere of the Earth must have been a mixture of the primordial gases with molecular weights greater than 10. Hydrogen and helium, with molecular weights of 2 and 4, should have been able to escape. Because hydrogen is the most abundant element in the solar system (see Table 1), it is thought that the most abundant forms of the other volatile elements were their compounds with hydrogen. If so, methane, ammonia, and water vapour, together with the noble gas neon, would have been the most abundant volatiles with molecular weights greater than 10 and, thus, the major constituents of the Earth's primordial atmosphere. The atmospheres of the four giant outer planets (Jupiter, Saturn, Uranus, and Neptune) are rich in such components, as well as in molecular hydrogen and, presumably, helium, which those more massive and colder bodies were apparently able to retain.

Possible major components of the Earth's initial atmosphere

Table 1: Abundances of Elements

	solar system*	Earth*	collection efficiency (percent)
H	27,000,000,000	9,500	0.00003
^{4}He	2,200,000,000	0.00005	0.000000000002
C	12,000,000	360	0.003
N	2,500,000	79	0.003
O	20,000,000	3,400,000	17
^{20}Ne	3,300,000	0.000093	0.000000003
Mg	1,100,000	1,100,000	100
S	520,000	98,000	19
^{36}Ar	88,000	0.00018	0.0000002
^{40}Ar	0.55	0.053	—†
Fe	900,000	1,200,000	133
^{84}Kr	26	0.0000036	0.00001

*Abundances indicate how many atoms of each element (or, in the case of the noble gases, isotopes) would accompany 10^6 silicon (Si) atoms. For example, the abundance of nitrogen (N) in the solar system is 2.5 times greater than that of Si, whereas its abundance in the Earth is less than that of Si by a factor of 0.000079. The Table includes the eight most abundant volatile elements, together with others. †See text.

Outgassing of the solid planet. The release of gases during volcanic eruptions is one example of outgassing; releases at submarine hydrothermal vents are another. Although the gas in modern volcanic emanations commonly derives from rocks that have picked up volatiles at the Earth's surface and then have been buried to depths at which high temperatures remobilized the volatile material, a very different situation must have prevailed at the earliest stages of the Earth's history.

The planet accreted from solid particles that formed as the primordial gas cloud cooled. Long before the volatile components of the cloud began to condense to form massive solid phases (*e.g.*, long before water vapour condensed to form ice), their molecules would have coated the surfaces of the solid particles of rocky material that were forming. As these solid particles continued to grow, a portion of the volatiles coating their surfaces would have been trapped and carried thereafter by the particles. If the solids were not remelted by impact as they collected to form the planet, the volatiles they carried would have been incorporated in the solid planet. In this way, even without collecting an enveloping gaseous atmosphere, a newly formed planet could include—as material occluded in its constituent grains—a substantial inventory of volatiles.

At some point in its early history, the Earth became so hot that much of the iron dispersed among the solid particles melted, became mobile, and collected to form the core. Related events led to the formation of rocky layers that were the precursors of the Earth's present-day mantle and crust. As part of this process of differentiation, volatiles present in the particles would have been released through outgassing. The outgassing must have occurred

Planetary differentiation

on a colossal scale if the accreting particles had retained their volatiles right up to the time of differentiation.

An atmosphere created by retention of these outgassing products would derive ultimately from nebular gases. Its chemical composition, however, would be expected to differ in two principal respects from that of an atmosphere formed by the capture of primordial gases: (1) whereas the captured atmosphere would contain all gases that were moving slowly enough (*i.e.,* that were sufficiently cold and/or of sufficient molecular weight) so that it was possible for the planet to retain them gravitationally, the outgassed atmosphere would contain only those gases "sticky" enough to have been significantly retained in the rocky particles from which the planet formed; and (2) methane and ammonia, two presumed components of a captured atmosphere, would probably not be stable under the conditions involved in outgassing. Thus, the noble gases, which would be poorly held by particles, would be of low abundance relative to gases derived from chemically active elements. Further, the principal forms of carbon and nitrogen in an outgassed atmosphere would be carbon monoxide or carbon dioxide together with molecular nitrogen.

Importation. A compromise between the extremes of direct capture and outgassing proposes that the Earth's inventory of volatiles was delivered to the planet late in its accretionary history—possibly after differentiation was nearly complete—by impact of a "last-minute" crop of solid bodies that were very strongly enriched in volatile materials (these were the last substances to condense as the solar nebula cooled). Such bodies might have had compositions similar to those of comets that still can be observed in the solar system. These last-minute condensates may have coated the planet as a surface veneer that yielded gases only when heated during differentiation, or they may have released their volatiles on impact.

Incorporation of volatile-rich solids

Because such bodies would have been relatively small, they would not have been able to retain primordial gases by means of a substantial gravitational field. Their complement of volatiles, retained by cold trapping in ices and on particle surfaces, would be expected to resemble the "sticky" (*i.e.,* polar and reactive) gases occluded by solid particles at earlier stages of cooling of the gas cloud but possibly lost during earlier higher temperature phases of the Earth's accretion.

Sinks. The dominant pathways by which gases are removed from the present atmosphere are discussed below in the section *Biogeochemical cycles.* Apart from those processes, three other sinks merit attention and are described here.

Photochemical reactions. Sunlight can provide the energy required to drive chemical reactions that consume some gases. Due to a rapid and efficient photochemical consumption of CH_4 and NH_3, a methane–ammonia atmosphere, for example, would have a maximum lifetime of about 1,000,000 years. This finding is of interest because it has been suggested that life originated from mixtures of organic compounds synthesized by nonbiological reactions starting from methane and ammonia. Recognition of the short atmospheric lifetimes of these materials poses grave difficulties for such a theory. Water, too, is not stable against sunlight that has not been filtered by overlying layers containing ozone or molecular oxygen, which very strongly absorb much of the Sun's ultraviolet radiation. Water molecules that rise above these layers are degraded to yield, among other products, hydrogen atoms, $H\cdot$.

Escape. Hydrogen and helium, or products like $H\cdot$, tend to have velocities high enough so that they are not bound by the Earth's gravitational field and are lost to space from the top of the atmosphere. The importance of this process extends beyond the very earliest stages of the Earth's history because continuous sources exist for these light gases. Helium is continually lost as it is produced by the decay of radioactive elements in the crust.

A combination of photochemical reactions and the subsequent escape of products can serve as a source for molecular oxygen (O_2), a major component of the modern atmosphere that, because of its reactivity, cannot possibly have derived from any of the other sources so far discussed. In this process, water vapour is broken up by ultraviolet light and the resulting hydrogen is lost from the top of the atmosphere, so that the products of the photochemical reaction cannot recombine. The residual oxygen-containing products then couple to form O_2.

Source of O_2

Solar-wind stripping. The Sun emits not only visible light but also a continuous flow of particles known as the solar wind. Most of these particles are electrically charged and interact only weakly with the atmosphere, because the Earth's magnetic field tends to steer them around the planet. Prior to the formation of the Earth's iron core and consequent development of the geomagnetic field, however, the solar wind must have struck the top layers of the atmosphere with full force. It is postulated that the solar wind was much more intense at that time than it is today and, further, that the young Sun emitted a powerful flux of extreme ultraviolet radiation. In such circumstances, much gas may have been carried away by a kind of atomic sandblasting that may have had a marked effect on the earliest phases of atmospheric development.

Biogeochemical cycles. Interactions with the crust and, in particular, with living things, the biosphere, can strongly affect the composition of the atmosphere. These interactions, which form the most important sources and sinks for atmospheric constituents, are viewed in terms of biogeochemical cycles, the most prominent and central being that of carbon. The carbon cycle, outlined schematically in Figure 1, includes two major sets of processes: biologic and geologic.

Biologic carbon cycle. The biologic processes of photosynthesis and respiration mediate the exchange of carbon between the atmosphere or hydrosphere and the biosphere,

Photosynthesis and respiration

$$CO_2 + 2H_2A \underset{\text{respiration}}{\overset{\text{photosynthesis}}{\rightleftharpoons}} CH_2O + 2A \text{ (or } A_2) + H_2O.$$

In these reactions, CH_2O crudely represents organic material, the biomass of bacteria, plants, or animals; and A represents the "redox partner" for carbon (reduction + oxidation → redox), the element from which electrons are taken during the biosynthesis of organic material and which accepts electrons during respiratory processes. In the present global environment, oxygen is the most prominent redox partner for carbon (*i.e.,* $A = O$ in the above equation), but sulfur (S) also can serve as a redox partner, and modified cycles based on other partners (*e.g.,* hydrogen) are possible. Imbalances in the biologic carbon cycle can change the composition of the atmosphere. For example, if O is the principal redox partner and if photosynthesis exceeds respiration, the amounts of O_2 will increase. The carbon cycle can in this way serve as a source for O_2. The strength of this source is dependent on the degree of imbalance between photosynthesis and respiration.

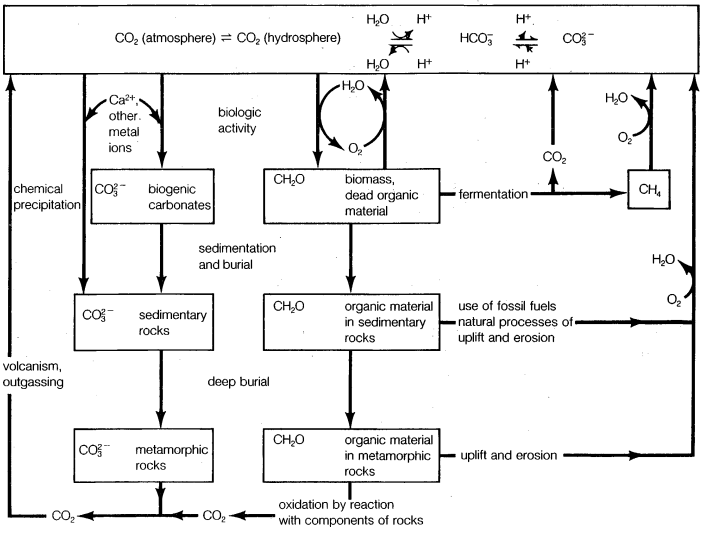


Figure 1: A schematic representation of the biogeochemical cycle of carbon.

The biologic degradation of organic material and the release of products to the atmosphere need not involve an inorganic redox partner such as oxygen or sulfur. Communities of microorganisms found in sediments are capable of carrying out the process of fermentation, in which electrons are shuffled among organic compounds. Many individual steps catalyzed by a variety of organisms are involved, but the overall reaction amounts to

Major source of methane in the atmosphere

$$2CH_2O \rightarrow CO_2 + CH_4.$$

This process is an important source of atmospheric methane.

Geologic carbon cycle. The geologic portions of the carbon cycle can be described most conveniently by following a carbon atom from the moment of its injection into the atmosphere in the form of carbon dioxide released from a volcano. The carbon dioxide—any CO_2 in the atmosphere—will come in contact with water in the environment and is likely to dissolve to form carbonic acid:

$$CO_2 + H_2O \rightleftharpoons H_2CO_3.$$

This weak acid is an important participant in weathering reactions that tend very slowly to dissolve rocks exposed to precipitation and groundwater at the Earth's surface. An exemplary reaction showing the conversion of a solid mineral to soluble products would be

$$CaSiO_3(s) + 2H_2CO_3(aq) \rightarrow Ca^{2+}(aq) + 2HCO_3^-(aq) + H_2SiO_3(aq),$$

where *s* indicates solid and *aq* stands for aqueous solution. Along with the other products of this reaction, the bicarbonate (HCO_3^-) derived from the volcanic CO_2 would eventually be transported to the ocean. At all points in the hydrosphere, the bicarbonate would be in equilibrium with other forms of dissolved CO_2 through chemical reactions that could be depicted as follows:

$$2HCO_3^-(aq) \rightleftharpoons CO_3^{2-}(aq) + H_2O + CO_2(aq).$$

In settings where its concentration was enhanced, carbonate (CO_3^{2-}) produced in this way could unite with the calcium ion Ca^{2+}, which is naturally present in seawater due to weathering reactions, to form solid $CaCO_3$, calcite, the principal mineral in limestone. The dissolved carbon dioxide might return to the atmosphere or remain in the hydrosphere. In either case, it eventually could enter the biologic carbon cycle and be transformed into organic matter. If the $CaCO_3$ and the organic matter sank to the bottom of the ocean, they both would be incorporated in sediments and could eventually become part of the rocky material of the crust. Uplift and erosion, or very deep burial and melting with subsequent volcanic activity, would eventually return the carbon atoms of the $CaCO_3$ and the organic matter to the atmosphere.

Interaction of biologic and geologic cycles. The pace of the biologic carbon cycle is measured in the lifetimes of organisms, while that of the geologic cycle is measured in the lifetimes of sedimentary rocks (which average about 600,000,000 years). Each interacts strongly with the atmosphere, the biologic cycle exchanging CO_2 and redox partners and the geologic cycle supplying CO_2 and removing carbonate minerals and organic matter—the eventual source of fossil fuels (*e.g.*, coal, oil, and natural gas)—in sediments. An understanding of the budgets and pathways of these cycles in the present global environment enables investigators to estimate their effects in the past, when conditions (the extent of evolution of the biota, the composition of the atmosphere, and so on) may have been quite different.

Quantitative significance of the processes of the biogeochemical cycle

The quantitative importance of these processes, now and over geologic time, can be summarized by referring to Table 2. Carbon in the atmosphere as carbon dioxide is almost the smallest reservoir considered in this tabulation, but it is the central point from which processes of the biogeochemical cycle have distributed carbon throughout the Earth's history. Reconstructions of atmospheric development must recognize that the very large quantities of carbon now found in sedimentary carbonates and organic carbon have flowed through the atmosphere and that the organic carbon (which includes all fossil fuels as well as far more abundant, ill-defined organic debris) represents material produced by photosynthesis but not recycled by respiration. The latter process must have been accompanied by the accumulation of the oxidized forms (*e.g.*, molecular oxygen, O_2) of carbon's redox partners.

Table 2: Carbon in the Earth's Crust

form	total amount (Pg* C)
Atmospheric CO_2 (as of 1978)	696
Oceanic CO_2, HCO_3^-, and CO_3^{2-}	34,800
Limestones, other carbonate sediments	64,800,000
Carbonate in metamorphic rocks	2,640,000
Total biomass	594
Organic carbon in ocean water	996
Organic carbon in soils	2,064
Organic carbon in sedimentary rocks	12,000,000
Organic carbon in metamorphic rocks	3,480,000

*One Pg (abbreviation for petagram) equals 10^{15} grams. Entries refer to amounts of carbon.

Table 2 also emphasizes the dissolution of atmospheric gases by the ocean. The carbon dioxide in the atmosphere is in equilibrium with, and far less abundant than, carbon dioxide, bicarbonate ion (HCO_3^-), and carbonate ion (CO_3^{2-}) in the ocean. If all carbon dioxide were somehow suddenly removed from the atmosphere, the ocean would replenish the supply within a few thousand years (the so-called stirring time of the ocean). Likewise, any change in the concentration of CO_2 in the atmosphere is accompanied by a quantitatively far larger change in the amount of CO_2, HCO_3^-, and CO_3^{2-} in the ocean. Similar equilibriums prevail for molecular nitrogen (N_2) and molecular oxygen (O_2). The atmosphere contains about 3,940,000 petagrams (Pg; one petagram equals 10^{15} grams) of nitrogen as N_2, with about 22,000 Pg being dissolved in the ocean. Oxygen is distributed in such a way that 1,200,000 Pg of O_2 are in the atmosphere while 12,390 Pg are in the ocean.

Weathering reactions. No matter what their origins, reactive gases in the atmosphere are likely to interact with other parts of the crust through what are termed weathering reactions. Not just carbonic acid associated with the carbon cycle but any acid becomes involved in acidic dissolution of susceptible rocks. As it does so, its concentration in the atmosphere declines, eventually reaching zero unless some process keeps replenishing the supply.

Even if respiration were suddenly to cease, oxygen produced by photosynthesis, or any oxidant in the atmosphere, would be consumed if oxidizable materials were present. The corrosion of metals is the most familiar example of this process in the modern world, but there are other examples involving natural forms of iron, sulfur, and carbon as well. Much of the iron bound in minerals is in the ferrous form (Fe^{2+}). As this material is exposed by uplift and erosion, it consumes atmospheric oxidants to form ferric iron (Fe^{3+}), the red, fully oxidized form of iron popularly identified as rust. Sulfide minerals (pyrite, or fool's gold, being the most familiar example) also consume oxidants as the sulfur is oxidized to produce sulfate. Finally, natural exposure of sedimentary organic matter, including coal beds or oil seeps, results in the consumption of atmospheric oxidants as the organic carbon is oxidized to produce carbon dioxide.

Oxidation

SEQUENCE OF EVENTS IN THE DEVELOPMENT OF THE ATMOSPHERE

Absence of a captured primordial atmosphere. If the planet grew large (and had, therefore, a substantial gravitational field) before all gases were dispersed from its orbit, it ought to have captured an atmosphere of nebular gases. The size and composition of such an atmosphere would depend on temperature as well as planetary mass. If the solid planet had reached full size and if temperatures were greater than 2,000 K, the minimum molecular weight that could be retained might have been high enough that the very abundant gases with molecular weights between 10 and 20 (methane, ammonia, water, and neon) would have been collected inefficiently, if at all. A thinner primordial atmosphere consisting of nebular gases with higher

molecular weights (*e.g.,* argon and krypton; see Table 1), however, ought still to have been captured.

In spite of this, characteristics of the present atmosphere (see below) show clearly that a primordial atmosphere either never existed or was completely lost. Explanations offered for both of these possibilities are linked to the development of the Sun itself. Astronomical observations of developing stars (*i.e.,* bodies similar to the early Sun) have shown that their early histories are marked by phases during which the gas in their surrounding nebulas is literally blown away by the pressure of light and particles ejected from the stars as they "turn on." (After this initial intense activity, young stars begin life with an energy output significantly below their mid-life maximum.) If the removal of gases occurred in the solar system after involatile solids had condensed but before the inner planets (Mercury, Venus, Earth, and Mars) accreted, it would have been impossible for the Earth to capture a primordial atmosphere. Alternatively, if planetary accretion preceded ejection of gases and the Earth had accumulated a primordial atmosphere, perhaps the early solar radiation, particularly the solar wind, was so intense that it was able to strip all gases from the inner planets, meeting the second condition described above—namely, complete loss.

Secondary atmosphere. The atmosphere that developed after primordial gases had been lost or had failed to accumulate is termed secondary. Although the chemical composition of the atmosphere has changed significantly in the billions of years since its origin, the inventory of volatile elements on which it is based has not.

Evidence for an atmosphere of secondary origin

Origin. The elemental composition of the volatile inventory reveals its secondary origin. Abundances are given in Table 1 for 12 nuclides that can be associated with four groups: (1) chemically active volatiles (H, C, N, O, S), (2) primordial noble gases (^{4}He, ^{20}Ne, ^{36}Ar, ^{84}Kr), (3) elements that form involatile minerals (O, Mg, S, Fe), and (4) a noble gas derived by the radioactive decay of an involatile element (^{40}Ar, derived from potassium). A comparison of entries in Table 1 shows that these groups have been collected by the Earth with sharply varying efficiencies. The column headed "collection efficiency" has been derived by the division of the abundance of each element on Earth by its abundance in the solar system and multiplying by 100. If the collection efficiency is close to 100 percent, the abundances are nearly equal and the transfer of this element from the solar system's initial reservoir to the planet was highly efficient. If the collection efficiency is low, most of the element was lost and is "missing" from the Earth's inventory. It is evident from Table 1 that efficiencies of collection are correlated primarily with chemical characteristics, not mass. This is the pattern expected if volatiles were retained by chemical interactions that yielded involatile phases rather than by gravitational attraction. Collection efficiencies for O, Mg, S, and Fe (which are included here only as representatives of the broad range of elements that were largely bound in involatile solid phases as the solar nebula cooled) are high. Those for the chemically active volatiles that could not form minerals stable at high temperatures (H, C, and N) are much lower. Spectacularly decreased efficiencies of collection are associated with the primordial noble gases.

The evidence points decisively to a process in which the elements to be retained in the terrestrial inventory were separated from those to be lost by a separation of solids from gases. The chemically active volatile elements could be incorporated in solids by formation of nitrides and carbides, by hydration of minerals, and by inclusion in crystal structures (*e.g.,* as ammonium and hydroxide ions) and could form some relatively involatile materials independently (organic compounds with high molecular weights are found in meteorites and were probably abundant in the cooling solar nebula); yet, none of these mechanisms was available to the noble gases. Formation of a group of solids rich in chemically active volatiles but not large enough to retain noble gases, followed by a loss of all materials still in the gas phase and an incorporation of the volatile-rich solids in the Earth, would be consistent with the chemical evidence and with the processes described above as outgassing and importation.

Evidence for outgassing as the primary source of the atmosphere

The special case of ^{40}Ar (Table 1) is particularly indicative of the derivation of the atmosphere through outgassing. Whereas the other noble-gas isotopes noted in Table 1 (^{4}He, ^{20}Ne, ^{36}Ar, ^{84}Kr) are primordial in origin, ^{40}Ar derives primarily from the radioactive decay of the isotope potassium-40. Therefore, even though the solar system abundance of ^{40}Ar is much lower than that of ^{36}Ar, its abundance on Earth is much higher because, uniquely among the noble-gas isotopes listed in Table 1, its source—the rock-forming element potassium (K)—is part of the solid planet. As radioactive potassium in rocks decayed over the Earth's history, the ^{40}Ar produced first became trapped within mineral crystals at sites formerly occupied by K^+, then was released when the crystals were melted in the course of igneous activity, and eventually reached the surface through outgassing. Given the abundance of potassium in the Earth's crust, it would be impossible to attribute the origin of the atmosphere to outgassing if the abundance of ^{40}Ar was far lower than that of ^{36}Ar, as in the solar system.

Early composition. The most critical parameter pertaining to the chemical composition of an atmosphere is its level of oxidation or reduction. At one end of the scale, an atmosphere rich in O_2 (like the present one of the Earth) is termed highly oxidizing, while one containing molecular hydrogen, H_2, is termed reducing. These gases themselves need not be present. Modern volcanic gases are located, for example, toward the oxidized end of the scale. They contain no O_2, but all hydrogen, carbon, and sulfur are present as H_2O, CO_2, and SO_2 (oxidized forms), while nitrogen is present as N_2 (not NH_3). A relationship prevails between the oxidation or reduction of outgassing volatiles and the inorganic material with which they come in contact: any hydrogen, carbon, or sulfur brought into contact with modern crustal rocks at volcanic temperatures will be oxidized by that contact.

The abundance of hydrogen in the solar nebula, the common occurrence of metallic iron in meteorites (representative of primitive solids), and other lines of geochemical evidence all suggest that the Earth's early crust was much less oxidized than its modern counterpart. Although all iron in the modern crust is at least partly oxidized (to Fe^{2+} or Fe^{3+}), metallic iron may have been present in the crust as outgassing began. If the earliest outgassing products were equilibrated with metallic iron, hydrogen would have been released as a mixture of molecular hydrogen and water vapour, carbon as carbon monoxide, and sulfur as hydrogen sulfide. The presence of metallic iron during the last stages of outgassing is, however, unlikely, and, because H_2 is not gravitationally bound, it would have been lost rapidly. At an early point, hydrogen would have been almost completely in the form of water vapour and carbon in the form of carbon dioxide. Nitrogen would have been outgassed along with the carbon and hydrogen. As carbon dioxide was consumed by weathering reactions and water vapour condensed to form the oceans, molecular nitrogen must have become the most abundant gas in the atmosphere. It is certain that molecular oxygen was not among the products of outgassing.

Absence of metallic iron in ancient rocks

Among the oldest rocks are water-laid sediments with an age of 3,800,000,000 years. Neither they nor any other ancient rocks contain metallic iron, though nearly all contain oxidized iron (Fe^{2+}). Carbon is present both as organic material and in a variety of carbonate minerals. The existence of these sediments requires atmospheric pressures and temperatures consistent with the presence of liquid water. The nature of the iron minerals and their abundance suggest that Fe^{2+} was a significant component of ocean water and that concentrations of O_2 had to have been essentially zero because Fe^{2+} reacts very rapidly with O_2.

The presence of organic carbon and carbonate minerals in the sediments dated 3,800,000,000 years old would be consistent with the development of a biologically mediated carbon cycle by that point in time, but the degree of preservation of these materials (which were heated to temperatures near 500° C [932° F] for millions of years at some point in their history) is so poor that the question cannot be settled. Relatively well-preserved sediments with

Rise of a biologically controlled carbon cycle

an age of 3,500,000,000 years are far more abundant. In addition to abundant organic carbon and carbonate minerals, they contain microfossils and sedimentary features demonstrating convincingly that life had arisen on Earth by that time. The distribution of the stable isotopes of carbon (carbon-12 and carbon-13) in sedimentary materials younger than 3,500,000,000 years demonstrates that living organisms were effectively in control of the global carbon cycle from that time onward.

The existence of sedimentary carbonates is direct evidence that carbon dioxide was present in the atmosphere. Its precise abundance is not known, but the best estimates are that it was substantially, perhaps 100 times, higher than the present atmospheric level. A strongly enhanced greenhouse effect (see below *Atmospheric heat budget and energy transfer*), leading to more efficient retention of heat derived from solar radiation, would be expected. For many students of the Earth's history, the fact that the early oceans did not freeze in spite of the dim Sun is evidence that the abundance of atmospheric carbon dioxide was high enough to provide the enhanced greenhouse effect.

Rise of molecular oxygen. Recognition of the nature of the Earth's pre-oxygenic environment is critical to consideration of this problem. If humans could somehow take a spaceflight not to another planet but to the Earth of 3,000,000,000 years ago, they would find that space suits would have been required on their home planet at that time. More dramatically, if those time-traveling astronauts were somehow able to take with them all of the oxygen from the modern atmosphere, they would find that it would disappear soon after release. Not only was oxygen absent in the early atmosphere but potent sinks for O_2 were abundant as well. Oxidizable materials such as ferrous iron, sulfides, and organic compounds littered environments in which they are now absent. These chemicals absorbed O_2 almost immediately after its release. Moreover, as the oxygen-absorbing capacity of such compounds was exhausted, new material that had been eroded from the unoxidized crust took their place. This process continued until the rock cycle (sedimentation, burial, igneous activity, uplift, and erosion) had exposed all oxidizable materials in the crust. No matter what the supply of O_2, the process must have taken time (about half the rock volume of the crust is recycled every 600,000,000 years). It is, therefore, very important to distinguish clearly between the first biologic production of O_2 and its persistent accumulation in the atmosphere. It is conceivable, even likely, that these events were separated by hundreds of millions of years. Evidence bearing on the oxygenation of the Earth's atmosphere is summarized in Figure 2 and discussed in the following sections. The abundance of O_2 at each point is expressed in terms of its approach to the present atmospheric level (PAL). For example, because the pressure of O_2 in the present atmosphere is 0.21 atm (abbreviation for atmosphere, a unit of pressure equal roughly to 14.7 pounds per square inch), a planetary atmosphere containing 10 percent of that amount, 0.021 atm, would be described as having an oxygen level of 0.1 PAL.

Photochemical reactions as a limited source of O_2

Photochemical production. The strength of this source is limited by the requirement that water vapour rise in the atmosphere to altitudes at which solar ultraviolet radiation capable of cleaving water molecules has not yet been absorbed by other atmospheric constituents. The transport of water vapour to high altitudes is severely impeded by a cold layer in the atmosphere. Water vapour freezes in this layer, and the rate of photochemical production of O_2 is thus limited. The severity of this limitation is not precisely known, but it is evident that atmospheric levels of oxygen did not rise until oxygenic photosynthesis was well established. This does not indicate that photochemical production of O_2 was insignificant. Rather, it demonstrates that the strength of the process as a source was exceeded by the strength of the contemporary oxygen sinks (chiefly oxidative weathering reactions at the Earth's surface) and that residence times for O_2 were so short that significant atmospheric concentrations could not accumulate. The best estimate is that pressures of O_2 at sea level and ground level were less than 5×10^{-8} PAL.

Onset of oxygenic photosynthesis. The development of a biologically mediated carbon cycle prior to 3,500,000,000 years ago virtually requires that some form of photosynthesis had arisen by that time, but the possibility remains that sulfur or hydrogen, not oxygen, was serving as the redox partner. It also has been noted that some sediments 3,500,000,000 years in age contain microfossils with shapes resembling those of modern oxygenic photosynthesizers. This is suggestive, though not compelling, evidence that oxygenic photosynthesis had developed by 3,500,000,000 years ago. Shape is an infamously imprecise indicator of biochemical characteristics of microorganisms. More specifically, while it might be possible to recognize a photosynthetic organism from its shape, it is very difficult to determine exactly what redox partners that organism employed.

Geochemical and paleontological features of sedimentary rocks 2,800,000,000 years in age offer stronger evidence

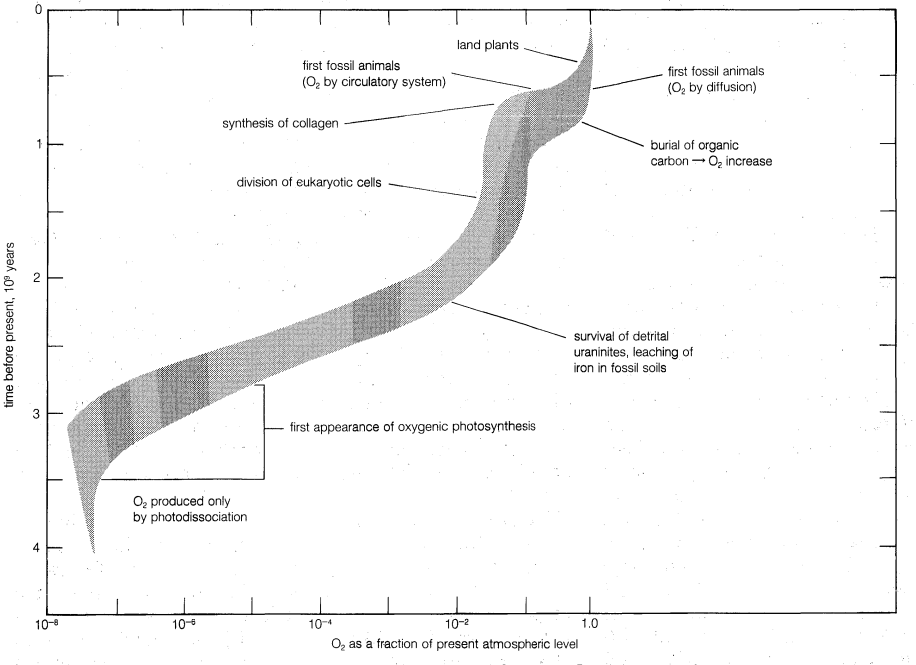


Figure 2: A "best guess" reconstruction of the abundance of O_2 in the Earth's atmosphere as a function of time. The O_2-abundance axis is logarithmic.

that oxygenic photosynthesis had arisen by that time. At 2,800,000,000 years, the abundance of carbon-13 in sedimentary organic carbon decreases sharply from levels maintained between 3,500,000,000 and 2,900,000,000 years ago, then slowly rises, regaining those levels about 2,200,000,000 years ago. This has been interpreted in terms of a transient in the biogeochemical carbon cycle in which biogenic methane (which is strongly depleted in carbon-13) served as an important mobile constituent of the cycle during the interval from 2,800,000,000 to 2,200,000,000 years ago. According to this interpretation, methane was able to play this role only after O_2 became available and facilitated its metabolism. As O_2 sinks decreased in strength and the atmosphere became oxidizing, however, the mobility of methane was reduced and the methane cycle took on its modern form, which seldom leads to strongly decreased abundances of carbon-13 in sedimentary organic matter.

Microfossils resembling modern oxygenic photosynthesizers also appear in sediments of this age, and they are accompanied by sedimentologic features (apparent "fossil gas pockets") that are interpreted as evidence of aerobic metabolism. Thus, evidence dating from about 2,800,000,000 years ago is more abundant and diverse (geochemically, morphologically, and sedimentologically) than that found in rocks 3,500,000,000 years of age. In spite of these points of consistency, this evidence is not decisive.

Evidence from younger sediments (see below) indicates that oxygenic photosynthesis almost certainly developed earlier than 2,200,000,000 years ago. Whatever the precise moment of development, it marked the origin of the first so-called oxygen oasis, a restricted environment in which the abundance of O_2 rose above 5×10^{-8} PAL probably quite significantly. Within such oases, aerobic metabolism could occur. At their margins, the delivery of oxidizable materials from the surrounding global environment overwhelmed the local supply of O_2. Overall, the atmosphere did not become oxidizing, but, as oxygenic photosynthesizers proliferated, the number and size of the oases grew.

Formation of oxygen oases

Transition to an aerobic environment. Pyrite and uraninite are minerals of iron and uranium, respectively, that are not stable in the presence of O_2. Though they can be found in some modern river sediments, neither can survive in them for thousands of years. Yet, many sediments older than about 2,200,000,000 years contain well-rounded grains of these minerals. Their shapes and locations indicate prolonged exposure and tumbling in ancient rivers or as beach deposits, but there is no evidence of chemical attack by oxygen. The precise significance of this observation is best considered together with measurements of the movement of iron in fossil soil profiles.

If soil gases (in equilibrium with the atmosphere) contain O_2, iron exposed during the breakdown of soil minerals will be immobilized by oxidation and will not be leached from near-surface soil horizons. Conversely, if O_2 is absent during soil development, chemical analysis of fossil soils will reveal depletion of iron near the former soil surface. Rates of the dissolution of uraninite and the leaching of iron in soil profiles also depend on the abundance of carbon dioxide (CO_2). Because the patterns of dependence are different, the combination of evidence based on both phenomena allows for the estimation of abundances of both CO_2 and O_2. This line of interpretation leads to the conclusion that about 2,200,000,000 years ago the ratio of the molecular abundance of O_2 to that of CO_2 was about 1.3 (at present it is 635), and that the pressure of O_2 was near 0.01 PAL while that of CO_2 was about nine times higher than at present. Other workers agree that the uraninite and fossil soil data indicate the development of oxidizing conditions at the surface by 2,200,000,000 years ago, but they place the most probable level of O_2 lower by a factor of 10 or more.

The consumption of oxidizing power by the crust is recorded by the inorganic constituents of sedimentary rocks. Iron-bearing sediments, or iron formations, are of particular interest because the collection of substantial quantities of iron in a sedimentary basin requires that iron be mobile in the world ocean. Mobility requires solubility, and, while Fe^{2+} is soluble, Fe^{3+}, the form of iron that results if O_2 comes in contact with Fe^{2+}, is highly insoluble.

Significance of iron-bearing sediments

Three states can be distinguished: (1) The existence of iron formations containing only Fe^{2+} suggests a complete absence of oxygen. (2) The existence of iron formations containing Fe^{2+} and Fe^{3+} indicates that levels of oxygen were low enough—essentially zero in the deep ocean—so that iron was mobile, but it also suggests that O_2 (perhaps at an oxygen oasis) was important in triggering deposition of the iron, though other means of oxidation—photochemical processes, for example—are quite conceivable. (3) The disappearance of iron formations from the sedimentary record suggests persistent oxygenation of the ocean. This sequence of possibilities is represented in the geologic record as follows: (1) The oldest sedimentary rocks are iron formations that contained only Fe^{2+} at the time of their deposition. (2) The first appearance of primary Fe^{3+} (produced during the formation of the rock rather than in later weathering) was in iron formations about 2,700,000,000 years ago. (3) Iron formations disappeared almost completely from the record about 1,700,000,000 years ago (with a few isolated and very small recurrences about 1,000,000,000 years ago). Moreover, the abundance of iron formations increased significantly from 2,700,000,000 to 2,200,000,000 years ago, suggesting that some new factor, possibly oxidative precipitation of Fe^{3+}, was enhancing the rate of deposition. It is for this same time interval that carbon-isotopic evidence indicates the operation of an O_2-dependent methane cycle.

Evidence for the evolution of eukaryotic organisms (those containing a membrane-bound nucleus and other organelles) first appears in the microfossil record of about 1,400,000,000 years ago. Biochemical reactions that occur during the growth and division of such cells require oxygen levels of 0.02 PAL. Attainment of that level by 1,400,000,000 years ago apparently led to oxygenation of the deep sea and the cutoff of deposition of iron formations about 1,700,000,000 years ago.

Attainment of the modern O_2 level. The abundance of carbon-13 in sedimentary organic materials and in carbonates from 900,000,000 to 600,000,000 years ago indicates that unusually large quantities of organic carbon were buried without reoxidation during that interval. The burial of this carbon must have been accompanied by the accumulation of oxidized forms of carbon's redox partners. The quantities released were adequate to raise the level of O_2 to 1.0 PAL or more.

Reconstruction of O_2 levels on the basis of biologic evidence

It has been calculated that oxygen requirements of the earliest animals, which developed about 700,000,000 years ago, would have been met—if the animals had circulatory systems that incorporated oxygen carriers like hemoglobin—by O_2 abundances as low as 0.1 PAL. If circulatory systems had not yet evolved, an O_2 abundance of 1.0 PAL would have been required. Studies of fossils indicate that the animals were very thin (one to six millimetres [0.04–0.24 inch]) in spite of great breadth and length (up to 1,000 millimetres [39 inches]). Such a shape seems optimized for transport of O_2 by diffusion from the surrounding water to the cells in which it was needed, thus pointing to the latter higher value (namely, an O_2 abundance of 1.0 PAL). Other reconstructions of O_2 levels based on biologic evidence suggest that the widespread development of land plants about 400,000,000 years ago must have driven O_2 to levels near 1.0 PAL, and they show O_2 levels rising smoothly from levels near 0.1 PAL at 650,000,000 years ago to 1.0 PAL at 400,000,000 years ago.

Variation in abundance of carbon dioxide. The approximately hundredfold decline of atmospheric CO_2 abundances from 3,500,000,000 years ago to the present has apparently not been monotonous. During that interval, numerous ice ages have come and gone. Significant changes in climate can result from geographic changes, but it is generally concluded that modulation of the efficiency of the Earth's greenhouse effect is also required to produce the extreme variations associated with widespread continental glaciations. In recognition of this, broad climatic variations during the past 750,000,000 years have been described in terms of alternating "icehouse" and "greenhouse" episodes.

Association with extreme climatic changes

Icehouse conditions—apparently associated with the depletion of atmospheric CO_2, the principal greenhouse gas—have prevailed since about 65,000,000 years ago and during two earlier periods, 650,000,000–530,000,000 and 360,000,000–240,000,000 years ago. It is suggested that intervening greenhouse episodes have been associated with higher abundances of CO_2 in the atmosphere. The hypothesis is far from proved. Nonetheless, its details are being explored aggressively amid concerns that the accelerated production of CO_2 due to industrial combustion of fossil fuels may reverse climatic conditions with catastrophic effect. It is feared that such a reversal could result in the melting of the polar ice caps and a consequent flooding of coastal areas (see below *Climate modification*).

(J.M.Ha.)

Structure of the present atmosphere

GENERAL CHARACTERISTICS

The atmosphere extends from the surface of the Earth to heights of thousands of kilometres, where it gradually merges with the solar wind. The composition of the atmosphere as measured by its mean density (the average mass per unit volume) is more or less constant with height to altitudes of about 100 kilometres. This state of approximate uniformity arises as a result of motion and as a consequence of the high frequency with which molecules of a particular species are involved in collisions with their neighbours. A representative oxygen molecule, O_2, for example, encounters a nitrogen molecule, N_2, on average once every 10^{-9} second at the surface. Even at heights of 100 kilometres, where the density of air molecules is much lower, the encounter time is still comparatively brief, about 10^{-3} second. A force imparted to one molecule is rapidly transferred to all.

The atmosphere tends to behave as though it were composed of a single molecular species with an effective molecular mass set by its mean composition. The bulk of the lower atmosphere is composed of N_2 and O_2, with relative abundances of, respectively, 0.78 and 0.21 based on the average number of molecules present in a representative volume of air. The mass of the hypothetical mean molecule of the lower atmosphere is 28.97 atomic units (one atomic unit corresponds to the mass of a hydrogen atom, 1.66×10^{-24} gram). This value is intermediate between that of N_2 (28 atomic units) and that of O_2 (32 atomic units) and reflects the presence in the atmosphere of trace quantities of water (18 atomic units), argon (40 atomic units), carbon dioxide (44 atomic units), and other less abundant compounds as well.

The collisional interaction between individual molecules becomes progressively less efficient at altitudes above 100 kilometres. Molecules begin to experience a force of gravity proportional to their individual molecular masses. Heavy gases are bound more closely to the Earth, whereas lighter gases are free to float higher. The average molecular mass of the atmosphere therefore declines steadily with increasing altitude, as illustrated in Figure 3. Atomic oxygen is more abundant than N_2 above about 160 kilometres. In turn, atomic oxygen gives way to helium above 600 kilometres and hydrogen is the major constituent at altitudes higher than 1,000 kilometres. The region above 100 kilometres is referred to as the heterosphere, a name intended to emphasize the importance of the change in composition as a function of altitude. In the same vein, the region lower than 100 kilometres was given the name homosphere.

Decrease of average molecular mass with increasing altitude

DIVISION BASED ON THERMAL STRUCTURE

A second classification, based on thermal structure, provides a more detailed and, in many respects, more useful scheme for the division of the atmosphere into distinct layers (Figure 4).

The temperature decreases rapidly above the surface of the Earth to an altitude of about 17 kilometres. The air is relatively unstable, a consequence of the decrease of temperature with altitude. Warmer air is comparatively light and has a tendency to rise. Conversely, colder air is dense and tends to sink. The atmosphere is poised to turn over, to convect much like water in a kettle heated from below. This region is known as the troposphere, a term derived from the Greek words *tropos,* "turning," and *sphaira,* "ball." Most of the weather of the planet is confined to the troposphere. The upper boundary of the troposphere is called the tropopause.

Troposphere

The temperature begins to increase slowly with altitude above the tropopause in a region known as the stratosphere, from the Latin word *stratus,* meaning "stretched out" or "layered." Vertical motions are strongly inhibited in the stratosphere. An air parcel that attempts to rise becomes rapidly colder and denser than the air it displaces. Buoyancy forces in this environment act to suppress vertical motion. Motions in the stratosphere are thus largely confined to the horizontal, accounting for the layered structure of high-altitude stratus clouds. The increase of

Stratosphere

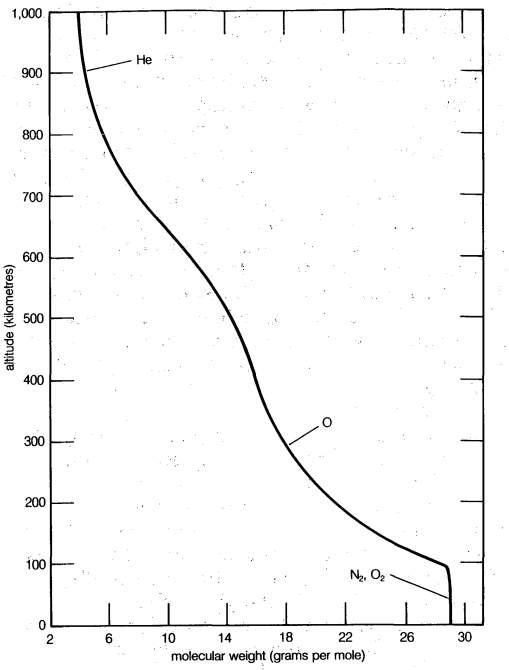


Figure 3: Average molecular mass of the atmosphere in atomic units (one atomic unit corresponds to the mass of a hydrogen atom) illustrating changes in composition with altitude.

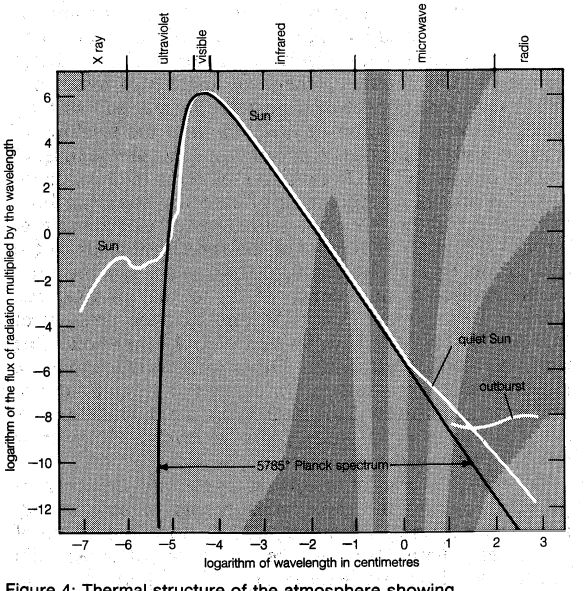


Figure 4: Thermal structure of the atmosphere showing depths of penetration for sunlight of different wavelengths (λ).

temperature with altitude persists to about 50 km, at which point the temperature is about as high as at the surface. This marks the upper boundary of the stratosphere, the stratopause.

The temperature resumes its general decrease with altitude above the stratopause in the mesosphere (*mesos* denoting "middle"). It reaches a minimum near 85 km at the mesopause, which is the coldest region of the atmosphere. The temperature increases again with altitude above the mesopause in the thermosphere, so named because of the importance of thermal conduction in this region. A large portion of the heat deposited in the thermosphere is conducted downward and is radiated out to space from the vicinity of the mesopause.

The thermal structure of the atmosphere reflects in part the influence of energy deposited directly by the absorption of sunlight. It is determined, though, to a much larger extent by a complex suite of processes important to redistributing energy vertically. The Sun is the ultimate source of energy. Approximately 45 percent of the energy incident from the Sun is absorbed by the surface. A comparable amount, about 33 percent, is reflected back to space, either by clouds, 26 percent, or by the surface itself, 7 percent, as shown in Figure 5. The atmosphere absorbs only 22 percent of the incident energy; most of this is captured by dust in the troposphere. The atmosphere is bathed in two more or less distinct radiation fields. The first field, originating in the Sun, has the majority of its energy in the visible and ultraviolet portions of the electromagnetic spectrum. The second, emanating from the surface of the Earth and its lower atmosphere, has most of its energy at longer wavelengths—namely, in the infrared.

The solar spectrum at visible wavelengths is about what would be expected for a blackbody radiating at a temperature of 5,785 K, the temperature of the photosphere from which most of the solar radiation is emitted. (A blackbody is a hypothetical ideal body or surface that absorbs and reemits all radiant energy falling upon it.) Radiation at shorter wavelengths is more intense (see Figure 6). Light at ultraviolet and X-ray wavelengths emanates from the outermost regions of the solar atmosphere, the chromosphere and corona. Temperatures there climb to values above 10^6 K.

Viewed from space, the spectrum of the Earth would be similar to that shown schematically in Figure 7. At longer wavelengths the radiation would be emitted by the atmosphere and surface and derived more or less equally from the dayside and nightside of the planet. At shorter wavelengths the spectrum would be dominated by sunlight reflected by clouds and by the surface on the dayside.

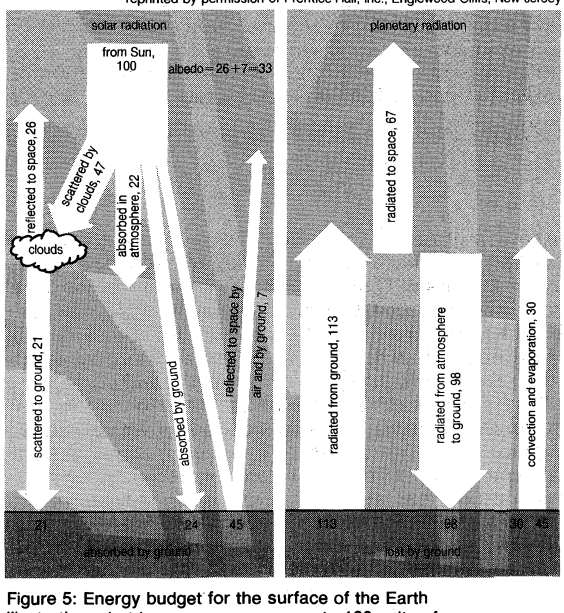


Figure 5: Energy budget for the surface of the Earth illustrating what happens, on average, to 100 units of energy incident from the Sun.

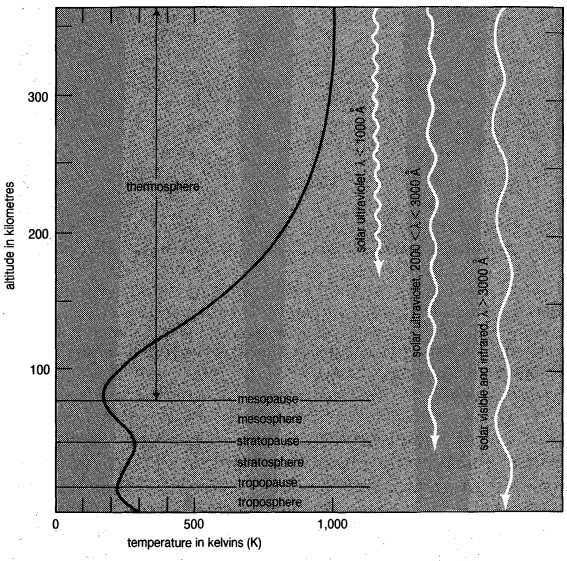


Figure 6: Spectrum of the Sun compared with energy emitted by an ideal blackbody at 5,785 K.

From C.W. Allen, *Quarterly Journal Royal Meteorological Society* (1958)

ATMOSPHERIC HEAT BUDGET AND ENERGY TRANSFER

The overall heat budget of the atmosphere and surface is summarized in Figure 5. The surface, on average, receives 17 percent of its heat directly from the Sun, 15 percent from solar radiation scattered by clouds, and the balance, 68 percent, from absorption of infrared radiation emitted by the atmosphere. The greater part of the energy absorbed by the surface, 79 percent, is returned to the atmosphere in the form of radiation, with spectral properties determined by the local ground temperature. The remainder, 21 percent, is transmitted to the atmosphere by conduction and as a by-product of the exchange of water, H_2O. The surface can cool by evaporation of H_2O, and the associated heat is transmitted to the air as vapour, which recondenses to form clouds and either rain or snow. Phase changes of H_2O play a major role in the energy budget of the lower atmosphere. It is, in fact, the importance of H_2O that sets the Earth apart from all of its neighbours in the solar system.

The atmosphere can be conceived of as a compressible fluid of infinite extent that is heated from below by a moist radiating surface and perturbed locally by energy absorbed from sunlight. Direct absorption of solar radiation is important primarily for the stratosphere and thermosphere. Transfer of energy by infrared radiation is a dominant mode for heat transmission between the different atmospheric layers, with an additional contribution due to motions generated by spatial heterogeneities in heating rates.

Transfer of energy by radiation is effected mainly by trace constituents of the atmosphere, primarily H_2O, CO_2, and O_3. In contrast to the major constituents, N_2 and O_2, these gases are able to absorb the longer wavelengths of the planetary radiation field. They thus assume an importance out of proportion to their abundance. They act to trap heat radiated by the surface, much as the glass panes of a greenhouse do. Like the glass, the atmosphere is transparent to sunlight but is essentially opaque to longer wavelengths. The infrared-active gases return heat to the ground, accounting for about 70 percent of the net input of energy to the surface. If the atmosphere were devoid of water and carbon dioxide, the surface temperature would be about 40 K colder than it is today, and large portions of the planet would be covered by ice.

The so-called greenhouse effect

Since the early 1980s there has been growing concern over the possibility that an increase in the abundance of carbon dioxide caused by combustion of fossil fuels could lead to a general warming of the global climate. Similar effects can arise from increases in the abundances of methane, nitrous oxide, and various chlorofluorocarbons (CFC's) such as CCl_2F_2 and CCl_3F. These species are re-

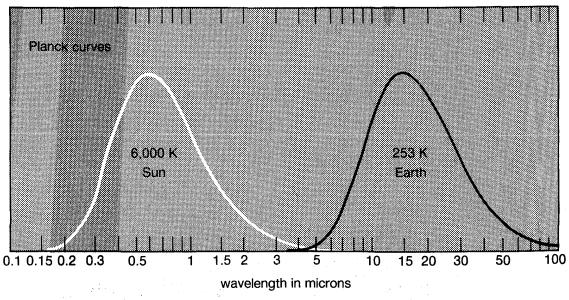


Figure 7: Spectrum of the Earth as viewed from space showing distinction between reflected sunlight and planetary radiation. The Earth is assumed to emit as a blackbody at an average temperature of 253 K.
Royal Meteorological Society

ferred to collectively as the greenhouse gases in recognition of their ability to trap heat. Their importance to climate is explored in greater detail below in *Climate modification.*

The increase of temperature with altitude above the tropopause is due primarily to absorption of solar radiation by ozone, O_3. As noted above, O_3 is a minor constituent of the atmosphere, a product of the interaction of molecular oxygen, O_2, with sunlight. It plays a critical role in the global life-support system, however, absorbing most of the light incident on the Earth in the ultraviolet portion of the spectrum with wavelengths between about 200 and 300 nanometres (one nanometre [nm] equals 10^{-9} metre). The absorption process results in the dissociation of O_3. A portion of the photon energy ($h\nu$) is deposited directly as heat; a larger fraction appears initially as internal excitation of O and O_2. The internal energy is degraded rapidly to heat in the stratosphere by collisions with O_2 and N_2. Finally, the chemical potential energy represented by $O + O_2$ is itself converted to heat as O_3 is reformed by reaction of O with O_2.

ecom-
ination
rocess

The recombination process involves two separate reactions,

$$\begin{aligned} O + O_2 &\rightleftharpoons O_3^* \\ O_3^* + M &\rightarrow O_3 + M, \end{aligned} \qquad (1)$$

where O_3^* denotes an unstable, energetic intermediate form of O_3. The excess energy in O_3^* is removed by collisions with atmospheric molecules M in the second step. The pair of reactions is usually considered as a single reaction,

$$O + O_2 + M \rightarrow O_3 + M. \qquad (2)$$

The rate of recombination, the number of O_3 molecules formed per unit volume per unit time, is proportional to the product of the concentrations of O, O_2, and M. The constant of proportionality, the reaction rate constant, may be determined experimentally in the laboratory.

To an excellent approximation in the stratosphere, it may be assumed that all of the radiative energy absorbed by O_3 is converted locally to heat. The heating rate—and ultimately the course of temperature with altitude—depends on the details of the distribution of O_3 with height.

The inversion of temperature above 80 kilometres, in the thermosphere, is due to energy extracted from sunlight at wavelengths below 200 nanometres. There are several important processes. Between 100 and 200 nanometres, absorption of solar radiation leads to dissociation of O_2,

$$h\nu + O_2 \rightarrow O + O. \qquad (3)$$

At shorter wavelengths, absorption is associated primarily with the ionization of O, O_2, and N_2,

$$\begin{aligned} h\nu + O &\rightarrow O^+ + e \\ h\nu + O_2 &\rightarrow O_2^+ + e \\ h\nu + N_2 &\rightarrow N_2^+ + e. \end{aligned} \qquad (4)$$

(Here, $h\nu$ represents a photon and e is the electron emitted.) Integrated over altitude, dissociation of O_2 is balanced by recombination. Molecular oxygen is reformed, either directly by

$$O + O + M \rightarrow O_2 + M, \qquad (5)$$

or indirectly by reactions such as

$$\begin{aligned} O + O_2 + M &\rightarrow O_3 + M \\ H + O_3 &\rightarrow OH + O_2 \\ O + OH &\rightarrow O_2 + H. \end{aligned} \qquad (6)$$

Sequence (6) is equivalent in effect to reaction (5). It can proceed more rapidly, however, under appropriate conditions in the atmosphere, depending on the efficiency of the step associated with the production of O_3 and the availability of hydrogen, H. The recombination path (5) is said to be catalyzed by the presence of H. As will be seen, catalytic reaction chains play a major role in much of the chemistry of the atmosphere below 80 kilometres.

The path from the absorption of solar energy by reactions (3) and (4) to the ultimate disposal of this energy as heat is less direct for the thermosphere than for the stratosphere. Recombination of O atoms, mainly by reactions (5) and (6), can occur at altitudes quite different from those at which O_2 is dissociated. The lower densities of the thermosphere allow O atoms to diffuse downward, with a compensating upward flow of O_2. Recombination requires relatively high densities and is confined mainly to altitudes below 100 kilometres. Dissociation of O_2, on the other hand, can proceed at any level, limited solely by the supply of O_2 and by the availability of photons with energy sufficient to fragment the bond in O_2. This spatial separation of dissociation and recombination results in a vertical redistribution of energy. Energy absorbed at one level by O_2 is converted to heat at another and is transferred in between in the form of chemical potential energy represented by a downward flow of O. Further, only a fraction of the available chemical potential energy is converted to heat. The balance is released as radiation, contributing to the phenomenon of airglow (see below).

Vertical redistribution of energy

The fate of the energy deposited in (4) is similarly complex. A fraction of the photon energy is imparted initially to the photoelectron (an electron ejected by the action of a single photon). The photoelectron loses energy by collisions, both elastic and inelastic, with atmospheric molecules. It can cause further ionization and contribute to the production of excited states and the associated emission of airglow. Photoelectrons share energy efficiently with ambient electrons. As a consequence, the electron temperature tends to rise above the temperature of the neutral gas. There is a flow of energy from electrons to ions to neutrals. To some extent the thermosphere must be treated as though it were composed of three characteristic thermal reservoirs, each one having its own distinct temperature.

The heating rate depends on a complex suite of atomic and molecular processes and is difficult to evaluate precisely. To further complicate matters, the chemical potential energy represented by the photo-ion can contribute to heating at levels far removed from its original source.

Electrons are removed by dissociative recombination of molecular ions such as the oxygen ion O_2^+ and that of nitric oxide, NO^+:

$$e + O_2^+ \rightarrow O + O \qquad (7)$$

$$e + NO^+ \rightarrow N + O. \qquad (8)$$

A portion of the potential energy is converted to heat in reactions (7) and (8). Some is used to produce excited, radiating states of O and N, and some is stored in excited states and degraded to heat subsequently by collisions. The balance is represented by the atoms N and O and is released only when the atoms recombine at a lower altitude to reform the stable molecules N_2 and O_2.

An atom or molecule of mass m is bound to the Earth by the force of gravity. The work done in moving a vertical distance ΔZ is $mg\ \Delta Z$, where g defines the gravitational acceleration, 9.8 metres per second per second. The work that must be done to escape the gravitational field is mgR, where R is the radius of the Earth—6,400 kilometres. The average kinetic energy of the atom is $(3/2)kT$, where k is the Boltzmann constant, 1.38×10^{-16} erg per degree Kelvin and T is temperature. The atom can escape the gravitational field if kT is much larger than mgR. In practice, escape is impeded by collisions, except at the highest

levels where the density of the atmosphere is low and collisions are comparatively rare.

Scale height

The atmosphere extends to great heights, with density declining by a factor of e (2.72) over an altitude interval given by $(kT)/(mg)$. The quantity $(kT)/(mg)$ is known as the scale height, denoted customarily by H. As noted above, it is proportional to the ratio of thermal kinetic energy to gravitational potential energy and measures the capacity of the atmosphere to support itself thermally in opposition to the confining force of gravity. The scale height is about seven kilometres in the lower atmosphere, rising to values in excess of 50 kilometres in the thermosphere where the mean molecular mass is smaller, reflecting the importance of lighter gases such as oxygen and helium. The scale height provides a useful measure of the vertical extent of the atmosphere: the atmosphere above a height z contains a total mass equivalent to that represented by a hypothetical layer of vertical extent H containing gas with density equal to the density at z.

If the mean distance traveled by an atom or molecule between collisions exceeds H, it can be assumed that the effect of collisions will be relatively unimportant. The average distance between collisions, known as the mean free path and denoted by λ, is given by $\{Qn\}^{-1}$, where Q is the collision cross section, the target area offered by a typical atom or molecule, and n is the number of atoms or molecules (targets) per unit volume. The level at which $\lambda = H$ is defined as the critical level, z_c. Effects of collisions may be ignored for heights greater than z_c. The region above z_c is known as the exosphere. An atom or molecule moving upward through z_c with speed greater than about 11 kilometres per second has kinetic energy in excess of its gravitational potential energy and in the absence of collisions is free to escape from the Earth.

Hydrogen loss

Atoms or molecules of a particular species have a range of speeds determined by their temperature, as described by the Maxwell–Boltzmann distribution law. For temperatures near the critical level, 750–2,000 K, significant numbers of hydrogen atoms have velocities greater than 11 kilometres per second. These atoms readily escape. There is an upward flow of hydrogen at all levels of the atmosphere to supply the flux leaving the Earth. Hydrogen is lost at a rate of about 10^8 atoms per square centimetre per second averaged over the surface of the planet. These escaping hydrogen atoms are derived ultimately from H_2O in the oceans. Integrated over the 4.5×10^9 years of the Earth's history, the escape of hydrogen has removed about two metres depth of H_2O from the surface of the oceans, liberating a quantity of O_2 roughly equivalent to that present in the atmosphere today. The escape of hydrogen is not, however, thought to be the major factor in regulating the level of atmospheric O_2. Burial of organic carbon in sedimentary rocks provides a more immediate source, though the escape of hydrogen may have been important at an earlier stage in the life of the Earth.

Thermal escape, as discussed here, is important mainly for hydrogen. There is, however, also a significant loss of helium. The loss mechanism in this case is believed to involve positively charged helium ions, He^+, formed by photoionization. These ions are thought to be accelerated by electric fields and ejected from the Earth at higher latitudes where magnetic field lines are, on occasion at least, open to space. Helium is a transitory constituent of the atmosphere. It originates in the Earth's crust and is quickly lost to space. Its concentration reflects a dynamic balance between the source from within and loss to the outside.

Composition of the present atmosphere

MAJOR COMPONENTS OF THE LOWER ATMOSPHERE

The atmosphere contains a bewildering array of gases, with relative abundances for important species ranging from 78 percent (N_2) to less than one part in 10^{12} (the hydroxyl radical, OH). The longer lived gases—*e.g.*, N_2 and O_2—are distributed more or less homogeneously around the Earth. The shorter lived species—*e.g.*, carbon monoxide, nitric oxide, and ozone—can vary considerably both in time and space. In many respects, the atmosphere can be considered an extension of the biosphere: almost all of the major constituents, with the exception of the noble gases, are either directly or indirectly under the influence of life.

The lower atmosphere as an extension of the biosphere

There are several natural linkages, as, for example, O_2, CO_2, CH_4, and H_2 (molecular hydrogen). Oxygen is a product of photosynthesis, summarized conveniently by the bulk reaction

$$h\nu + CO_2 + H_2O \rightarrow CH_2O + O_2. \quad (9)$$

The notation is qualitative rather than quantitative. The formula CH_2O denotes any of a variety of organic compounds formed in the primary life-giving photosynthetic event: the stoichiometry is approximately 1:2:1, C:H:O.

Aerobic respiration and decay involve the reverse of reaction (9),

$$CH_2O + O_2 \rightarrow CO_2 + H_2O + \text{energy}. \quad (10)$$

This reaction satisfies the energy needs of the human population and of most of the other higher animals. All life on Earth ultimately depends on the ability of plants to capture solar energy and to store this energy in the form of potential food, CH_2O. It is a two-way street. In the absence of reaction (10), carbon would accumulate in organic form and the fuel for photosynthesis, atmospheric CO_2, would be depleted. Bacteria play a major role in recycling carbon primarily by reactions analogous to (10).

Respiration and decay can proceed even when the supply of O_2 is limited. This can arise, for example, in the sediments of organic-rich swamps and in the stomachs of ruminants. The product of anaerobic decay is methane (CH_4) in this case. Oxidation of carbon may occur photochemically in the atmosphere, initiated by reaction with the OH radical: molecular hydrogen is a product of the oxidation of CH_4 and other hydrocarbons; like CH_4, H_2 is removed from the atmosphere mainly by reaction with OH.

Distribution of carbon, nitrogen, and oxygen compounds. *Carbon compounds.* The bulk of the Earth's volatile carbon resides in sediments, either as organic carbon or as a component of carbonate minerals such as calcite, $CaCO_3$. Carbon is carried to the sediments in detrital material. The fraction lost from the oceans during subsequent burial represents but a small fraction of the total net primary productivity—less than 1 percent.

The life cycle is efficient (see above Figure 1). Carbon atoms are exchanged back and forth between the atmosphere, biosphere, soils, and oceans. Even the sediments provide only a temporary, albeit in human terms long (100,000,000-year), residence for the restive atom. The atom returns to the atmosphere as sediments are uplifted and weathered. The transit time from weathering to eventual return to the sediments is about 100,000 years, most of this spent as a component of the bicarbonate ion, HCO_3^-, in solution in the deep sea.

On time scales of a few hundred years or longer, the abundance of CO_2 in the atmosphere is determined by the dynamics, chemistry, and biology of the oceans. Most of the carbon in the oceans is present in cold, relatively stagnant water at depth. It returns to the atmosphere in association with slow upwelling motion at low latitudes. As surface waters cool and sink at high latitude, they draw carbon from the atmosphere, roughly balancing the source at low latitude. Falling fecal material provides an additional important means for transporting carbon from the surface to the deep.

The dynamics of this complex exchange are just beginning to be understood. It is clear, though, that the level of atmospheric CO_2 is not immutable. Studies of gases trapped in polar ice indicate that CO_2 has fluctuated from about 200 to roughly 280 parts per million (ppm) over the past 100,000 years. Low levels of CO_2 are associated with cold, ice-age conditions at the surface; high values correspond to times when the climate was relatively warm during interglacials. It appears that this behaviour has persisted over at least the past 750,000 years.

It is against this background that assessments must be made of the impact of the recent change in the CO_2 level caused by the burning of fossil fuels. The level of CO_2 has risen since the Industrial Revolution from about 280 ppm in 1850 to approximately 350 ppm today. It is expected to climb to values in excess of 600 ppm by the early part of

the 21st century. This poses a double challenge. Is it possible to predict accurately the effects of an increasing level of CO_2 on climate? If so, can this knowledge be used to influence the course of action over the next few decades? Humankind has developed the ability to change the Earth on a global scale in a single lifetime. Yet, it remains to be seen whether scientific knowledge can develop apace.

Nitrogen compounds. In contrast to carbon, which finds its most stable home in sediments, most of the Earth's nitrogen is in the atmosphere as the relatively inert gas N_2. The N–N bond in N_2 is very strong and is not easily fractured in the atmosphere, except at high altitudes where the molecule is exposed to energetic ultraviolet radiation or at low altitudes where it may be raised to a temperature exceeding 2,000 K near a lightning stroke. The bond can be broken by biologically mediated reactions, however, by photosynthetic blue-green algae or by bacteria functioning in symbiosis with plants of the legume family. Dissociation of N_2 is essential for life: amino groups, represented by NH_2, are indispensable components of living tissue. The N–N bond must be broken before the atom can be incorporated in living organisms. The class of compounds containing odd numbers of N atoms is referred to collectively as fixed nitrogen. When the bond in N_2 is broken, the molecule is said to be fixed.

Fixed nitrogen

Fixed nitrogen occurs in a variety of oxidation states, as shown in Figure 8. Its importance for the biosphere may be attributed in no small measure to this aspect of its chemistry. Oxidation of ammonium, NH_4^+, represented by

$$NH_4^+ + \frac{3}{2}O_2 \rightarrow NO_2^- + H_2O + 2H^+, \quad (11)$$

involves a change in free energy of 51.8 kilocalories at 298 K and one atmosphere. It denotes the first step in the oxidation of nitrogen, primary nitrification, in which nitrogen is transformed from oxidation state −3 to +3. The energy liberated in reaction (11) is utilized by bacteria, *Nitrosomonas* and similar genera, which effect the reaction. A different group of bacteria, *Nitrobacter,* is able to oxidize nitrogen further, from nitrite (NO_2^-) to nitrate (NO_3^-),

$$NO_2^- + \frac{1}{2}O_2 \rightarrow NO_3^-, \quad (12)$$

a process known as secondary nitrification. Reaction (12) involves a release of free energy of 20.1 kilocalories per mole.

Plants can use the ammonium ion, NH_4^+, or NO_2^- or NO_3^- to satisfy their need for fixed nitrogen but generally tend to prefer NH_4^+. Nitrogen in NO_2^- and NO_3^- must be reduced from oxidation state +3 or +5 to −3 before it can be assimilated into living tissue. The reactions—in this case assimilatory reduction—are denoted by

$$NO_2^- + \frac{3}{2}CH_2O + 2H^+ \rightarrow NH_4^+ + \frac{3}{2}CO_2 + \frac{1}{2}H_2O, \quad (13)$$

and

$$NO_3^- + 2CH_2O + 2H^+ \rightarrow NH_4^+ + 2CO_2 + H_2O. \quad (14)$$

The necessary source of energy here is supplied by oxidation of organic matter, represented by CH_2O.

The atmosphere contains 4×10^{15} metric tons of nitrogen in the form of N_2. Approximately 10^8 metric tons are fixed annually by biologic agents, and about the same amount is transformed by various industrial processes associated with the combustion of fossil fuels and the manufacture of chemical fertilizer. In the absence of a source, the atmosphere would lose its reservoir of N_2 in about 2×10^7 years. The gas would be oxidized to NO_3^- and would tend to accumulate in the oceans. This obviously has not happened. There is a return of nitrogen from NO_3^- to N_2 through bacterially mediated denitrification, which is described by the reaction

Denitrification

$$NO_3^- + \frac{5}{4}CH_2O \rightarrow \frac{1}{2}N_2 + \frac{3}{4}H_2O + \frac{5}{4}CO_2 + OH^-. \quad (15)$$

Denitrification proceeds under anaerobic conditions (*i.e.,* in the absence of free oxygen) and provides another example of the splendid efficiency of the global life-support system. Waste for one species is opportunity for another. Reaction (15) is exothermic: it releases energy equivalent to 124 kilocalories per mole. It represents an essential link in the nitrogen cycle, providing a respiratory path for bacteria when O_2 is deficient and serving at the same time to maintain a relatively constant level of atmospheric N_2. The general features of the nitrogen cycle are shown in Figure 9.

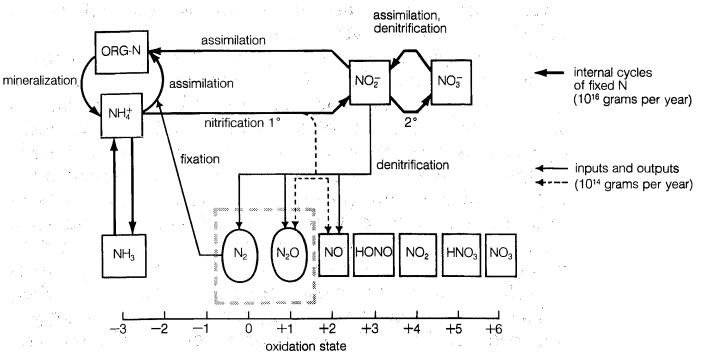


Figure 8: Major paths for oxidation and reduction of nitrogen. The influence of terrestrial life on atmospheric composition is apparent from the biologic transformations of nitrogen shown here.

Nitrous oxide, N_2O, is the second most abundant nitrogen constituent of the atmosphere, as indicated in Table 3. It is formed as a by-product of denitrification, by reduction of NO_3^- and by oxidation of NH_4^+, the first step in nitrification. Additional production of N_2O occurs during fossil fuel combustion. The lifetime of N_2O in the atmosphere is about 150 years; it is removed mainly by photolysis (chemical decomposition by the action of radiant energy) in the stratosphere. Given its relatively short lifetime, as compared with N_2, one might expect N_2O to vary, and indeed there is evidence for a contemporary increase in the concentration of the gas in the atmosphere. This apparent increase is attributed in part to the burning of fossil fuels and in part to the general anthropogenic disturbance of the global nitrogen cycle. The human role in nitrogen fixation is now comparable to that of nature. Other links in the nitrogen cycle might be expected to adjust accordingly. The atmosphere contains various forms of fixed nitrogen besides N_2 and N_2O. As mentioned above, fixed nitrogen is introduced into the air by lightning at a

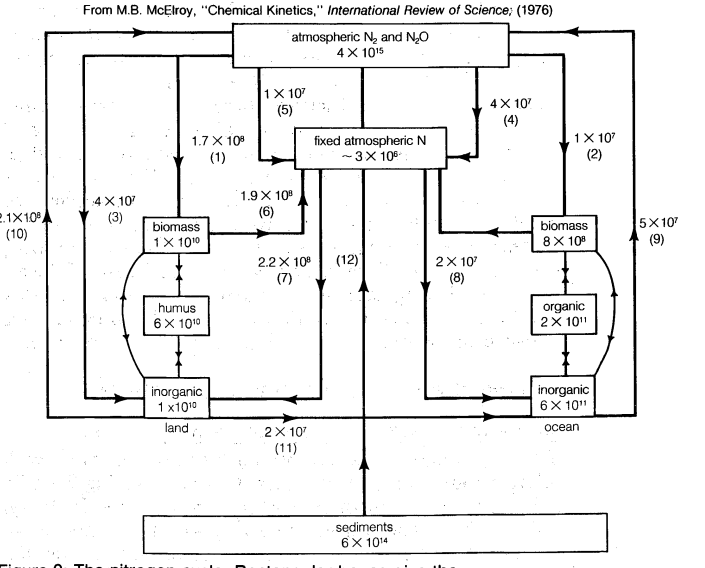


Figure 9: The nitrogen cycle. Rectangular boxes give the major reservoirs, summarizing the approximate contents of nitrogen in units of 10^{15} grams. Exchanges are indicated by lines with arrows. Exchange rates are given in units of 10^6 grams per year.

Table 3: Composition of the Atmosphere

species	relative abundance (parts per billion by volume)	source	comment
N_2	7.81×10^8	biologic	long lived
O_2	2.01×10^8	biologic	long lived
H_2O	10^6–10^7	physical	long lived
Ar	9.34×10^6	radiogenic	permanent
CO_2	3.5×10^5	biologic, industrial	variable, increasing
Ne	1.8×10^4	interior	permanent
He	5.2×10^3	radiogenic	escaping
CH_4	1.6×10^3	biologic	variable, increasing
Kr	1.0×10^3	interior	permanent
H_2	5.0×10^2	biologic, photochemical	variable
N_2O	3.0×10^2	biologic, industrial	increasing
CO	1.0×10^2	photochemical, industrial	variable, increasing
SO_2	$<10^2$	industrial, photochemical	variable
O_3	$<10^2$	photochemical	variable
Xe	9×10^1	interior	permanent
NO, NO_2, NO_x	variable	industrial, biologic	—*
CH_3Cl	6.0×10^{-1}	biologic	short lived
CCl_2F_2	2.9×10^{-1}	industrial	increasing
CCl_3F	1.7×10^{-1}	industrial	increasing
CCl_4	1.2×10^{-1}	industrial	increasing
CH_3CCl_3	9.8×10^{-2}	industrial	increasing
CF_4	7.0×10^{-2}	industrial	increasing
CH_3Br	1.0×10^{-2}	biologic, industrial	possibly increasing

*Relatively short lived, with an average lifetime of roughly one month.

rate of about 10^7 tons N yr^{-1} (metric tons of nitrogen per year). It also is released as a product of the combustion of fossil fuels and the burning of biomass. In addition, fixed nitrogen emanates from soil as a result of the microbial oxidation of NH_4^+, and it is formed in the stratosphere by a process involving the reaction of the metastable oxygen atom $O(^1D)$ with N_2O,

$$O(^1D) + N_2O \rightarrow NO + NO. \quad (16)$$

In all cases, the primary source of fixed nitrogen is nitric oxide, NO. About 5×10^7 tons N yr^{-1} of NO are introduced directly into the troposphere, much of this in continental areas, with a particularly large contribution from industrial sources and automobiles and trucks in cities. The stratospheric source is smaller, about 10^6 tons N yr^{-1}.

Reactions in the atmosphere result in a variety of secondary nitrogen species. Important compounds in remote regions include nitrogen dioxide (NO_2), the unstable intermediate nitrogen trioxide (NO_3), dinitrogen pentoxide (N_2O_5), nitrous acid (HNO_2), and nitric acid (HNO_3). Significant, too, is the set of reactive nitrogen compounds, $NO + NO_2 + NO_3 + N_2O_5 + HNO_2 + HNO_3 + NO_2NO_2$, designated NO_x. The subset $NO + NO_2$ plays a particularly important role in the photochemistry of ozone and is identified separately as NO_y.

The compound $CH_3CO.O_2NO_2$, known as peroxyacetyl nitrate, or PAN, is thought to provide an important storage reservoir for NO_x in the unpolluted troposphere. It decomposes thermally in the troposphere, with production of NO_2 and the peroxyacetyl radical, $CH_3CO.O_2$,

$$CH_3CO.O_2NO_2 \rightleftarrows CH_3CO.O_2 + NO_2. \quad (17)$$

Reaction (17) describes an equilibrium, which favours PAN at low temperatures and products at high temperatures. Conditions are such that PAN is relatively stable in the upper troposphere, but it decomposes rapidly in the warm environment near the ground. The compound is believed to form in urban air, a product of the complex set of reactions involved in the oxidation of hydrocarbons. From there, it can be exported via air motions to the upper troposphere, allowing urban air pollution to spread over much wider geographic areas. A typical reaction scheme for the production of PAN is presented in Table 4.

Formation of PAN in urban air

Fixed nitrogen is removed from the atmosphere by solution in rainwater or snow and by deposition on surfaces. The lifetime of NO_x is relatively short—a month or so on average. Investigators have observed considerable temporal and spatial variability in NO_x, with concentrations as high as 100 parts per billion (ppb) in polluted urban environments and as low as 50 parts per trillion (ppt) at remote locations, as, for instance, in the central Pacific. Fixed nitrogen is removed from the atmosphere mainly as HNO_3, contributing to the acidity of precipitation. This matter will be discussed at greater length below in connection with the phenomenon of acid rain.

Table 4: A Scheme for Oxidation of Alkanes Leading to Production of PAN*

$OH + RCH_3 \rightarrow H_2O + RCH_2$	(4.1)
$RCH_2 + O_2 \rightarrow RCH_2O_2$	(4.2)
$RCH_2O_2 + NO \rightarrow RCH_2O + NO_2$	(4.3)
$RCH_2O + O_2 \rightarrow RCHO + HO_2$	(4.4)
$OH + RCHO \rightarrow RCO + H_2O$	(4.5)
$RCO + O_2 \rightarrow RCO.O_2$	(4.6)
$RCO.O_2 + NO_2 \rightarrow RCO.O_2.NO_2$	(4.7)

*R denotes a typical methyl radical such as CH_3.

Carbon monoxide. Photochemical processes play a major role not only for NO_x but also for CO. Carbon monoxide is released to the atmosphere as a product of incomplete combustion. It also is formed as an intermediate in the oxidation of hydrocarbons, including methane. Present understanding of the path for decomposition of CH_4 is shown in Figure 10. The hydroxyl radical, OH, is central to this process. Reaction with OH is the first step in the removal of many hydrocarbons, including CH_4. Production of CO by oxidation of CH_4 proceeds at a rate proportional to the concentration of OH, as does the removal of CO, which is effected by the reaction

$$CO + OH \rightarrow CO_2 + H. \quad (18)$$

The radical OH is formed by the reaction of the metastable oxygen atom $O(^1D)$ with H_2O,

$$O(^1D) + H_2O \rightarrow OH + OH, \quad (19)$$

with $O(^1D)$ produced by photolysis of O_3,

$$h\nu + O_3 \rightarrow O(^1D) + O_2. \quad (20)$$

Reactions with CH_4 and CO are the primary paths for removal of OH.

Relationship between increases in carbon monoxide and methane concentrations

This leads to an interesting situation. An increase in carbon monoxide production due to fossil fuel combustion or biomass burning is expected to cause a decrease in OH concentration, with a resulting increase in the lifetime of gases normally removed by reaction with OH. As a consequence, one may expect an increase in CO concentration that is even larger than that due to growth in the magnitude of the source alone, with a related rise in the concentration of CH_4. The increases in the global abundances of CO and CH_4 that have been observed may be attributed, at least in part, to the complexities and nonlinearities of the chemistry affecting the concentration of OH.

Carbon monoxide has a lifetime in the atmosphere of about a month. The concentration in the Northern Hemisphere is larger than in the Southern Hemisphere by about a factor of 2. Surface concentrations at unpolluted sites range from 100 to 200 ppb and vary seasonally, with the highest values in winter when the OH concentration is low. The relative abundance, or mixing ratio, is large near the surface, decreasing aloft in the Northern Hemisphere, with an opposite trend in the south. Mixing ratios of CO in the Southern Hemisphere average about 60 ppb at the surface, climbing to roughly 70 ppb in the upper troposphere. Concentrations in cities reach values as high as one part per million and can pose a problem for public health from time to time.

Suite of sulfur compounds. In addition to the carbon, nitrogen, and oxygen compounds discussed above, the atmosphere contains a suite of sulfur-bearing gases. Like nitrogen, sulfur can occur in a variety of oxidation states. Compounds observed in the atmosphere range from the reduced gases carbonyl sulfide (COS), carbon disulfide (CS_2),

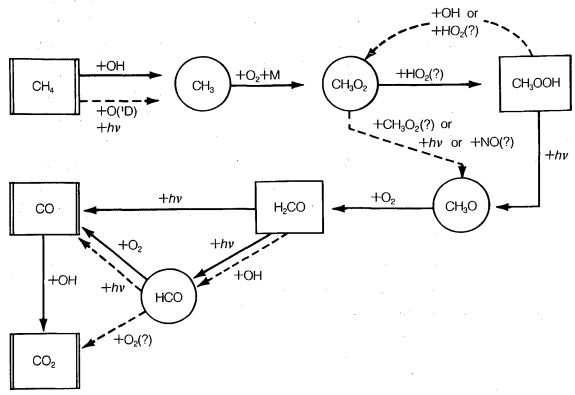


Figure 10: Oxidation path for methane, CH_4.
From M.B. McElroy, "Chemical Kinetics," *International Review of Science*; (1976)

hydrogen sulfide (H_2S), and dimethyl sulfide [$(CH)_2S$], where sulfur is present in the oxidation state +2, to the more oxidized form of sulfur dioxide (SO_2), where sulfur is in the oxidation state +4. The reduced species H_2S and $(CH_3)_2S$ are produced by reactions involving living organisms. Hydrogen sulfide originates mainly from anaerobic environments on land (*e.g.*, swamps and marshes). The ocean is an important source of $(CH_3)_2S$. The origin of COS and CS_2 is less clear; it may include contributions from a variety of sources, both natural and anthropogenic.

Reduced sulfur gases are oxidized at varying rates in the atmosphere and contribute to the global budget of SO_2. The first step in oxidation for most involves reaction with the OH radical. Lifetimes of H_2S and $(CH_3)_2S$ are relatively brief, and the distribution of these species in the atmosphere is correspondingly variable. Carbonyl sulfide is longer lived, and this species is distributed more or less uniformly with latitude, with a mixing ratio of about 500 ppt. The lifetime of carbon disulfide is intermediate; its mixing ratio varies from about 3 to 30 ppt.

Formation of sulfur dioxide

Sulfur dioxide is formed not only by oxidation of reduced gases in the atmosphere but also as a by-product of the combustion of coal and the smelting of ores such as copper. The anthropogenic component is comparable to the natural one on a global scale and is dominant over large areas of the industrially developed Northern Hemisphere. The lifetime of SO_2 is about a week. Sulfur in SO_2 is oxidized rapidly by both homogeneous and heterogeneous reactions and is removed from the atmosphere by precipitation and by dry deposition on surfaces, mainly as sulfuric acid (H_2SO_4) in which sulfur is present with oxidation state +6. The abundance of SO_2 is variable, as would be expected given its relatively short lifetime. On a global basis, the mixing ratio averages about 200 ppt.

Suite of noble gases. Table 3 includes, in addition to the carbon, hydrogen, nitrogen, oxygen, and sulfur species discussed above, a suite of noble gases, with mixing ratios ranging from almost 1 percent in the case of argon to about 1 ppb for xenon. There are two components to the Earth's noble gas inventory. Some of the noble gases are primitive in the sense that the corresponding elements were captured from the solar nebula when the Earth formed some 4,500,000,000 years ago. Others, notably helium-4 and argon-40, are of more recent origin, produced by the radioactive decay of such elements as uranium, thorium, and potassium in the Earth's crust and mantle. Noble gases are released from the interior of the planet at ocean-spreading centres (*i.e.*, the mid-oceanic ridges) when new crust is formed, as well as from volcanoes and hot springs. With the exception of helium, which escapes, the noble gases are inert and accumulate in the atmosphere. Studies of the elemental and isotopic composition of the atmospheric gases, in combination with analogous data for meteorites and other planets, provide valuable clues to the origin and early history of the Earth.

Concentrations of halogenated hydrocarbons. The atmosphere contains a variety of halogenated hydrocarbons as summarized in Table 3. These compounds are for the most part industrial in origin, with the exception of methyl chloride (CH_3Cl) and methyl bromide (CH_3Br), which are produced at least in part by marine organisms. Halogenated hydrocarbons are used as solvents, foam-blowing agents, fire extinguishers, fumigants, refrigerants, and propellants in aerosol spray dispensers. They are among the most ubiquitous products of modern industrial society.

Concentrations of several of the more widely used chlorinated species, CCl_2F_2 (CFC-12, or F-12), CCl_3F (CFC-11, or F-11), $CHClF_2$ (CFC-22), CH_3CCl_3 (methyl chloroform), and $C_2Cl_3F_3$ (CFC-113), have been increasing at compound rates of between 5 and 10 percent over the past several decades. The fully halogenated methanes, CCl_2F_2 and CCl_3F, are exceptionally stable, with lifetimes in the range of 60 to 130 years. They are removed mainly in the stratosphere by photolysis in the ultraviolet. They break down, releasing their constituent chlorine atoms to form reactive radicals, such as Cl and ClO. Chlorine radicals play an important role in the chemistry of stratospheric ozone (see below). There is concern that an increasing burden of stratospheric chlorine could result in a significant global reduction in the column abundance of ozone with an attendant increase in the transmission of ultraviolet solar radiation.

Formation of reactive chlorine radicals

Fluorine is relatively inert in the stratosphere. Atoms of fluorine formed by the decomposition of the halocarbons react with methane to form hydrogen fluoride (HF). In contrast to hydrogen chloride (HCl), which decomposes rapidly by reaction with the OH radical, HF is stable in the stratosphere. The concentration of reactive fluorine radicals is much less than that of chlorine. Consequently, fluorine plays a negligible role in the chemistry of stratospheric ozone.

The abundance of brominated halocarbons is much less than the abundance of the chlorinated species. Hydrogen bromine, however, is less stable than HCl. A relatively large fraction of the bromine in brominated compounds that decompose in the stratosphere ends up in the form of the radicals Br and BrO. It is thought that these species can contribute significantly to removal of ozone. They tend to amplify the effect of chlorine and may become even more significant in the future.

THE ROLE OF PHOTOCHEMISTRY

The preceding section dealt chiefly with gases that enter the atmosphere from below. As explained, the three primary sources for such gases are reactions effected by living organisms excluding man, reactions specifically attributable to human activity, and processes relating to the abiotic metabolism of the Earth. In this section, attention is focused on the role of photochemistry, the synthesis of gases in the atmosphere as a consequence of the interaction with sunlight. Ozone is by far the most important product of the photochemical process.

Ozone as a photochemical product

The bulk of the atmosphere's O_3 is found in the stratosphere at altitudes of about 25 kilometres. It is this reservoir that is primarily responsible for protecting the surface from otherwise harmful rays of ultraviolet sunlight. The abundance of O_3 in the troposphere also is environmentally significant. Direct exposure to high concentrations of O_3 can reduce productivity in plants. It also can have a variety of undesirable effects on public health, causing, for example, respiratory problems.

The processes regulating the abundance of O_3 in the stratosphere and troposphere are quite distinct. In the former case, production of O_3 is initiated by photolysis of O_2; in the latter, the O_2 bond is broken through a complex set of reactions associated with the oxidation of hydrocarbons in the presence of NO_x.

Stratospheric ozone. Studies of O_3 have a long history, dating to about 1930. Early research emphasized the chemistry of a hypothetical pure oxygen system. Four reactions were considered:

$$h\nu + O_2 \rightarrow O + O, \quad (21)$$

$$O + O_2 + M \rightarrow O_3 + M, \quad (22)$$

$$h\nu + O_3 \rightarrow O + O_2, \quad (23)$$

and

$$O + O_3 \rightarrow 2O_2. \quad (24)$$

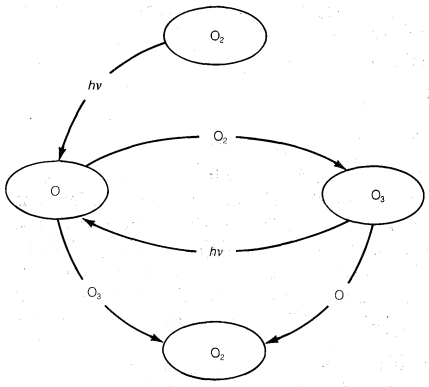


Figure 11: Schematic view of ozone chemistry in a pure oxygen environment.

Reactions (22) and (23) serve to establish the relative abundances of atomic oxygen and ozone. The oxygen atom formed in reaction (21) cycles many times through (22) and (23) before it is eventually removed by (24). It is thus convenient to consider O and O_3 in combination, to think of a family of odd oxygen compounds, $O + O_3$. Reaction (21) is the source of odd oxygen, and reaction (24) is the sink. In equilibrium, source and sink must balance. This simple scheme is illustrated in Figure 11.

Family of odd oxygen compounds

The concentration of O is controlled by reactions (22) and (23) at all levels below about 70 kilometres. Thus,

$$[O] = \frac{J_{23}[O_3]}{k_{22}[O_2][M]}. \tag{25}$$

Balancing (21) and (24), using (25), gives

$$[O_3] = \frac{J_{21}^{1/2} k_{22}^{1/2} f_{O_2} [M]^{3/2}}{k_{24}^{1/2} J_{23}^{1/2}}, \tag{26}$$

where J and k denote rate constants for (21)–(24) with the subscripts identifying the relevant reactions; f_{O_2} is the mixing ratio for O_2, $[O_2] = f_{O_2}[M]$. (Concentration is signified by the bracketing of symbols.)

The essential features of the photochemistry of stratospheric O_3 are summarized by reaction (26). The rate for photolysis of O_2, J_{21}, is vanishingly small at low altitude: photons with sufficient energy to dissociate O_2 are absorbed by O_2 and O_3 above 25 kilometres. Rate constants are more or less invariant with height at high altitude, and the abundance of O_3 decreases accordingly with altitude, as $[M]^{3/2}$ at high elevation. The concentration of O_3 is small at low altitude, limited by the small and declining value of J_{21}. It follows that O_3 must assume a maximum value at some intermediate level. In practice, the maximum is observed at a height of about 25 kilometres. Height variations of J_{21}, J_{23}, [M], [O], and [O_3] are shown schematically in Figure 12.

Much of the complexity of stratospheric chemistry, and indeed much of its recent history, is concerned with working out the paths for removal of odd oxygen besides reaction (24). For example, the pair of reactions

$$NO + O_3 \rightarrow NO_2 + O_2 \tag{27}$$
$$NO_2 + O \rightarrow NO + O_2 \tag{28}$$

have a net effect equivalent to (24). In the presence of nitric oxide, NO, the rate for (24) is faster than would be otherwise the case; in short, reaction (24) is catalyzed by NO. Similarly, (24) can be catalyzed by chlorine, Cl,

$$Cl + O_3 \rightarrow ClO + O_2 \tag{29}$$
$$ClO + O \rightarrow Cl + O_2, \tag{30}$$

by bromine, Br,

$$Br + O_3 \rightarrow BrO + O_2 \tag{31}$$
$$BrO + O \rightarrow Br + O_2, \tag{32}$$

or by the hydroxyl radical, OH,

$$OH + O_3 \rightarrow HO_2 + O_2 \tag{33}$$
$$HO_2 + O \rightarrow OH + O_2. \tag{34}$$

Catalytic paths occur for which there is no direct reaction analogue. The sequence

$$OH + O_3 \rightarrow HO_2 + O_2 \tag{35}$$
$$HO_2 + O_3 \rightarrow OH + O_2 + O_2 \tag{36}$$

is equivalent to

$$O_3 + O_3 \rightarrow 3O_2, \tag{37}$$

as is

$$Cl + O_3 \rightarrow ClO + O_2 \tag{38}$$
$$Br + O_3 \rightarrow BrO + O_2 \tag{39}$$
$$ClO + BrO \rightarrow Cl + Br + O_2. \tag{40}$$

The rate constant for (37) is vanishingly small, and this reaction has a negligible effect on stratospheric ozone. The catalytic paths, (35) + (36) in particular, can be important, however, and provide additional means for the removal of odd oxygen, especially at low altitude where the abundance of atomic oxygen is small.

It is convenient to extend the definition of odd oxygen to include NO_2, ClO, BrO, and HO_2 (hydroperoxyl). The catalog of sinks for odd oxygen expands then to encompass reactions (28), (30), (32), (34), and (40). Note that each of these loss reactions is responsible for the removal of two compounds of odd oxygen. Similarly, (21), the only source reaction, is responsible for two odd oxygen products.

Significance of the odd oxygen concept

The concept of odd oxygen is useful in that it provides a systematic scheme for classifying the several hundred reactions involved in a comprehensive model of stratospheric O_3. Only a few of these reactions are directly implicated in the removal of O_3. The balance is important mainly in regulating the levels of NO_2, ClO, BrO, and HO_2.

It is useful to associate these compounds with specialized families in addition to odd oxygen. The choice of association is to some extent arbitrary. The family of NO_x compounds mentioned earlier provides a natural link for NO_2. In similar fashion, ClO is identified with Cl_x, $Cl + ClO + ClNO_3 + HOCl + HCl$; BrO with Br_x, $Br + BrO + BrNO_3 + HBr$; and HO_2 with HO_x, $OH + HO_2 + H_2O_2$. It is important to note that a particular compound may belong to more than one family.

Reaction of the metastable oxygen atom $O(^1D)$ with nitrous oxide, N_2O,

$$O(^1D) + N_2O \rightarrow NO + NO, \tag{41}$$

is the primary source of stratospheric NO_x. Decomposition of the halocarbons provides the dominant source of Cl_x and Br_x, while reaction

$$O(^1D) + H_2O \rightarrow OH + OH \tag{42}$$

is the most important source of HO_x. Transport to the troposphere is the major means for removal of NO_x, Cl_x, and Br_x from the stratosphere. HO_x is removed in situ by

$$OH + HO_2 \rightarrow H_2O + O_2 \tag{43}$$

and by

$$OH + HNO_3 \rightarrow H_2O + NO_3. \tag{44}$$

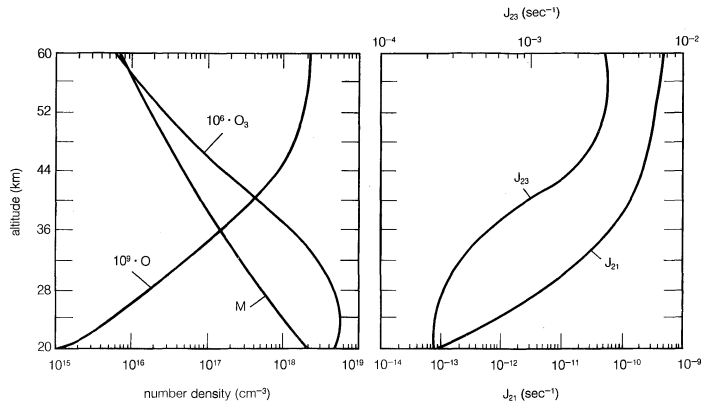


Figure 12: Height profiles for O_3, O, and M illustrating important terms in the equation for O_3, equation (26).

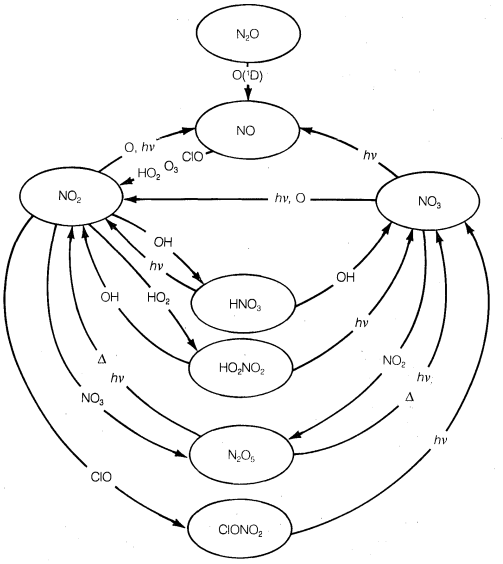


Figure 13: Transformations of nitrogen compounds in the atmosphere.

From *Atmospheric Ozone 1985*, World Meteorological Organization, Report No. 16, (1986), by courtesy of the National Aeronautics and Space Administration

Sources, sinks, and interchange mechanisms are illustrated schematically for NO_x, Cl_x, Br_x, and HO_x in Figures 13–16.

The distribution of O_3 with altitude and latitude is determined not only by chemistry but also by transport. The lifetime of odd oxygen, mainly O_3, is short at high altitude and low latitude where the abundance of atomic oxygen is relatively high. Chemistry is the dominant control in this case. The influence of transport is important at low altitude and high latitude; O_3 tends to accumulate in these regions where it is protected from chemical loss. This accounts for the gross features of the latitudinal distribution displayed in Figure 17. There is a net flow of O_3 from the stratosphere to the troposphere, where it is removed heterogeneously at the surface by gas phase reaction with HO_2.

Tropospheric ozone. Tropospheric O_3 is supplied either by in situ oxidation of hydrocarbons and CO or by downward transport from the stratosphere. The path for production due to oxidation of CO is particularly simple and may be summarized as follows:

$$CO + OH \rightarrow CO_2 + H, \quad (45)$$

$$H + O_2 + M \rightarrow HO_2 + M, \quad (46)$$

$$HO_2 + NO \rightarrow OH + NO_2, \quad (47)$$

$$h\nu + NO_2 \rightarrow NO + O, \quad (48)$$

and

$$O + O_2 + M \rightarrow O_3 + M. \quad (49)$$

Reactions (45)–(49) are equivalent to the bulk reaction

$$CO + 2O_2 \rightarrow CO_2 + O_3. \quad (50)$$

From *Atmospheric Ozone 1985*, World Meteorological Organization, Report No. 16, (1986), by courtesy of the National Aeronautics and Space Administration

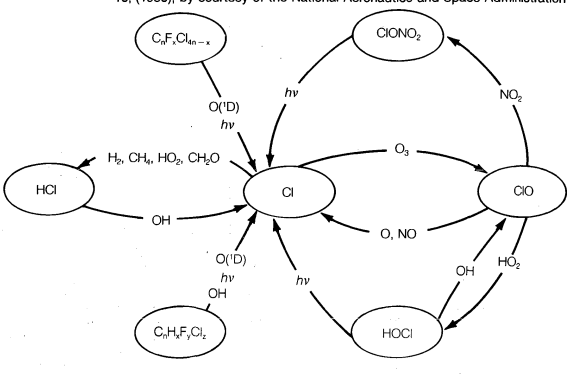


Figure 14: Transformations of chlorine compounds in the atmosphere.

An alternate path for oxidation of CO involves reactions (45) and (46) followed by

$$HO_2 + O_3 \rightarrow OH + 2O_2. \quad (51)$$

The bulk reaction in this case is

$$CO + O_3 \rightarrow CO_2 + O_2. \quad (52)$$

In the first path O_3 is produced, while in the second it is consumed. The choice of oxidation scheme depends on the relative abundance of NO (nitric oxide) and O_3, on the relative importance of reactions (47) and (51). Production of O_3 is favoured if the abundance of NO is high; loss is predominant if NO is low. Similar bifurcation occurs in the oxidation of CH_4 and other more complex hydrocarbons.

The chemistry of tropospheric O_3 is in general less well understood than the chemistry of stratospheric O_3, because the troposphere is intrinsically more complex than the stratosphere. Distributions of NO_x and hydrocarbons are quite inhomogeneous and, as yet, not well defined. Ozone can be produced in one region and consumed in another. In any case, further research is required to understand better the intricacies of tropospheric chemistry.

EFFECTS OF HUMAN ACTIVITY ON ATMOSPHERIC COMPOSITION AND THEIR RAMIFICATIONS

Change has been a fact of life for the Earth for most of its history. Changes in the past have been both fast and slow, regular and episodic, driven by factors internal to the planet and by forces from without. The evolution of an oxygen-mediating enzyme altered the composition of

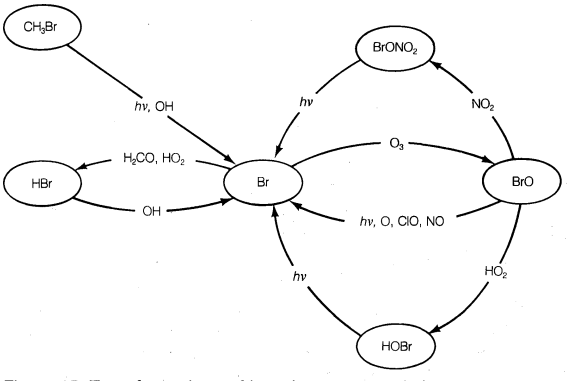


Figure 15: Transformations of bromine compounds in the atmosphere.

the atmosphere irreversibly some 2,000,000,000 years ago, allowing free oxygen to accumulate and thereby changing the nature of the biota for all subsequent time. The impact of a large asteroid (or possibly a comet), it is thought, led to the demise of the dinosaurs and a variety of other species at the end of the Cretaceous Period (from about 136,000,000 to 65,000,000 years ago). Climate has adjusted throughout geologic time to the changing position of the continents driven by the slow motion of giant lithospheric plates, rigid blocks of the Earth's surface rocks (see PLATE TECTONICS). Ice sheets have waxed and waned with regularity in response to the changing orbit and rotation axis of the Earth, and, as noted above, there have been associated changes in the carbon dioxide level of the atmosphere. What sets the present era apart, however, is the unique importance of a single species, man. Since about the mid-19th century, humankind has developed the capacity to harness the powers of nature, to mine the reserves of organic carbon laid down in sediments over millions of years. It has used this abundant source of energy to alter its way of life, with effects, for the most part inadvertent though undoubtedly large, for the global environment. Some of those effects associated with the atmosphere are considered here.

Climate modification. Trace gases, carbon dioxide and water vapour in particular, play an essential role in terrestrial climate. In the absence of these molecules that absorb strongly in the infrared, the surface temperature would be about 40 K colder than it is today. The oceans would be

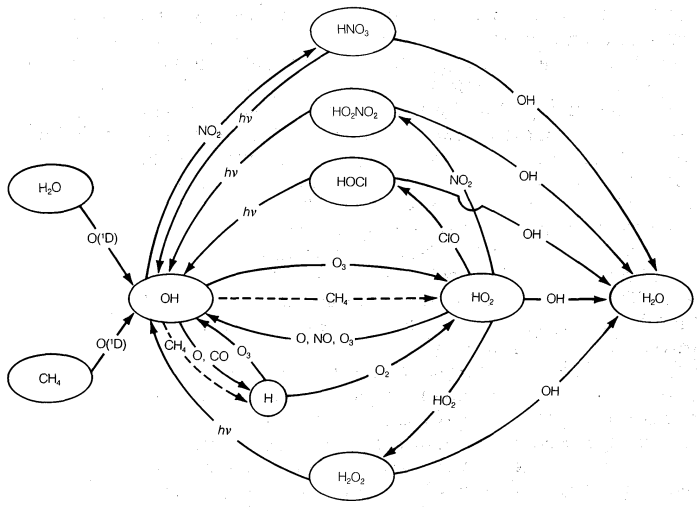


Figure 16: Transformations of hydrogen compounds in the atmosphere.

From *Atmospheric Ozone 1985*, World Meteorological Organization, Report No. 16, (1986), by courtesy of the National Aeronautics and Space Administration

frozen over, and life as we know it would be impossible. There is an important synergism between CO_2 and H_2O. Carbon dioxide itself absorbs only a small fraction of the heat radiated by the Earth's surface. Water vapour is much more significant. The abundance of H_2O, however, is controlled ultimately by temperature. An increase in CO_2 may be expected to cause an increase in temperature, which allows more water vapour to enter the atmosphere and leads to a further increase in temperature.

It has been suggested that the climate of the Earth could be ultimately unstable. Addition of gases capable of trapping heat could accelerate the release of H_2O and raise the temperature to a point where the oceans would evaporate, transforming the atmosphere into a hot steamy vapour. The surface pressure of the atmosphere would then be about 1,000 times higher than it is today, and its dominant constituent would be H_2O. Some believe that such changes may have occurred on Venus, with H_2O subsequently escaping from the top of its atmosphere. In any event, contemporary Venus is a striking example of the importance of the greenhouse effect. Its atmosphere contains a large concentration of CO_2, almost 10^6 times more than that of the Earth. Moreover, the Venusian surface temperature is much hotter than the Earth's—about 780 K—in spite of the fact that Venus absorbs less energy from the Sun because of its ubiquitous cloud cover and associated high albedo.

Increase in CO_2 concentration

The concentration of CO_2 in the Earth's atmosphere has risen steadily over the past 140 or so years, from about 280 parts per million in 1850 to about 350 parts per million. The change is due largely to the combustion of fossil fuels. Since the Industrial Revolution, nearly 1.5×10^{11} metric tons of organic carbon have been mined and consumed in the form of coal, oil, and natural gas. Carbon dioxide is the largest single waste product of modern society. The average person is responsible for the release of almost four tons of CO_2 each year; the amount is even larger in the developed countries. The total global emission in 1985 was close to 5×10^9 metric tons of C.

Approximately half of the carbon emitted since the Industrial Revolution persists in the atmosphere today. The balance is presumed to have made its way into the oceans or to have been incorporated into organic matter on land. Uptake of CO_2 by the oceans is limited by the supply of carbonate ions, CO_3^{2-}, in surface waters. The bulk reaction is

$$CO_2 + CO_3^{2-} + H_2O \rightarrow 2HCO_3^-. \quad (53)$$

Carbon dioxide is a weak acid, and inclusion of CO_2 in the oceans leads to a reduction in pH, transforming carbon from CO_3^{2-} to bicarbonate ions, HCO_3^-, and dissolved neutral carbon. There is a limit, however, to this switching. The total negative charge carried by dissolved carbon compounds is fixed by the alkalinity of the ocean waters:

$$[\text{Alk}] = [HCO_3^-] + 2[CO_3^{2-}]. \quad (54)$$

Alkalinity can be altered only by the addition of salts, which are supplied, for example, by the dissolution of calcite, $CaCO_3$, in sediments. The CO_3^{2-} content of waters at the ocean surface is small, and sustained uptake of CO_2 requires that CO_3^{2-} be supplied continuously to the surface. Over a 100-year period about 10 percent of the water in the oceans is exposed at the surface. If such is the case, only 30 percent of the carbon dioxide released to the atmosphere by burning fossil fuels can be incorporated into the sea.

A continuing rise in carbon dioxide is inevitable. If current estimates for the reserve of fossil fuels—about 4×10^{12} metric tons of carbon—are considered, and if it is assumed that half of this reserve is used up over the next 100 years, the level of CO_2 could exceed 1,000 parts per million. If one assumes more conservatively that the consumption of fossil fuels will double over the next 100 years, CO_2 may be expected to grow to about 600 parts per million—approximately twice the value in 1850.

Impact of other gases on the climate

Carbon dioxide is not the only gas with the ability to affect climate. The trace gases F-11, F-12, methane, and nitrous oxide are even more potent on a molecule-per-molecule basis. It has been estimated that a molecule of F-11 or F-12, for example, can produce warming equivalent to that caused by 10^4 molecules of CO_2. Perturbations to stratospheric ozone and tropospheric ozone also can have a significant impact. In short, a comprehensive theory of climate change must allow for all of these various changes in atmospheric composition.

Most assessments of climatic change are based on simple, one-dimensional energy-balance models. These models, it is hoped, will provide an estimate for the globally averaged change in surface temperature. Studies suggest that the rise in carbon dioxide since 1850 should have resulted in an increase in the contemporary surface temperature of about 0.4° C. The trace gases F-11, F-12, methane, nitrous oxide, ozone, and stratospheric water vapour are estimated to have added to heating by CO_2 an increment equivalent of an additional 0.16° C, for a net change of 0.56° C. A change of this magnitude is generally consistent with observation, as shown in Figure 18.

From *Atmospheric Ozone 1985*, World Meteorological Organization, Report No. 16, (1986), by courtesy of the National Aeronautics and Space Administration

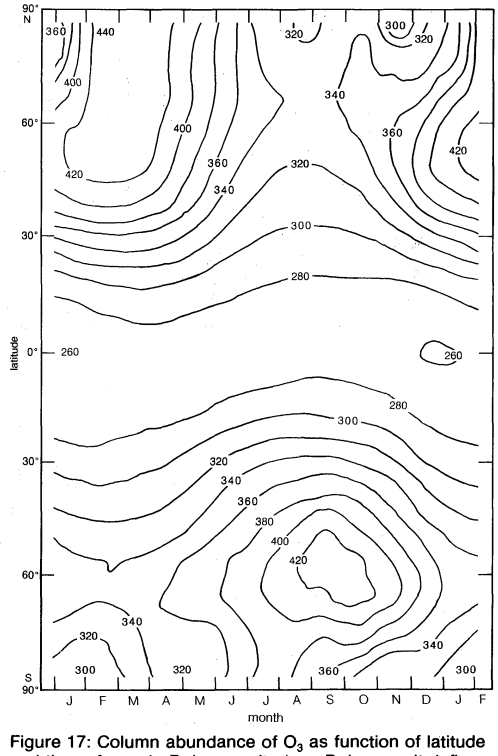


Figure 17: Column abundance of O_3 as function of latitude and time of year in Dobson units (one Dobson unit defines the content of O_3 in a column of one-millimetre length at standard air pressure and temperature).

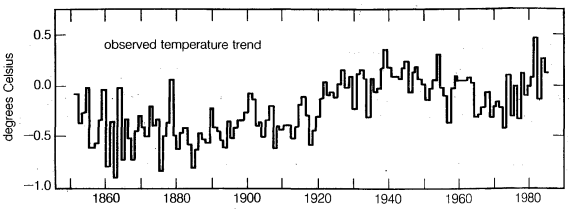


Figure 18: Change in global mean temperature since 1850.
From *Atmospheric Ozone 1985*, World Meteorological Organization, Report No. 16, (1986), by courtesy of the National Aeronautics and Space Administration

The effects of trace gases are expected to grow relative to CO_2 in the years ahead. Results from one particular simulation are given in Figure 19. In this case, CO_2 was predicted to cause an increase in surface temperature of about 0.9° C by the year 2030. With other trace gases included, it was estimated that the temperature increase could be as large as 2.2° C. A change of this magnitude would be unprecedented in recent Earth history.

A prediction of an increase in global mean temperature, whatever its magnitude, has at best limited use for government policymakers. Scientific concern is more with the geographic and temporal distribution of the change. Will it rain more or less, and when? How will the change in climate affect food production? What changes are to be expected in the location of deserts, forests, grasslands, and agricultural zones? Benefits may be expected for some and problems for others. Comprehensive assessment requires a model incorporating all of the important feedbacks between the atmosphere, oceans, and biosphere, with spatial resolution sufficient to distinguish at least the major biomes. Such a model is as yet unavailable, though efforts to develop it are under way. Preliminary studies with three-dimensional models similar to those used for predicting weather suggest that climatic zones can be expected to shift to higher latitudes on a warmer Earth. Further work is required, however, before models of this kind can be used with confidence for quantitative planning.

Depletion of stratospheric ozone. The importance of stratospheric O_3 has been recognized in a general way for almost 50 years. In the absence of O_3, the surface of the Earth would be exposed to lethal ultraviolet radiation with wavelengths as short as 240 nanometres. It was only in 1970, however, that scientists began to focus on the fact that even small changes in O_3 can have a significant impact on humans. Investigators observed that migration of people to lower latitudes—the shift in population from the northeastern part of the United States to the Sun Belt (roughly the southern and southwestern regions of the country), for example—was accompanied by an alarming rise in the incidence of skin cancer. Not all of this increase could be attributed to enhanced sunlight. There appeared to be an underlying factor to the smaller abundance of O_3 at lower latitudes and the associated increase in exposure of fair-skinned people to solar radiation with wavelengths near 300 nanometres. Epidemiological studies suggested that the incidence of skin cancer would rise by about $3x$ percent for every x percent decrease in the column density of O_3. This led to inevitable questions concerning the stability of O_3 as the human influence began to extend upward through the tropopause to the stratosphere.

Shield against shortwave ultraviolet solar radiation

It has long been known that the stratosphere turns over very slowly. Debris from the testing of nuclear bombs in the late 1950s and early 1960s was readily detectable a decade later. Plans were under way in the early 1970s to develop a commercial fleet of supersonic aircraft. These planes were projected to cruise at altitudes of about 20 kilometres. It seemed inevitable that gases from their exhaust would accumulate in the stratosphere, and nitric oxide became a particular concern. It was suggested that nitric oxide from a fleet of 500 supersonic aircraft could lead to a reduction in ozone abundance by as much as 3 percent. This led to a major research program coordinated by the U.S. Department of Transportation, and results from the program played an instrumental role in the decision by the federal government to suspend funds for the development of the large supersonic transport (SST). The British and French, however, continued their work on a smaller version of a supersonic transport, the Concorde, which was eventually introduced for limited service from Europe to North America and the Middle East. Flying at a lower altitude than the SST would have and releasing smaller quantities of nitric oxide, the Concorde has had a negligible effect on stratospheric ozone. It remains as a reminder of a vitriolic debate that served, if nothing else, to draw attention to the stratosphere, to highlight the potential vulnerability of even the most remote regions of the atmospheric environment.

The debate concerning the environmental impact of the SST has had a lasting effect on the development of atmospheric science, spawning a new interdisciplinary program of research linking chemists, physicists, and biologists in a common effort to understand the stratosphere. The program, with international participation, has been remarkably successful and has led to a new view of the interdependence of the atmosphere, hydrosphere, and biosphere.

The concern over the effects of exhaust gases from supersonic aircraft was soon followed by a new issue: the possibility that chlorine atoms released by decomposition of chlorofluorocarbons could have a larger and more persistent effect on stratospheric ozone. CFC's were developed first in the 1930s but found widespread use only in the years following World War II. They were employed with great success by U.S. troops in the Pacific to dispense insecticides from aerosol spray cans. This led to many commercial uses, from propellants and refrigerants to foaming agents and degreasers and a host of other applications. Moreover, the use of CFC's rapidly spread from the United States to Europe and the Far East. All this changed in 1975, when it was recognized that the release of CFC's to the atmosphere could pose a serious problem for stratospheric ozone. Production of F-12 declined from a peak of about 4.5×10^5 metric tons in 1975 to about 3.4×10^5 metric tons in 1982. A similar drop was registered for F-11.

Much of the work undertaken since the mid-1970s has focused on the effects of CFC's on the assumption that the composition of the atmosphere was otherwise constant. It has become clear, however, that the response of ozone depends not simply on the abundance of CFC's but also on the abundances of methane, nitrous oxide, and carbon monoxide. These species, too, are changing. Current mod-

From *Atmospheric Ozone 1985*, World Meteorological Organization, Report No. 16, (1986), by courtesy of the National Aeronautics and Space Administration

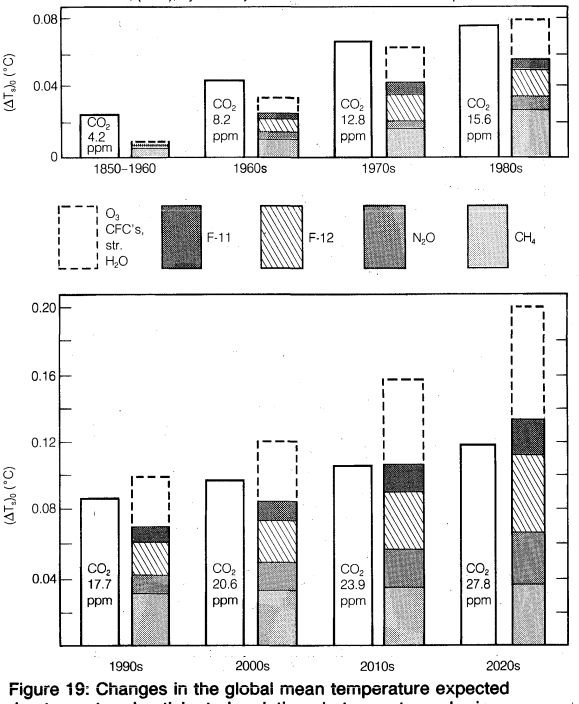


Figure 19: Changes in the global mean temperature expected due to past and anticipated variations in trace atmospheric gases. (Top) Increases in greenhouse forcing per decade from 1850 through the mid-1980s. (Bottom) Decadal increments of greenhouse forcing projected for the period from 1990 to 2030.

els suggest that a continuing release of CFC's at the rate registered in 1980, other gases remaining constant, would lead to a reduction in stratospheric ozone by about 5 percent. Maximum impact is predicted to occur at altitudes above 25 kilometres. An increase in nitrous oxide of 20 percent is expected to cause a reduction in ozone of about 2 percent. An increase in carbon dioxide should lead to a reduction in stratospheric temperatures with a consequent reduction in the anticipated impact of CFC's and nitrous oxide on ozone. The effect of an increasing burden of methane is more complex. Oxidation of methane provides a source of ozone at low altitude, while the reaction of chlorine with methane converts chlorine radicals to hydrogen chloride, resulting in a reduction in the impact of CFC's between 30 and 40 kilometres.

Models suggest that the change in the column density of stratospheric ozone to date should be relatively small. Reductions in ozone at high altitude, near 40 kilometres, ought to be balanced by excess production at low altitudes due in part to the higher level of methane and in part to NO_x released by high-altitude aircraft. Observational evidence is consistent with this view. A statistical analysis concluded that the change in the ozone column from 1970 to 1983 averaged −0.003 percent per decade. It showed, however, that a small though statistically significant drop in ozone—a decline of about 2 percent—occurred at altitudes above 30 kilometres between 1970 and 1980.

Recurrent decrease of ozone concentration over Antarctica

There has been a new development since 1985. That year, Joseph C. Farman and his associates at the British Antarctic Survey reported that the level of ozone over Antarctica had dropped precipitously every October since 1982, with the first such change apparent as early as 1978. A number of theories, or more properly hypotheses, have been advanced to account for this phenomenon. Several implicate effects of anthropogenic chlorine, enhanced by small quantities of bromine. Others suggest that the reduction may be due to a diminished supply of ozone from low latitudes, reflective of a change in stratospheric dynamics. In any case, the phenomenon was quite unexpected and serves as a powerful warning that current scientific understanding of the stratosphere is still rudimentary.

Acid rain and allied problems. The acidity of rain, or of any liquid, depends on the concentration of protons (positively charged hydrogen ions), H^+, in solution. It is measured conventionally by pH, defined by the relation

$$H = -\log_{10} [H^+], \tag{55}$$

where $[H^+]$ denotes the molar abundance of H^+—*i.e.*, the concentration expressed in moles per litre. Pure water is characterized by a pH of 7.0. A liquid is said to be acidic if pH is less than 7.0; otherwise it is considered neutral or alkaline.

Exposed to an atmosphere containing carbon dioxide, water absorbs carbon, releasing protons by

$$CO_2 + H_2O \rightarrow HCO_3^- + H^+. \tag{56}$$

Similar reduction in pH occurs with absorption of nitric acid (HNO_3) or sulfuric acid (H_2SO_4). Reactions in this case may be represented by

$$HNO_3 \rightarrow H^+ + NO_3^- \tag{57}$$

and

$$H_2SO_4 \rightarrow 2H^+ + SO_4^{2-}. \tag{58}$$

Absorption of less oxidized gases such as nitrogen dioxide (NO_2) or sulfur dioxide (SO_2) may be described by

$$4NO_2 + O_2 + 2H_2O \rightarrow 4NO_3^- + 4H^+ \tag{59}$$

and

$$2SO_2 + O_2 + 2H_2O \rightarrow 2SO_4^{2-} + 4H^+. \tag{60}$$

In each case, protons are liberated and pH declines. This result may be viewed as an inevitable consequence of the chemical nature of carbon, nitrogen, and sulfur. These elements are poised to form negative ions in their oxidized states. The resulting negative charge must be balanced, to some extent at least, by production of H^+.

The pH of rain in equilibrium with atmospheric CO_2 would have a value of 5.65. The addition of HNO_3 SO_2 with concentrations typical of the clean environment remote from concentrated sources, about 100 parts per trillion, would lower the pH to approximately 5.3. (This assumes lifetimes for nitric acid and sulfur dioxide of roughly 10 days with precipitation at an average rate of one metre per year.) A concentration of sea salt or alkaline aerosol equivalent to about 50 parts per million, a typical value, would raise the pH to about 6.0. On this basis, it is concluded that the natural range of pH should lie between 5.3 and about 6.0. The data in Table 5 are consistent with this view, except for the measurement in Brazil indicating a value of 4.6. The acid content of rain in the Amazon River basin may reflect the importance of local biologic sources of sulfur, particularly hydrogen sulfide (H_2S). As mentioned above, hydrogen sulfide is oxidized readily in the atmosphere to sulfur dioxide, with a lifetime of no more than a few days.

Table 5: The pH of Rain at a Number of Representative Remote Locations

place	pH
New Zealand	5.5 ± 0.5
Pago Pago	5.3 ± 0.3
Greenland	5.8 ± 0.65
American Samoa	5.3 ± 0.36
Hawaii	5.0 ± 0.4
Amazon	4.6 ± 0.2
El Salvador	5.2 ± 0.4

Rains in the northeastern part of the United States show median values for pH in the range of 4.1 to about 4.5. Values are lowest in a band extending from the Ohio River Valley to upstate regions of New York. Information on the composition of precipitation is particularly illuminating. Table 6 shows data from sites in Minnesota, Georgia, and New York obtained in 1979. Sulfate accounts for about 60 percent of the negative charge in rains at all three sites; nitrate ranges from 20 percent in Georgia to 34 percent in New York. The pH in New York was 4.34, as compared with the high value, 6.31, measured in Minnesota. The difference may be attributed to the presence of alkaline aerosol in rain reaching the site at Minnesota. The addition of calcium carbonate, $CaCO_3$, results in removal of H^+ according to

$$CaCO_3 + H^+ \rightarrow Ca^{2+} + HCO_3^-. \tag{61}$$

Soils in the western part of the United States are generally alkaline, and the aerosol content of the atmosphere tends to reflect this condition. Soils in the eastern United States are more acidic, and the buffering capacity is reduced accordingly.

The low pH and high sulfate content of precipitation in the Northeast is scarcely surprising. Approximately 12,000,000 metric tons of sulfur were vented to the atmosphere from industrial sources east of the Mississippi River in 1978. The sulfur content of rain falling on the same region amounted to only 20 percent of the quantity released. The balance was either exported or deposited by dry processes on various surfaces—vegetation and so forth.

Table 6: Average Ion Concentrations* Observed in Precipitation at Three U.S. Sites (in 1979)

ion	west central Georgia	southwest Minnesota	northwest New York
SO_4^-	38.9	45.8	44.8
NO_3^-	11.6	24.2	25.0
Cl^-	8.2	4.2	4.2
HCO_3^-	0.3	10.3	0.1
Anions	59.0	84.5	74.1
NH_4^+	5.5	37.7	8.3
Ca^{2+}	5.0	28.9	6.5
Mg^{2+}	2.4	6.1	1.9
K^+	0.7	2.0	0.4
Na^+	17.6	13.7	4.9
H^+	17.8	0.5	45.7
Cations	49.3	88.9	67.7
pH	4.75	6.31	4.34

*Microequivalents per litre.

The industrial source grew by a factor of 2.5 between 1950 and 1970, with the rate of growth most rapid during the 1950s. The decline in subsequent growth reflected a shift by utilities from coal to oil and nuclear power and from high-sulfur to low-sulfur coal, a trend that has been reversed somewhat since the early 1980s. The global industrial output of sulfur in 1978 was about 65,000,000 metric tons, which may be compared with estimates of 64,000,000 metric tons per year for production by the unperturbed global biosphere. The biospheric production in the eastern United States is about 100,000 metric tons per year and is surely less than the industrial contribution by at least a factor of 10.

Industrial sources of oxidized nitrogen are similarly dominant. The global industrial output of NO_x (see above) is about 20,000,000 metric tons N per year, distributed more or less equally between North America, Europe, and the rest of the world. The industrial output is thought to exceed production by natural agents in North America by a factor of from 3 to 13. Combustion of fossil fuels accounts for about 40 percent of the NO_x. Biomass burning accounts for another 25 percent, with the remainder from lightning and microbial processes in soils.

The relation between emission of potentially acidic compounds and ultimate delivery of acids to specific geographic regions is by no means straightforward. It depends on the nature of the atmospheric circulation and characteristics of the source—*e.g.,* height of smokestacks. Chemical factors controlling dry deposition are poorly understood, and there are gaps in scientific understanding of the relevant aspects of atmospheric chemistry. Oxidation of nitrogen dioxide to nitric acid is thought to proceed by the reaction

$$OH + NO_2 + M \rightarrow HNO_3 + M. \quad (62)$$

Oxidation of sulfur is initiated similarly by reaction with the hydroxyl radical, OH,

$$OH + SO_2 + M \rightarrow HSO_3 + M, \quad (63)$$

or it may proceed heterogeneously with hydrogen peroxide, H_2O_2, formed from HO_2. Ammonia, NH_3, is important in neutralizing acid aerosols. A complete description of the relevant chemistry must thus include NH_3 and O_3 in addition to oxides of nitrogen and sulfur. Interconnections between various chemical cycles can lead to subtle, unexpected nonlinearities in the composite effects, with important consequences for the geographic pattern of the resulting deposition of acid.

Acids can affect receiving systems in a variety of ways, some good, others bad. Addition of nitrate to ecosystems deficient in nitrogen can lead to enhanced biologic productivity. The benefits, however, may be offset by toxic effects associated with the release of aluminum and magnesium from clay minerals if the pH of soils falls below about 5.5. Leaching of metals from soils and sediments also can have adverse effects on aquatic systems. Fish may be killed by aluminum even under conditions where the pH is considered otherwise benign. Given the current rates of acid deposition, the pH of a number of lakes in the northeastern United States and in southeastern Canada is expected to average 4.2 to 4.8, with occasional depressions to near 3.8 during periods of snowmelt or heavy rain. Most fish species and almost all mollusks perish if the pH falls below 4.8; planktonic communities are simplified in this case, dominated by a few acid-tolerant taxa. Adverse effects of acid precipitation are most obvious for lakes, and it is not surprising that this aspect of the issue has received the most attention in the press. The effects on the terrestrial biosphere are more ambiguous. It appears that elevated levels of ozone in the troposphere may pose a more serious hazard for plant life even at locations removed from major industrial sites. Effects in this case, for a variety of agronomic and forest species, include foliar injury and significant reduction in growth and yield.

Toxic effects of metals leached from soils and sediments

In summary, there are deficiencies in current scientific understanding of acid rain, both in its atmospheric and biospheric dimensions. Uncertainties can be removed or reduced by a suitably directed research program. It is clear that the chemistry of the troposphere, both polluted and clean, should receive attention in this effort.

The ionosphere and phenomena of the upper atmosphere

THE IONOSPHERE

General characteristics. The bulk of the atmosphere consists of electrically neutral atoms and molecules. At high altitudes, however, a significant fraction of the atmosphere is electrically charged. This region is generally called the ionosphere. It extends throughout the mesosphere and thermosphere but is most important and distinct at altitudes above about 80 kilometres. The name was introduced during the 1920s and was formally defined in 1950 by a committee of the Institute of Radio Engineers. Members of the committee identified the region as "the part of the earth's upper atmosphere where ions and electrons are present in quantities sufficient to affect the propagation of radio waves."

Much of the early research on the ionosphere was carried out by radio engineers and was stimulated by the need to define the factors influencing long-range radio communication. Priorities have changed in recent years. Today, the need is to understand the ionosphere as the environment for Earth-orbiting satellites and ballistic missiles. The emphasis is on processes. Scientific knowledge of the ionosphere has grown tremendously, fueled by a steady stream of data from spacecraft-borne instruments and enhanced by measurements of relevant atomic and molecular processes in the laboratory.

Division of the ionosphere into layers

Historically, the ionosphere was thought to be composed of a number of relatively distinct layers. The most important layers were identified by the letters D, E, and F, with the F layer subsequently divided into regions F_1 and F_2. The nomenclature is somewhat peculiar. It appears that Edward V. Appleton of Britain, a pioneer in early radio probing of the ionosphere, was accustomed to using the symbol E to describe the electric field of the wave reflected from the first layer of the ionosphere he studied. Later, Appleton identified a second layer at higher altitude and used the symbol F in this case. Suspecting a layer at lower altitude, he adopted the additional symbol D. In time the letters came to be associated with the layers themselves rather than with the field of the reflected waves. It is now known that Appleton's layers are not particularly distinct. The electron density increases more or less uniformly with altitude from the D region, reaching a maximum in the F_2 region. In spite of the peculiarities of the original rationale for the naming of the layers, the nomenclature employed to describe the different regions of the ionosphere continues in wide use. Though the names persist, the definitions have evolved to reflect present-day understanding of the underlying physics and chemistry.

Ionospheric physics and chemistry. Most of the ionization in the ionosphere is effected by photoionization. Photons of short wavelength (*i.e.,* high energy) are absorbed by atmospheric gases. A portion of the energy is used to eject an electron, converting a neutral atom or molecule to a pair of charged species: an electron, which is negatively charged, and a companion positive ion. Ionization in the F_1 region is produced mainly by ejection of electrons from O_2, O, and N_2. The threshold for ionization of O_2 corresponds to a wavelength of 102.7 nanometres. Thresholds for O and N_2 are at 91.1 and 79.6 nanometres, respectively.

Positive ions can react with neutral gases and change their identity. There is a tendency for these reactions to favour production of more stable ions. Thus, ionized oxygen, O^+, can react with O_2 and N_2, resulting in ionized molecular oxygen, O_2^+, and ionized nitric oxide, NO^+:

$$\begin{aligned} O^+ + O_2 &\rightarrow O + O_2^+ \\ O^+ + N_2 &\rightarrow NO^+ + N. \end{aligned} \quad (64)$$

Similarly, ionized molecular nitrogen, N_2^+, can react with O and O_2, forming NO^+ and O_2^+:

$$\begin{aligned} N_2^+ + O &\rightarrow NO^+ + N \\ N_2^+ + O_2 &\rightarrow N_2 + O_2^+. \end{aligned} \quad (65)$$

The most stable, and consequently most abundant, ions in the E and F_1 regions are O_2^+ and NO^+, the latter more so than the former. At lower altitude, O_2^+ can be converted

to NO^+ by reactions with the minor species N (nitrogen) and NO (nitric oxide),

$$O_2^+ + N \rightarrow O + NO^+$$
$$O_2^+ + NO \rightarrow O_2 + NO^+. \quad (66)$$

In the D region, NO^+ can be converted to H_3O^+, the hydronium ion, and to companion species such as $H_5O_2^+$ and $H_7O_4^+$, formed by the addition of H_2O. Production of hydrated ions is limited by the availability of H_2O. As a consequence, they are confined to altitudes below about 85 kilometres.

The electron density in the D, E, and F_1 regions reflects for the most part a local balance between production and loss. Electrons are removed mainly by dissociative recombination, a process in which electrons attach to positively charged molecular ions and form highly energetic, unstable neutral molecules. These molecules decompose spontaneously, converting internal energy to kinetic energy of fragments. The most important processes in the ionosphere involve recombination of O_2^+ and NO^+. The reactions in this case may be summarized as follows:

$$O_2^+ + e \rightarrow O + O$$
$$NO^+ + e \rightarrow N + O. \quad (67)$$

A portion of the energy released in reaction (67) may appear as internal excitation of either nitrogen, oxygen, or both. The excited atoms can radiate, emitting faint visible light in the green and red regions of the spectrum and thereby contribute to the phenomenon of the airglow.

Airglow

The airglow originates principally from altitudes above 80 kilometres and is responsible for the diffuse background light that makes it possible to distinguish objects at the Earth's surface on dark, moonless nights. Airglow is produced for the most part by reactions involved in the recombination of molecular oxygen. The contribution from reaction (67) is readily detectable, however, and provides a useful technique with which to observe changes in the ionosphere from the ground. Studies of the airglow have a long and checkered history in atmospheric science. Over the years they have contributed significantly to scientific understanding of processes in the upper atmosphere.

As indicated above, dissociative recombination provides an effective path for the removal of molecular ions. There is no comparable means for the removal of atomic ions. Direct recombination of O^+ with an electron requires that the excess energy be radiated as light. Radiative recombination is inefficient, however, compared with dissociative recombination and plays a small role in the removal of ionospheric electrons. There is a complication at high altitudes where atomic oxygen is the major constituent of the neutral atmosphere and where electrons are produced primarily by photoionization of O. The atomic ion O^+ may be converted to NO^+ and O_2^+ through reactions with N_2 and O_2, but the abundances of N_2 and O_2 decline relative to O as a function of increasing altitude. In the absence of competing reactions, the concentration of O^+ and the density of electrons would increase steadily with altitude, paralleling the rise in the relative abundance of O. This occurs to some extent but is limited eventually by vertical transport.

Ions and electrons produced at high altitude are free to diffuse downward, guided by the Earth's magnetic field. The lifetime of O^+ is long at high altitude where the densities of O_2 and N_2 are very small. As ions move downward, the densities of O_2 and N_2 increase. Eventually the time constant for reaction of O^+ with O_2 and N_2 becomes comparable to the time for diffusion: O^+ is converted to either O_2^+ or NO^+ before it can move much farther. The O^+ density exhibits a maximum in this region. Competition between chemistry and transport is responsible for the formation of an electron-density maximum in the F_2 layer. The dominant positive ion is O^+.

The density of O^+ decreases with decreasing altitude below the peak, reflecting a balance between production by photoionization of O and removal by reaction (65). The density of O^+ also decreases above the peak. In this case, the removal of photo-ions is regulated by downward diffusion rather than by chemistry. The distribution of O^+ with altitude above the peak reflects a balance of forces, a pressure gradient that acts to support O^+ in opposition to gravitational and electrostatic forces that combine to pull O^+ down. The electrostatic force is set up to preserve charge neutrality. In its absence, the density of ions, which are much more massive than electrons, would tend to fall off more rapidly with altitude than that of electrons. The abundance of electrons would quickly exceed the density of ions, and the atmosphere would accumulate negative charge. The electric field redresses the imbalance, drawing electrons down and providing additional upward support for positively charged ions. The abundance of O^+ falls with increasing altitude as though O^+ had a mass of 8 atomic units rather than 16 atomic units (the electric field exerts a force equivalent to the gravitational force on a body of mass 8 atomic units, directed upward for ions and downward for electrons). The density of electrons falls off with altitude at precisely the same rate as O^+, preserving the balance of positive and negative charge.

Ionization at any given level depends on three factors: (1) the availability of photons of a wavelength capable of effecting ionization, (2) a supply of atoms and molecules necessary to intercept this radiation, and (3) the efficiency with which the atoms and molecules are able to do so. The efficiency is relatively large for O, O_2, and N_2 from about 10 to 80 nm. This is the portion of the spectrum responsible for the production of electrons and ions in the F_1 region. Photons between 90 and 100 nm are absorbed only by O_2. They therefore penetrate deeper and are responsible for producing about half the ionization in the E layer. The balance is derived from soft X rays (those of longer wavelength), which are absorbed with relatively low efficiency in the F region and so are able to penetrate to altitudes of about 120 kilometres under conditions of high solar elevation. Hard X rays (those of shorter wavelength), notably below about 5 nm, reach even deeper. This portion of the spectrum accounts for the bulk of the ionization in the D region, with an additional contribution from wavelengths longer than 102.6 nm—mainly from photons in the strong solar emission line at Lyman α, 121.6 nm. Lyman α emission is absorbed weakly by the major components of the atmosphere, O, O_2, and N_2. It is absorbed readily by NO and has sufficient energy to ionize this relatively unstable compound. In spite of the low abundance of NO, the high flux of solar radiation at Lyman α is able to provide a significant source of ionization for the D region near 90 kilometres.

Ionospheric variations. The ionosphere is variable in space and time. Some of the changes are chemical in origin and can be readily understood on the basis of the general considerations outlined above. There is a systematic variation, for example, as a function of time of day. In the early morning the Sun is relatively low in the sky. Radiation must penetrate a large column of air before reaching a given level of the atmosphere. Ionization rates are less than at noon, and the profile of ionization is shifted to higher altitudes. The heights of the D, E, and F_1 layers change in response to solar elevation. The layers are lowest and the densities of electrons are highest at noon. Ionization in the D, E, and F_1 regions tends to disappear at night as a result of recombination in the absence of a compensating source.

The diurnal variation of the F_2 layer is less dramatic. Ionization produced at high altitudes during the day maintains a sizable density of electrons at the F_2 peak throughout the day and diffuses downward at night. This accounts for the fact that radio reception (both in the broadcast and shortwave bands) is generally best at night. Ionization at lower altitudes—primarily those corresponding to the D region—tends to interfere with radio transmissions during the day. Interference is minimal at night, since ionization in the D layer effectively disappears with the setting of the Sun.

Impact of solar activity on the density of ionization

The density of ionization varies as a function of solar activity in response to changes in the intensity and spectral properties of solar radiation. The output of solar energy is relatively constant in the visible and near-ultraviolet portions of the spectrum. It varies appreciably, however, at shorter wavelengths, reflecting changes in the temperature of the outermost regions of the solar atmosphere.

The changes are particularly large, in excess of a factor of 10, at X-ray wavelengths. Variations in the D region are correspondingly large, with smaller though still significant changes in the E and F layers. The temperature of the thermosphere varies as a function of solar activity in response to the changing input of energy. Exospheric temperatures range from about 750 K at solar minimum to as high as 2,000 K when activity is particularly intense.

Solar activity varies on a characteristic time scale of about 11 years. It is not entirely periodic, however, as successive cycles can differ significantly. There are indications that activity can be low for periods as long as centuries. For example, the Sun was quiet for more than 200 years from about 1600 to roughly 1850.

As mentioned earlier, ionization above the F_2 peak is removed mainly by downward diffusion. Ionization is constrained, however, to move along the magnetic field. The field is oriented horizontally at the magnetic equator, so vertical diffusion is inhibited at low latitudes. The density of O^+ and electrons at low latitudes is controlled by chemistry to a larger extent than at high latitudes. The F_2 peak is correspondingly higher in altitude and the density of electrons is elevated accordingly.

Ions and electrons formed at high altitudes and low latitudes are transported to higher latitudes by thermospheric winds. As a result, the highest density of electrons at the F_2 peak is observed at intermediate latitudes, displaced by about 10° relative to the magnetic equator.

Transport also can affect the distribution of ionization at lower altitudes. The diurnal pattern of heating in the troposphere and stratosphere excites a spectrum of waves, some of which are free to propagate vertically. The amplitude of the waves grows significantly as the disturbance enters regions of lower density. The passage of the waves is associated with strong alternating horizontal winds. Ionization can be driven up inclined magnetic field lines at one altitude, while winds blowing in an opposite direction at higher altitudes can induce simultaneous downward motion. This can lead to bunching of ionization, causing a local enhancement of the electron density. The mechanism is particularly important in the E region and is responsible for a phenomenon known as sporadic E.

The buildup of ionization is normally limited by dissociative recombination of molecular ions. The ionosphere at D- and E-region altitudes, however, contains a small but variable concentration of atomic ions derived from the ionization of metals ablated from meteorites. The density of metallic ions, notably those of sodium, magnesium, and potassium is sometimes high enough to supply, after concentration, a layer of ionization with density comparable to that of the F layer. This can result in a major temporary disruption of radio communications.

Winds generated in the lower ionosphere by thermal forcing from below have characteristic periods expressed as submultiples of a day. Waves with a period of 24 hours dominate at low latitudes, whereas those with a characteristic period of 12 hours are more important at high latitudes. The waves are basically similar in physical origin to tides excited in the ocean by lunar gravity. The vertical motion to which they owe their origin is generated, however, by the diurnal pattern of heating and cooling rather than by gravity. Additional waves can arise due to irregular forcing associated, for example, with thunderstorms, motion over orographic features, and a range of other small-scale meteorologic disturbances. These small-scale disturbances are referred to as gravity waves to distinguish them from the more regular planetary-scale motions excited by the diurnal cycle of heating and cooling. The regular response to thermal forcing is known as the atmospheric tide.

Gravity waves and atmospheric tides

Tides and gravity waves have similar effects on ionization in the E region. They both are responsible for concentrating ionization in layers. In combination with the large-scale system of winds in the lower thermosphere, they also are effective in driving an irregular current that flows in the E and lower F regions of the ionosphere. The current owes its origin to differences in the facility with which motions of ions and electrons are constrained by the magnetic field. It is associated with an electric field and results in a modulation of the magnetic field that can be readily detected at the surface. The current is particularly intense in the equatorial region where it is known as the electrojet. The region of strong current flow is called the dynamo region.

Protons, H^+, and helium ions, He^+, are important components of the ionosphere above the F_2 peak. They increase in abundance relative to the atomic ion of oxygen, O^+, with increasing altitude. Protons are produced by the photoionization of atomic hydrogen,

$$h\nu + H \rightarrow H^+ + e, \tag{68}$$

and by charge transfer from O^+ to H,

$$O^+ + H \rightarrow O + H^+. \tag{69}$$

Helium ions are formed by the photoionization of helium. The distribution of H^+ and He^+ with altitude reflects the influence of the polarization electric field set up to preserve charge neutrality. When O^+ is the dominant ion, the polarization field acts to lift H^+ and He^+ with a force equivalent, but in opposite direction, to that exerted by the gravitational field on a particle of mass 8 atomic units. Protons behave as though they had an effective gravitational mass of -7 atomic units $(1 - 8)$. The effective mass of He^+ is -4 atomic units $(4 - 8)$.

The abundance of H^+ and He^+ increases with altitude. Eventually H^+ becomes the dominant component of the ionosphere, and the polarization field is then diminished. A field equivalent to the gravitational force acting on a body of mass 0.5 atomic units, directed upward for ions and downward for electrons, is sufficient to maintain equal densities of H^+ and electrons. The effective masses of O^+ and He^+ shift to 15.5 atomic units $(16 - 0.5)$ and 3.5 atomic units $(4 - 0.5)$, respectively, and densities of O^+, He^+, and H^+ decline with further increases in altitude.

The overall structure of the outer ionosphere is influenced strongly by the configuration of the magnetic field. Ionization is removed by diffusion, but diffusion is limited to the directions imposed by the morphology of the field.

Close to the Earth, the magnetic field has a structure similar to that of an ideal dipole. Field lines are oriented more or less vertically at high latitudes, sweep back over the Equator when they are essentially horizontal, and connect to the Earth in a symmetrical pattern at high latitudes in the opposite hemisphere. The field departs from this ideal dipolar configuration, however, at high altitudes. There, the terrestrial field is distorted to a significant extent by the solar wind with its imbedded solar magnetic field. Ultimately the terrestrial field is dominated by the interplanetary field. The solar wind compresses the field of the Earth on the dayside at a distance of about 65,000 kilometres from the planet. The terrestrial field is stretched out on the nightside in a giant tail that reaches past the orbit of the Moon, extending perhaps to distances in excess of 1,000 Earth radii. The boundary between regions dominated by the solar and terrestrial magnetic fields is known as the magnetopause, and the region interior to the magnetopause is called the magnetosphere.

The magnetopause and the magnetosphere

The outermost regions of the magnetosphere are exceedingly complex, especially at high latitudes where the terrestrial field lines are open. Ionization from the solar wind can leak into the magnetosphere in a number of ways. It can enter by turbulent exchange at the dayside magnetopause or more directly at cusps in the magnetopause at high latitudes, where closed loops of the magnetic field on the dayside meet fields connecting to the magnetotail. In addition, it can enter at large distances on the nightside where the magnetic pressure is relatively low and the field lines are able to reconnect readily, providing easy access to the giant plasma sheet in the interior of the magnetotail. For further information about the magnetosphere and related subjects, see EARTH, THE: *The magnetic field of the Earth.*

AURORAS

Auroras are perhaps the most spectacular manifestations of the complex interaction of the solar wind with the outer atmosphere. The energetic electrons and protons responsible for auroras are thought to emanate from the plasma sheet inside the Earth's magnetotail.

Auroras occur in both hemispheres. They are confined for the most part to high latitudes, to oval-shaped regions maintaining a more or less fixed orientation with respect to the Sun. The centre of an auroral oval is displaced a few degrees to the nightside with respect to the geomagnetic pole. The midnight portion of an oval is, on average, at a geomagnetic latitude of 67°; the midday portion is at about 76°. An observer between 67° and 74° magnetic latitudes generally encounters auroras twice a day—once in the evening and once in the morning.

Auroral zone

The portion of the Earth that traverses the midnight portion of an oval is known as the auroral zone. In the Northern Hemisphere, for example, it lies along the curve extending from the northern regions of Scandinavia through Iceland, the southern tip of Greenland, the southern region of Hudson Bay, Alaska, and on to the coast of Siberia. This is the prime region from which to view an aurora. The phenomenon is by no means static, however. The auroral zone shifts poleward at times of low solar activity. It has been known to move as far south as 40° (geographic latitude) in the United States during periods of high solar activity. At low latitudes, an aurora assumes a characteristic red colour. In ancient times this feature was often interpreted as evidence of impending disaster. Auroras assume a variety of forms, depending on the vantage point from which they are observed. The luminosity of an aurora is generally aligned with the magnetic field. Field lines are close to vertical in polar regions, and so an aurora occurring there appears to stand on end, to hang from the sky in great luminous drapes (Figure 20). It is a spectacular sight indeed, especially if viewed from a distance either from the north or south. At lower latitudes, the magnetic field lines are inclined with respect to the vertical. There, an aurora appears as streamers radiating from the zenith. Such is the majesty of auroras that no two displays are totally alike. Light can move rapidly across the sky on some occasions, and at other times, it can appear to stand in place, flickering on and off.

Common cause of auroral displays

The most common type of aurora is associated with the bombardment of the atmosphere by electrons having energies of up to 10,000 electron volts (eV). The energy for these electrons originates ultimately from the Sun. It is propagated through space by the solar wind and is communicated to the magnetosphere, most probably to the plasma sheet, by complex electromagnetic interactions not as yet well understood. The energetic electrons enter the atmosphere along magnetic field lines. They produce a shower of secondary and tertiary electrons, approximately one for every 35 eV of energy in the primary stream.

V.P. Hessler

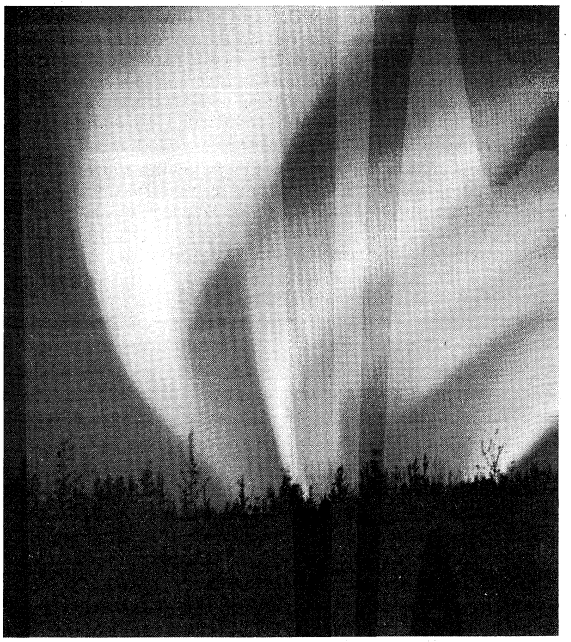


Figure 20: *Aurora borealis.*
Multiple auroral arcs photographed in Alaska.

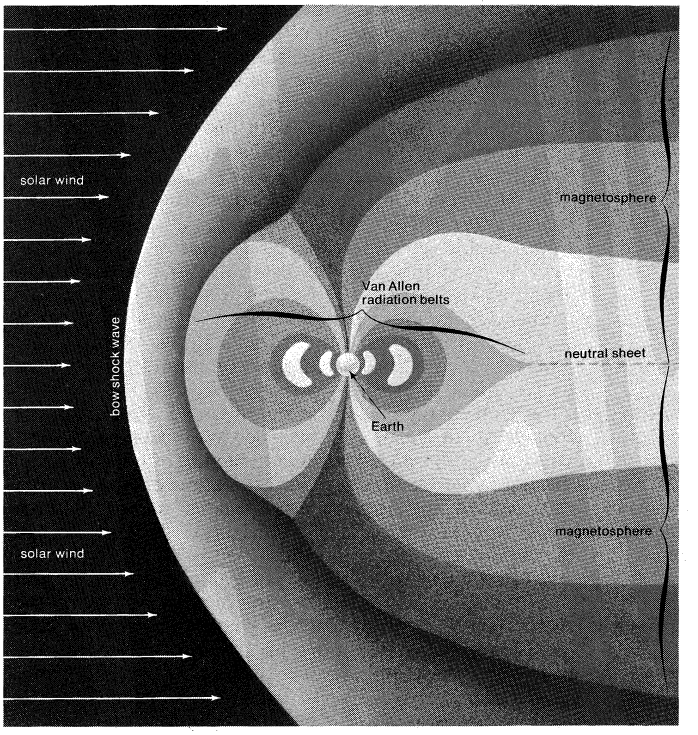


Figure 21: The Van Allen radiation belts contained within the Earth's magnetosphere.

Primaries can propagate to altitudes as low as 100 kilometres. Most of the luminosity is produced, however, by low-energy secondary and tertiary electrons. Prominent emissions are associated with the red line of atomic oxygen at 630 nm, the green line of atomic oxygen at 558 nm, the first negative bands of ionized molecular nitrogen at 391 nm and 428 nm, and a host of emissions from O, O_2, O_2^+ and N_2. Many of these features also are present in the day and night airglow. They are most notable in auroras because of their intensity and the rapidity with which they switch on and off in response to changes in the flux and energy of incoming primaries. An aurora has a characteristic red colour if the energy of primaries is relatively low. In this case, emission is dominated by atomic oxygen and is confined for the most part to altitudes above 200 kilometres. If the energy of the primaries is high, an aurora has a greenish blue colour and extends downward to altitudes as low as 90 kilometres.

Auroral displays also are produced by the bombardment of the atmosphere by energetic protons. Protons with energies of up to 200,000 eV are responsible for auroral activity in a diffuse belt shifted equatorward from the main auroral zone. These protons can be detected from the ground by observation of Doppler-shifted radiation emitted by fast hydrogen atoms formed by charge transfer from atmospheric atoms and molecules. Protons also play a role at higher latitudes, especially at times following major flares on the Sun. It is thought that the protons responsible for auroras at the polar caps are solar in origin. Associated energies may reach as high as 10^6 eV, and particles may penetrate as deep as 80 kilometres. Polar cap auroras can provide a significant transient source of mesospheric and stratospheric nitric oxide. They can be responsible for small but detectable short-term fluctuations in the abundance of stratospheric ozone.

THE VAN ALLEN RADIATION BELTS

Zones of energetic protons and electrons

The magnetosphere includes two doughnut-shaped belts, or zones, centred on the Equator that are occupied by appreciable numbers of energetic protons and electrons trapped in the magnetic field high above the sensible atmosphere (Figure 21). The inner belt extends from roughly 1,000 to 5,000 kilometres above the terrestrial surface and the outer belt from some 15,000 to 25,000 kilometres. The belts were named in honour of James A. Van Allen,

the American physicist who discovered them in 1958. Van Allen's find was a triumph of serendipity: he detected the presence of the trapped particles with a Geiger counter designed to measure the flux of cosmic rays in space. It was the first great discovery of the space age and was achieved by combining data obtained with instruments carried by three of the earliest U.S. scientific satellites—Explorer 1, Explorer 4, and Pioneer 3.

The flux of protons crossing a square centimetre of surface in the inner Van Allen belt can be as large as 20,000 per second, higher than the flux of cosmic radiation in space by a factor of 10^4. Protons in the inner belt have energies in excess of 7×10^8 electron volts, enough to enable them to penetrate several centimetres of lead. Spacecraft that have to fly through the belts thus must be protected; otherwise their electronic components would be subjected to irreparable damage.

Neutron decay as a source of high-energy protons in the inner belt

The high-energy protons in the inner Van Allen belt are thought to originate from the decay of neutrons. These neutrons are produced by the interaction of energetic cosmic rays of galactic origin with the atmosphere. Some of the neutrons are ejected upward, and a fraction of them decay into protons and electrons upon passing through the region occupied by the Van Allen belts. These protons and electrons travel in spiral paths along the flux lines of the Earth's magnetic field. At intermediate latitudes, the converging flux lines cause the particles to reverse their direction, so that they move back and forth between the magnetic poles. Collisions with atoms in the thin atmosphere eventually remove the particles from the belts, but they generally survive for about 10 years. This relatively long lifetime allows particles to accumulate in the radiation belts, providing high fluxes in spite of the small magnitude of the intrinsic source.

The inner belt merges gradually with the outer belt, where a portion of the ionization is derived from the solar wind. Clear evidence of this is provided by the observation of helium ions in addition to protons in the outer belt. It has been established that helium ions account for about 10 percent of the solar wind. (Unlike the outer zone, the inner belt contains no helium ions.) The flux of electrons in the outer belt can vary over time intervals as short as a few days. These changes appear to correlate with periods of strong magnetic disturbance. They are not, however, as yet well understood.

Studies of the outer magnetosphere and related phenomena continue to draw attention. Space physicists find it a challenging task to unravel the mysteries of the ever-changing cosmic laboratory and hope to continue the saga initiated by Van Allen and other pioneers of the space age.

(M.B.McE.)

BIBLIOGRAPHY

Development of the Earth's atmosphere: Astrophysical considerations bearing on the earliest stages of atmospheric development are reviewed in JOHN S. LEWIS and RONALD G. PRINN, *Planets and Their Atmospheres: Origin and Evolution* (1984). Aspects of subsequent development are discussed in JAMES C.G. WALKER, *Evolution of the Atmosphere* (1977); and HEINRICH D. HOLLAND, *The Chemical Evolution of the Atmosphere and Oceans* (1984). Numerous chapters in J. WILLIAM SCHOPF (ed.), *Earth's Earliest Biosphere: Its Origin and Evolution* (1983), discuss biologic controls of atmospheric composition and their development over time. The probable rise of oxygen just prior to the development of the earliest animals is discussed in detail by BRUCE RUNNEGAR, "The Cambrian Explosion: Animals or Fossils?," *Journal of the Geological Society of Australia*, 29(4):395–411 (1982). Development of the biogeochemical cycle of carbon and its interactions with the atmosphere are discussed in E.T. SUNDQUIST and W.S. BROECKER (eds.), *The Carbon Cycle and Atmospheric CO_2: Natural Variations Archean to Present* (1985).

(J.M.Ha.)

Structure, composition, and physical properties of the atmosphere: JOHN T. HOUGHTON, *The Physics of Atmospheres*, 2nd ed. (1986), is a readable account of the physical basis for the treatment of atmospheric radiation and a good introduction to atmospheric dynamics. Simpler discussions are given by RICHARD M. GOODY and JAMES C.G. WALKER, *Atmospheres* (1972). A useful account of atmospheric chemistry is given by RICHARD P. WAYNE, *Chemistry of Atmospheres: An Introduction to the Chemistry of the Atmospheres of Earth, the Planets, and Their Satellites* (1985). Atmospheric composition, including an account of the interaction of the atmosphere with the oceans, is discussed by HEINRICH D. HOLLAND, *The Chemistry of the Atmospheres and Oceans* (1978).

Effects of human activities on the atmosphere: An overview of the Earth's endangered, changing atmosphere is given in JOHN GRIBBIN (ed.), *The Breathing Planet* (1986), a collection of essays from *New Scientist;* and "The Changing Atmosphere: Implications for Mankind," *Chemical and Engineering News*, vol. 64, no. 47 (Nov. 24, 1986), a special issue on global warming and the depletion of stratospheric ozone. For a review of current scientific understanding of these problems, see WORLD METEOROLOGICAL ORGANIZATION, *Atmospheric Ozone: An Assessment of Our Understanding of the Processes Controlling Its Present Distribution and Change*, 3 vol. (1985), report no. 16 of the Global Ozone Research and Monitoring Project. See also WILLIAM C. CLARK (ed.), *Carbon Dioxide Review, 1982* (1982); UNITED STATES. CONGRESS. SENATE. COMMITTEE ON ENVIRONMENT AND PUBLIC WORKS. SUBCOMMITTEE ON TOXIC SUBSTANCES AND ENVIRONMENTAL OVERSIGHT, *Global Warming* (1986); and HAROLD W. BERNARD, JR., *The Greenhouse Effect* (1980). A survey of the history of acid deposition from the atmosphere and its sources is provided by EVILLE GORHAM, "Acid Rain: An Overview," ch. 1 in CHANDRAKANT M. BHUMRALKAR (ed.), *Meteorological Aspects of Acid Rain* (1984), pp. 1–18, with an extensive bibliography. More detailed information may be found in JON R. LUOMA, *Troubled Skies, Troubled Waters: The Story of Acid Rain* (1984); and THOMAS PAWLICK, *A Killing Rain: The Global Threat of Acid Precipitation* (1984). Further bibliographic information is available in an annotated bibliography of research, G. HARRY STOPP, JR., *Acid Rain* (1985).

The ionosphere and phenomena of the upper atmosphere: For a general review of the underlying physics of the upper atmosphere, see J.A. RATCLIFFE (ed.), *Physics of the Upper Atmosphere* (1960); and C.O. HINES *et al.* (eds.), *Physics of the Earth's Upper Atmosphere* (1965). A modern account of aeronomy is given in P.M. BANKS and G. KOCKARTS, *Aeronomy*, 2 vol. (1973). A classic treatment of the aurora and airglow is presented in JOSEPH W. CHAMBERLAIN, *Physics of the Aurora and Airglow* (1961). For a specialized account of magnetospheric processes, see WILMOT N. HESS, *The Radiation Belt and Magnetosphere* (1968). A readable summary article is J.G. ROEDERER, "The Particle and Field Environment of the Earth," *Astronautics and Aeronautics*, 7:22–28 (January 1969).

Current research is reported in the following journals: *Advances in Atmospheric Sciences* (quarterly); *Atmospheric Environment* (monthly); *Environmental Science & Technology* (monthly); *Journal of the Atmospheric Sciences* (semimonthly); *Science* (weekly); and *Scientific American* (monthly).

(M.B.McE.)

Atoms: Their Structure, Properties, and Component Particles

The atom is the smallest unit into which matter can be divided without the release of electrically charged particles. It also is the smallest unit of matter that has the characteristic properties of a chemical element. As such, the atom is the basic building block of chemistry.

Most of the atom is empty space. The rest consists of a positively charged nucleus of protons and neutrons surrounded by a cloud of negatively charged electrons. The nucleus is small and dense compared to the electrons, which are the lightest charged particles in nature. Electrons are attracted to any positive charge by their electric force; in an atom, electric forces bind the electrons to the nucleus.

It is easier to describe an atom mathematically than conceptually, and so physicists have developed several models to explain its various characteristics. In some respects, the electrons in an atom behave like particles orbiting the nucleus. In others, the electrons behave like waves frozen in position around the nucleus. Such wave patterns, called orbitals, describe the distribution of individual electrons. The behaviour of an atom is strongly influenced by these orbital properties, and its chemical properties are determined by orbital groupings known as shells.

This article opens with a broad overview of the fundamental properties of the atom and its constituent particles and forces. A more mathematical and technical discussion of its structure and nucleus is provided in subsequent sections. Included too is a historical survey of the most influential concepts about the atom that have been formulated through the centuries. For additional information pertaining to nuclear structure and elementary particles, see SUBATOMIC PARTICLES. For coverage of other related topics in the *Macropædia* and *Micropædia,* see the *Propædia,* sections 111, 112, 121, 122, 124, 125, and 128.

This article is divided into the following sections:

Components and properties of atoms

CONSTITUENT PARTICLES AND FORCES

Most matter consists of an agglomeration of molecules, which can be separated relatively easily. Molecules, in turn, are composed of atoms joined by chemical bonds that are more difficult to break. Each individual atom consists of smaller particles—namely, electrons and nuclei. These particles are electrically charged, and the electric forces on the charge are responsible for holding the atom together. Attempts to separate these smaller constituent particles require ever-increasing amounts of energy and result in the creation of new subatomic particles, many of which are charged.

As noted at the outset of this article, an atom consists largely of empty space. The nucleus is the positively charged centre of an atom and contains most of its mass. It is composed of protons, which have a positive charge, and neutrons, which have no charge. These constituent protons and neutrons collectively are called nucleons. Protons, neutrons, and the electrons surrounding them are long-lived particles present in all ordinary, naturally occurring atoms. Other subatomic particles may be found in association with these three types of particles. They can be created only with the addition of enormous amounts of energy, however, and are very short-lived.

Nucleons

All atoms are roughly the same size, whether they have three or 90 electrons. Approximately 50,000,000 atoms of solid matter lined up in a row would measure one centimetre (0.4 inch). A convenient unit of length for measuring atomic sizes is the angstrom (Å), defined as 10^{-10} metre. The radius of an atom measures 1–2 Å.

Compared to the overall size of the atom, the nucleus is even more minute. It is in the same proportion to the atom as a marble is to a football field. In volume, the nucleus takes up only 10^{-14} of the space in the atom—*i.e.*, one part in 100,000,000,000,000. A convenient unit of length for measuring nuclear sizes is the femtometre (fm), which equals 10^{-15} metre. The diameter of a nucleus depends on the number of particles it contains and ranges from about 4 fm for a light nucleus such as carbon to 15 fm for a heavy nucleus such as lead. In spite of the small size of the nucleus, virtually all the mass of the atom is concentrated there. The protons are massive, positively charged particles, whereas the neutrons have no charge and are nearly as massive as the protons. The fact that nuclei can have anywhere from one to about 250 nucleons accounts for their wide variation in mass. The lightest nucleus, that of hydrogen, is 1,836 times more massive than an electron, while heavy nuclei are nearly 500,000 times more massive.

Size of nuclei

Although electrons exhibit complicated behaviour within an atom, they are characterized completely by a few parameters. The intrinsic properties of an electron are its charge, mass, an internal motion called spin, and magnetic moment. All electrons have identical properties. As the lightest charged particles in existence, they are absolutely stable because they cannot decay into smaller units. Their charge and mass, which are important determinants of atomic properties, are listed in Table 1. The spin of the electron provides it with a directional orientation. The electron has a magnetic moment along its spin axis. (Magnetic moment is a property of a particle, which, like a compass needle, causes its axis to align in a magnetic field.) Electrons are subject not only to the electromagnetic force but also to the force of gravity and the so-called weak interaction, the force primarily manifested in the radioactive decay of nuclei.

Electrons

Table 1: Fundamental Atomic Constants

	symbol	value
Avogadro's number	N_A	6.022×10^{23} mol^{-1}
Fundamental charge	e	1.602×10^{-19} C
Faraday	F	9.649×10^{4} C mol^{-1}
Planck's constant	h	6.626×10^{-34} J·s
	$\hbar = h/2\pi$	1.055×10^{-34} J·s
Mass of electron	m_e	9.11×10^{-31} kg
Rest energy of electron	$m_e c^2$	8.2×10^{-14} J $= 5.11 \times 10^{5}$ eV
Mass of proton	m_p	1.673×10^{-27} kg
Rest energy of proton	$m_p c^2$	1.50×10^{-10} J $= 938.3$ MeV
Bohr radius	a_0	5.292×10^{-11} m $= 0.5292$ Å
Speed of light (in vacuum)	c	2.998×10^{8} m s^{-1}

Most properties of atoms—particularly those associated with chemical bonds, physical forces, and the properties of bulk matter—depend solely on the behaviour of the electrons surrounding the nucleus. The chemical properties of an atom depend on the arrangement of its electrons making up the cloud around the nucleus. The atoms of one element differ from those of other elements in the number of their electrons. Also, atoms form molecules by lending and sharing electrons. Some elements, such as alkali metals, have an electron that is loosely bound to the nucleus and thus easily removed in chemical reactions; other elements, such as the noble (or inert) gases, have very tightly bound electrons and little affinity for other electrons. Electrons that have been freed from their atoms can cause lightning, and freed electrons driven through wires make up ordinary electric currents.

The nucleus of an atom is characterized by the number of protons and neutrons in it. Besides a charge and mass, the nucleus also may have a spin and a magnetic moment of its own, depending on the internal arrangement of its protons and neutrons. The forces between nucleons include the three forces affecting electrons as well as the so-called strong force, which is much more powerful than any of the others. Because of the strong force, nuclear binding energies are 1,000,000 times the binding energies of electrons in atoms. The amounts of energy that can be released in a transformation of the nucleus are correspondingly larger than the chemical energies released by a transformation of the electron patterns in atoms. The protons and neutrons in the nucleus are governed by the laws of quantum mechanics (see below), which describe the complex internal structure of the nucleus.

Even the individual protons and neutrons that make up the nucleus have an internal structure of their own. The constituents of nucleons are called quarks. Unlike the particles composing the larger units of matter, the quark cannot be freed from the nucleon and studied in isolation. The strong force acting between quarks is so powerful that they can never be completely separated. Any attempt to probe the substructure of a nucleon releases particles of various types, but the particles produced contain quarks in fixed combinations, never single quarks. Particles containing quarks are collectively called hadrons. Hadrons are classified into two categories: baryons and mesons. Baryons, which are composed of three quarks, include protons and neutrons as their lightest examples. Mesons, which contain two quarks, are largely responsible for nuclear forces. Except for the nucleons, all such particles decay in a small fraction of a second after their creation.

Hadrons and quarks

There is one more broad category of subatomic particles, the leptons. Electrons and neutrinos are leptons. They have no detectable internal components and may be truly fundamental particles of nature. The neutrino is an uncharged particle with little or no mass that is created during the radioactive decay of nuclei.

Leptons

PROPERTIES OF ATOMS

Atomic number. The single most important characteristic of an atom is its atomic number, which is defined as the number of units of positive charge in the nucleus. A neutral atom has an equal number of protons and electrons, so that the positive and negative charges exactly balance. The atomic number determines the chemical properties of an atom, including the kinds of molecules that can be formed and their binding energies. Hence, the atomic number determines an atom's characteristics as an element. (An element is composed of atoms with the same atomic number.) Elements found in nature range from atomic number 1, hydrogen, to atomic number 92, uranium. In addition, artificial elements with atomic numbers beyond 100 have been produced.

Atomic mass number. The total number of nucleons (both protons and neutrons) in an atom is the atomic mass number, or mass number. Atoms with the same atomic number but different atomic masses are called

Differences between isotopes

isotopes. Isotopes have identical chemical properties, yet they can have very different nuclear properties (see below *Isotopes*). The nuclear properties of an atom include possible radioactivity (the propensity to become radioactive in nuclear reactions), magnetic properties, and weight. The element potassium, for example, has two natural isotopes, ^{39}K and ^{40}K. They form exactly the same compounds, but ^{40}K is radioactive and decays into another element. In scientific notation, the isotope of potassium with 19 protons, 20 neutrons, and a total of 39 nucleons can be written either as $_{19}K^{39}$ or as $^{39}_{19}K$.

Because isotopes have the same number of protons, all of the isotopes of a given element occupy the same place in the periodic table of elements. Most elements have stable isotopes. For example, hydrogen has three isotopes, each with one proton. The nucleus of ordinary hydrogen is an isolated proton, but the isotope deuterium has a neutron bound to the proton. Both of these isotopes are stable. The third hydrogen isotope, tritium, has two neutrons and is radioactive. Radioactive isotopes can be made for many elements; the more the number of neutrons deviates from the optimum number for that atomic mass, the shorter the life of the radioactive isotope.

Atomic weight. The term atomic weight, or atomic mass, refers to the mass of a fixed number of atoms of an element. The standard scientific unit for dealing with atoms in macroscopic quantities is the mole (mol), which is defined arbitrarily as the amount of a substance with as many atoms or other units as there are in 12 grams of the carbon isotope ^{12}C. The number of atoms in a mole is called Avogadro's number, the value of which is approximately 6×10^{23}. The atomic mass of an element expressed in daltons, or more commonly atomic mass units (amu's), is the number of grams in one mole of the element. The amu is convenient because atomic masses are nearly equal to atomic mass numbers and therefore are close to integer values.

Historically, the law for chemical combination according to molar weights was the primary evidence for the existence of atoms and molecules. For example, two grams of hydrogen combine with 16 grams of oxygen to form water. This represents two moles of hydrogen of atomic weight 1 combining with one mole of oxygen of atomic weight 16. Elements consisting of a mixture of several isotopes may not have an atomic mass close to an integer, because the mass will be the weighted average of the different isotopes. An example is chlorine, which has two common isotopes, ^{35}Cl and ^{37}Cl, and a weighted average mass of 35.5 amu.

Electric charge. The normal atom is electrically neutral, meaning that it carries a net electric charge of zero. Some atoms, however, have lost or gained electrons in chemical reactions or in collisions with other particles. Atoms with a net charge, either from the gain or loss of electrons, are called ions. If a neutral atom loses an electron, it becomes a positive ion; if it gains an electron, it becomes a negative ion.

Ions

The charge on any particle is a whole multiple of the electron's charge, either positive or negative. The quarks are an exception to this rule. They have charges of $+^2/_3e$ and $-^1/_3e$. However, they exist only in groups, and each group as a whole has an integral multiple of the electron's charge. The amount of charge in this fundamental unit is equal to 1.6×10^{-10} coulomb. This means that in a current of one ampere—roughly what a 100-watt light bulb uses in the ordinary 110-volt household circuit—about 6×10^{18} electrons pass through the wire every second.

Electron shells. The behaviour of electrons in atoms is quite subtle and is governed by the laws of quantum mechanics. According to these laws, electrons occupy various regions of the atom in frozen wave patterns called orbitals. The orbitals are most easily visualized as clouds surrounding the nucleus. The shape and size of the orbital, and the energy of the electron in it, are calculated by differential equations. The orbitals vary in shape from smooth and spherical for the electrons most tightly bound to the nucleus to rather diffuse and lumpy for the least bound electrons. The hydrogen atom has a single electron in a spherical cloud. The electron could go into other orbitals, but it would require additional energy in the atom to do so. The quantum theory provides that the energy of an atom can only change in definite amounts called quanta. The different possible states an atom can be in, each with its own definite energy, are called energy levels. The light emitted from an atom has specific frequencies associated with the energy quanta. Energy at the atomic level is often expressed in electron volts (eV). There are 2.26×10^{25} eV in one kilowatt-hour. To remove an electron from an atom requires several electron volts, depending on the atom. Visible light has a quantum energy of about 2 eV.

Energy levels and quanta

Each electron in a multielectron atom has its own orbital, according to a law of quantum mechanics known as the Pauli principle (also called the Pauli exclusion principle or exclusion principle). Thus, in atoms with many electrons, many different kinds of orbitals are occupied. A group of orbitals with the same or nearly the same energy is called a shell. The pattern of filled and unfilled shells in each element is different; this variety gives the elements their distinctive characteristics.

Chemical behaviour. The chemical behaviour of atoms depends on the shells of the more loosely bound electrons. The Pauli principle is responsible for chemical valence, the principle of chemistry according to which atoms of one element bond to a definite number of atoms in other elements according to simple counting rules. If these shells are completely filled, the electrons are tightly bound and the atom does not readily share or lend its electrons to form chemical bonds. If there is only one electron in the last shell, it is weakly bound and the atom can be easily ionized. Examples of these situations are helium, which has a filled shell and is an inert gas, and lithium, which has one more electron in the next shell and is a highly reactive metal.

Formation of chemical bonds

One kind of chemical bond is the ionic bond, in which a loosely bound electron from one atom transfers to a deeper shell in another atom. The two ions are held together by their electrical forces. Another kind of bond is the covalent bond. In this situation, the electron clouds of one atom are distorted by the presence of another atom. In the new cloud pattern, the outer electrons are more concentrated in the region between the two atoms. Thus, the atoms share their electrons. This allows atoms of the same element to form chemical bonds, which could not happen with ionic bonds. Chemical-bond energies typically measure several electron volts.

Nuclear properties. Like atoms, nuclei have a shell structure with the protons and neutrons in orbitals. Nuclei can exist in states of different energy, but ordinary stable nuclei are always in the most bound state. The scale of these energies is 1,000,000 times as large as atomic or chemical energies.

Nuclear transformations

Nuclei can undergo transformations that affect their binding energies. If a transformation leads to more tightly bound nuclei, the excess energy will be released in some form. If one mole of atoms undergoes a nuclear transformation and releases 1,000,000 electron volts (1 MeV) of energy per nucleus, the total energy will be 10^{11} joules.

Some transformations can take place spontaneously, and such a process is called radioactivity. In one form of radioactivity, a neutron in the nucleus is converted to a proton or vice versa. If an electron is emitted at the same time, the process is known as beta radioactivity and the electron is called a beta ray. In another form of radioactivity, the nucleus disintegrates into one of lower mass number with the excess nucleons being ejected as a small nucleus. The small nucleus is commonly helium-4. This process is called alpha radioactivity, and the emitted helium-4 nuclei are called alpha rays. A third kind of ray observed in radioactivity is the gamma ray. Such rays are quanta of light of very high energy that are emitted when the nucleus makes a transition from one energy state to another of lower energy (see below *Radioactivity*).

Nuclear transformations also take place in nuclear reactions, which are the processes that occur when a nucleus is struck by some external particle. In a fusion reaction, two light nuclei come together and merge into a single heavier nucleus. Another important reaction is fission, the division of a nucleus into two roughly equal parts. Fission can be induced in the heaviest elements by reactions with free

neutrons. Both fusion and fission can release energy by reforming the nuclei so that their atomic masses are closer to the middle range where nuclei have maximum binding energy (see below *Nuclear fission* and *Nuclear fusion*).

Development of atomic theory

The concept of the atom that Western scientists accepted in broad outline from the 1600s until about 1900 originated with Greek philosophers in the 5th century BC. Their speculation about a hard, indivisible fundamental particle of nature was replaced slowly by a scientific theory supported by experiment and mathematical deduction. It was 2,000 years before modern physicists realized that the atom is indeed divisible and that it is not hard, solid, or immutable.

THE ATOMIC PHILOSOPHY OF THE EARLY GREEKS

Leucippus of Miletus (5th century BC) is thought to have originated the atomic philosophy. His famous disciple, Democritus of Abdera, developed and named the building blocks of matter *atomos,* meaning literally "indivisible," about 430 BC. Democritus believed that atoms were uniform, solid, hard, incompressible, and indestructible and that they moved in infinite numbers through empty space until stopped. Differences in atomic shape and size determined the various properties of matter. In Democritus' philosophy, atoms existed not only for matter but also for such qualities as perception and the human soul. For example, sourness was caused by needle-shaped atoms, while the colour white was composed of smooth-surfaced atoms. The atoms of the soul were considered to be particularly fine. Democritus developed his atomic philosophy as a middle ground between two opposing Greek theories about reality and the illusion of change. He argued that matter was subdivided into indivisible and immutable particles that created the appearance of change when they joined and separated from others.

The speculations of Democritus

The philosopher Epicurus of Samos (341–270 BC) used Democritus' ideas to try to quiet the fears of superstitious Greeks. According to Epicurus' materialistic philosophy, the entire universe was composed exclusively of atoms and void, and so even the gods were subject to natural laws.

Most of what is known about the atomic philosophy of the early Greeks comes from Aristotle's attacks on it and from a long poem, *De rerum natura* ("On the Nature of Things"), which the Latin poet and philosopher Titus Lucretius Carus (*c.* 95–55 BC) wrote to popularize its ideas. The Greek atomic theory is significant historically and philosophically, but it has no scientific value. It was not based on observations of nature, measurements, tests, or experiments. Instead, the Greeks used mathematics and reason almost exclusively when they wrote about physics. Like the later theologians of the Middle Ages, they wanted an all-encompassing theory to explain the universe, not merely a detailed experimental view of a tiny portion of it. Science constituted only one aspect of their broad philosophical system. Thus, Plato and Aristotle attacked Democritus' atomic theory on philosophical grounds rather than on scientific ones. Plato valued abstract ideas more than the physical world and rejected the notion that attributes such as goodness and beauty were "mechanical manifestations of material atoms." Where Democritus believed that matter could not move through space without a vacuum and that light was the rapid movement of particles through a void, Aristotle rejected the existence of vacuums because he could not conceive of bodies falling equally fast through a void. Aristotle's conception prevailed in medieval Christian Europe; its science was based on revelation and reason, and the Roman Catholic theologians rejected Democritus as materialistic and atheistic.

THE EMERGENCE OF EXPERIMENTAL SCIENCE

De rerum natura, which was rediscovered in the 15th century, helped fuel the 17th-century debate between orthodox Aristotelian views and the new experimental science. The poem was printed in 1649 and popularized by Pierre Gassendi, a French priest who tried to separate Epicurus' atomism from its materialistic background by arguing that God created atoms.

Soon after Galileo Galilei expressed his belief that vacuums can exist (1638), scientists began studying the properties of air and partial vacuums to test the relative merits of Aristotelian orthodoxy and the atomic theory. The experimental evidence about air was only gradually separated from this philosophical controversy.

The Anglo-Irish chemist Robert Boyle began his systematic study of air in 1658 after he learned that Otto von Guericke, a German physicist and engineer, had invented an improved air pump four years earlier. In 1662 Boyle published the first physical law expressed in the form of an equation that describes the functional dependence of two variable quantities. This formulation became known as Boyle's law. From the beginning, Boyle wanted to analyze the elasticity of air quantitatively, not just qualitatively, and to separate the particular experimental problem about air's "spring" from the surrounding philosophical issues. Pouring mercury into the open end of a closed J-shaped tube, Boyle forced the air in the short side of the tube to contract under the pressure of the mercury on top. By doubling the height of the mercury column, he roughly doubled the pressure and halved the volume of air. By tripling the pressure, he cut the volume of air to a third, and so on.

Boyle's law

This behaviour can be formulated mathematically in the relation $PV = P'V'$, where P and V are the pressure and volume under one set of conditions and P' and V' represent them under different conditions. Boyle's law says that pressure and volume are inversely related for a given quantity of gas. Although it is only approximately true for real gases, Boyle's law is an extremely useful idealization that played an important role in the development of atomic theory.

Soon after his air-pressure experiments, Boyle wrote that all matter is composed of solid particles arranged into molecules to give material its different properties. He explained that all things are "made of one Catholick Matter common to them all, and . . . differ but in the shape, size, motion or rest, and texture of the small parts they consist of."

In France Boyle's law is called Mariotte's law after the physicist Edme Mariotte, who discovered the empirical relationship independently in 1676. Mariotte realized that the law holds true only under constant temperatures; otherwise, the volume of gas expands when heated or contracts when cooled.

Forty years later, Isaac Newton expressed a typical 18th-century view of the atom that was similar to that of Democritus, Boyle, and Gassendi. In the last query in his book *Opticks* (1704), Newton stated:

> All these things being considered, it seems probable to me that God in the Beginning form'd Matter in solid, massy, hard, impenetrable, moveable Particles, of such Sizes and Figures, and with such other Properties, and in such Proportion to Space, as most conduced to the End for which he form'd them; and that these primitive Particles being Solids, are incomparably harder than any porous Bodies compounded of them; even so very hard, as never to wear or break in pieces; no ordinary Power being able to divide what God himself made one in the first Creation.

By the end of the 18th century chemists were just beginning to learn how chemicals combine. In 1794 Joseph-Louis Proust of France published his law of definite proportions (also known as Proust's law). He stated that the components of chemical compounds always combine in the same proportions by weight. For example, Proust found that no matter where he got his samples of the compound copper carbonate, they were composed by weight of five parts copper, four parts oxygen, and one part carbon.

Law of definite proportions

THE BEGINNINGS OF MODERN ATOMIC THEORY

Experimental foundation of atomic chemistry. The British chemist and physicist John Dalton extended Proust's work and converted the atomic philosophy of the Greeks into a scientific theory between 1803 and 1808. His book *New System of Chemical Philosophy* (part I, 1808; part II, 1810) was the first application of atomic

Dalton's application of atomic theory to chemistry

theory to chemistry. It provided a physical picture of how elements combine to form compounds and a phenomenological reason for believing that atoms exist. His work, together with that of Joseph-Louis Gay-Lussac of France and Amedeo Avogadro of Italy, provided the experimental foundation of atomic chemistry.

On the basis of the law of definite proportions, Dalton deduced the law of multiple proportions, which stated that when two elements form more than one compound by combining in more than one proportion by weight, the weight of one element in one of the compounds is in simple, integer ratios to its weights in the other compounds. For example, Dalton knew that oxygen and carbon can combine to form two different compounds and that carbon dioxide (CO_2) contains twice as much oxygen by weight as carbon monoxide (CO). In this case, the ratio of oxygen in one compound to the amount of oxygen in the other is the simple integer ratio 2:1. Although Dalton called his theory "modern" to differentiate it from Democritus' philosophy, he retained the Greek term atom to honour the ancients.

Dalton had begun his atomic studies by wondering why the different gases in the atmosphere do not separate, with the heaviest on the bottom and the lightest on the top. He decided that atoms are not infinite in variety as had been supposed and that they are limited to one of a kind for each element. Proposing that all the atoms of a given element have the same fixed mass, he concluded that elements react in definite proportions to form compounds because their constituent atoms react in definite proportion to produce compounds. He then tried to figure out the masses for well-known compounds. To do so, Dalton made a faulty but understandable assumption that the simplest hypothesis about atomic combinations was true. He maintained that the molecules of an element would always be single atoms. Thus, if two elements form only one compound, he believed that one atom of one element combined with one atom of another element. For example, describing the formation of water, he said that one atom of hydrogen and one of oxygen would combine to form HO instead of H_2O. Dalton's mistaken belief that atoms join together by attractive forces was accepted and formed the basis of most of 19th-century chemistry. As long as scientists worked with masses as ratios, a consistent chemistry could be developed because they did not need to know whether the atoms were separate or joined together as molecules.

Gay-Lussac's law of combining gases

Gay-Lussac soon took the relationship between chemical masses implied by Dalton's atomic theory and expanded it to volumetric relationships of gases. In 1809 he published two observations about gases that have come to be known as Gay-Lussac's law of combining gases. The first part of the law says that, when gases combine chemically, they do so in numerically simple volume ratios. Gay-Lussac illustrated this part of his law with three oxides of nitrogen. The compound NO has equal parts of nitrogen and oxygen by volume. Similarly, in the compound N_2O, the two parts by volume of nitrogen combine with one part of oxygen. He found corresponding volumes of nitrogen and oxygen in NO_2. Thus, Gay-Lussac's law relates volumes of the chemical constituents within a compound, unlike Dalton's law of multiple proportions, which relates only one constituent of a compound with the same constituent in other compounds.

The second part of Gay-Lussac's law states that if gases combine to form gases, the volumes of the products are also in simple numerical ratios to the volume of the original gases. This part of the law was illustrated by the combination of carbon monoxide and oxygen to form carbon dioxide. Gay-Lussac noted that the volume of the carbon dioxide is equal to the volume of carbon monoxide and is twice the volume of oxygen. He did not realize, however, that the reason that only half as much oxygen is needed is because the oxygen molecule splits in two to give a single atom to each molecule of carbon monoxide. In his "Mémoire sur la combinaison des substances gazeuses, les unes avec les autres" (1809; "Memoir on the Combination of Gaseous Substances with Each Other"), Gay-Lussac wrote:

> Thus it appears evident to me that gases always combine in the simplest proportions when they act on one another; and we have seen in reality in all the preceding examples that the ratio of combination is 1 to 1, 1 to 2 or 1 to 3. . . . Gases . . . in whatever proportions they may combine, always give rise to compounds whose elements by volume are multiples of each other. . . . Not only, however, do gases combine in very simple proportions, as we have just seen, but the apparent contraction of volume which they experience on combination has also a simple relation to the volume of the gases, or at least to one of them.

Gay-Lussac's work raised the question of whether atoms differ from molecules and, if so, how many atoms and molecules are in a volume of gas. Amedeo Avogadro, building on Dalton's efforts, solved the puzzle, but his work unfortunately was ignored for 50 years.

Avogadro's hypotheses

In 1811 Avogadro proposed two hypotheses: (1) The atoms of elemental gases may be joined together in molecules rather than existing as separate atoms, as Dalton believed. (2) Equal volumes of gases contain equal numbers of molecules. These hypotheses explained why only half a volume of oxygen was necessary to combine with a volume of carbon monoxide to form carbon dioxide. Each oxygen molecule has two atoms, and each atom of oxygen joins one molecule of carbon monoxide.

Until the early 1860s, however, the allegiance of chemists to another concept espoused by the eminent Swedish chemist Jöns Jacob Berzelius blocked acceptance of Avogadro's ideas. (Berzelius was influential among chemists because he had determined the atomic weights of many elements extremely accurately.) Berzelius contended incorrectly that all atoms of a similar element repel each other because they have the same electric charge. He thought that only atoms with opposite charges could combine to form molecules.

Because early chemists did not know how many atoms were in a molecule, their chemical notation systems were in a state of chaos by the mid-19th century. Berzelius and his followers, for example, used the general formula MO for the chief metallic oxides, while others assigned the formula used today, M_2O. A single formula stood for different substances, depending on the chemist: H_2O_2 was water or hydrogen peroxide; C_2H_4 was marsh gas or ethylene. Proponents of the system used today based their chemical notation on an empirical law formulated in 1819 by the French scientists Pierre-Louis Dulong and Alexis-Thérèse Petit concerning the specific heat of elements. According to the so-called Dulong–Petit law, the specific heat of all elements is the same on a per atom basis. This law, however, was found to have many exceptions and was not fully understood until the development of quantum theory in the 20th century.

To resolve such problems of chemical notation, the Sicilian chemist Stanislao Cannizzaro revived Avogadro's ideas in 1858 and expounded them at the First International Chemical Congress, which met in Karlsruhe, Ger., in 1860. A noted German chemistry professor wrote later that, when he heard Avogadro's theory at the congress, "It was as though scales fell from my eyes, doubt vanished, and was replaced by a feeling of peaceful certainty." Within a few years, Avogadro's hypotheses were widely accepted in the world of chemistry.

Atomic weights and the periodic table. As more and more elements were discovered during the 19th century, scientists began to wonder how the physical properties of the elements were related to their atomic weights. During the 1860s several schemes were suggested. The Russian chemist Dmitry Ivanovich Mendeleyev based his system on the atomic weights of the elements as determined by Avogadro's theory of diatomic molecules. In his paper of 1869 introducing the periodic table, he credited Cannizzaro for using "unshakeable and indubitable" methods to determine atomic weights. "The elements, if arranged according to their atomic weights, show a distinct periodicity of their properties. . . . Elements exhibiting similarities in their chemical behavior have atomic weights which are approximately equal (as in the case of Pt, Ir, Os) or they possess atomic weights which increase in a uniform manner (as in the case of K, Rb, Cs)." Skipping

Mendeleyev's periodic table

hydrogen because it is anomalous, Mendeleyev arranged the 63 elements known to exist at the time into six groups according to valence. Valence, which is the combining power of an element, determines the proportions of the elements in a compound. For example, H_2O combines oxygen with a valence of 2 and hydrogen with a valence of 1. Recognizing that chemical qualities change gradually as atomic weight increases, Mendeleyev predicted a new element wherever there was a gap in atomic weights between adjacent elements. His system was thus a research tool and not merely a system of classification. Mendeleyev's periodic table raised an important question, however, for future atomic theory to answer: Where does the pattern of atomic weights come from?

Kinetic theory of gases. Whereas Avogadro's theory of diatomic molecules was ignored for 50 years, the kinetic theory of gases was rejected for more than a century. The kinetic theory relates the independent motion of molecules to the mechanical and thermal properties of gases—namely, their pressure, volume, temperature, viscosity, and heat conductivity. Three men—Daniel Bernoulli in 1738, John Herapath in 1820, and John James Waterston in 1845—independently developed the theory. The kinetic theory of gases, like the theory of diatomic molecules, was a simple physical idea that chemists ignored in favour of an elaborate explanation of the properties of gases.

The contributions of Bernoulli, Herapath, and Waterston

Bernoulli, a Swiss mathematician and scientist, worked out the first quantitative mathematical treatment of the kinetic theory in 1738, picturing gases as consisting of an enormous number of particles in very fast, chaotic motion (see Figure 1). He derived Boyle's law by assuming that gas pressure is caused by the direct impact of particles on the walls of their container. He understood the difference between heat and temperature, realizing that heat makes gas particles move faster and that temperature merely measures the propensity of heat to flow from one body to another. In spite of its accuracy, Bernoulli's theory remained virtually unknown during the 18th century and early 19th century for several reasons. First, chemistry was more popular than physics among scientists of the day, and Bernoulli's theory involved mathematics. Second, Newton's reputation insured the success of his more comprehensible theory that gas atoms repel one another. Finally, Joseph Black, another noted British scientist, developed the caloric theory of heat, which proposed that heat was an invisible substance permeating matter. At the time, the fact that heat could be transmitted by light seemed a persuasive argument that heat and motion had nothing to do with each other.

Herapath, an amateur ignored by his contemporaries, published his version of the kinetic theory in 1821. He also derived an empirical relation akin to Boyle's law but did not understand correctly the role of heat and temperature in determining the pressure of a gas.

Waterston's efforts met with a similar fate. A civil engineer and amateur physicist, he could not even get his work published by the scientific community, which had become increasingly professional throughout the 19th century. Nevertheless, Waterston made the first statement of the law of equipartition of energy, according to which all kinds of particles have equal amounts of thermal energy. He derived practically all the consequences of the fact that pressure exerted by a gas is related to the number of molecules per cubic centimetre, their mass, and their mean squared velocity. He derived the basic equation of kinetic theory, which reads $P = NMV^2$. Here P is the pressure of a volume of gas, N is the number of molecules per unit volume, M is the mass of the molecule, and V^2 is the average velocity squared of the molecules. Recognizing that the kinetic energy of a molecule is proportional to MV^2 and that the heat energy of a gas is proportional to the temperature, Waterston expressed the law as $PV/T =$ a constant.

During the late 1850s, a decade after Waterston had formulated his law, the scientific community was finally ready to accept a kinetic theory of gases. The studies of heat undertaken by the British physicist James Prescott Joule during the 1840s had shown that heat is a form of energy. This work, together with the law of the conservation of energy that he helped to establish, had persuaded scientists to discard the caloric theory by the mid-1850s. The caloric theory had required that a substance contain a definite amount of caloric (*i.e.*, a hypothetical weightless fluid) to be turned into heat; however, experiments showed that any amount of heat can be generated in a substance by putting enough energy into it. Thus, there was no point to hypothesizing such a special fluid as caloric.

At first, after the collapse of the caloric theory, physicists had nothing with which to replace it. Joule, however, discovered Herapath's kinetic theory and used it to calculate the velocity of hydrogen molecules in 1851. Then Rudolf Clausius developed the kinetic theory mathematically in 1857, and the scientific world took note. Clausius and two other physicists, James Clerk Maxwell and Ludwig Eduard Boltzmann (who developed the kinetic theory of gases in the 1860s), introduced sophisticated mathematics into physics for the first time since Newton. In his 1860 paper "Illustrations of the Dynamical Theory of Gases," Maxwell used probability to produce his famous distribution curve for the velocities of gas molecules. Employing Newtonian laws of mechanics, he also provided a mathematical basis for Avogadro's theory. Maxwell, Clausius, and Boltzmann assumed that gas particles were in constant motion, that they were tiny compared to their space, and that their interactions were very brief. They then related the motion of the particles to pressure, volume, and temperature. Interestingly, none of the three committed himself on the nature of the particles.

From Daniel Bernoulli, *Hydrodynamica* (1738), in *A Source Book of Physics*, trans., W.F. Magie (1935); Harvard University Press

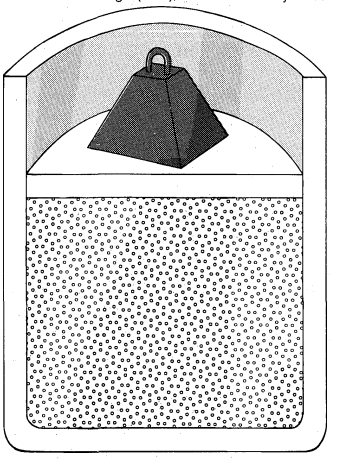


Figure 1: As conceived by Daniel Bernoulli in *Hydrodynamica* (1738), gases consist of numerous particles in rapid, random motion. He assumed that the pressure of a gas is produced by the direct impact of the particles on the walls of the container (see text).

STUDIES OF THE PROPERTIES OF ATOMS

Size of atoms. The first modern estimates of the size of atoms and the numbers of atoms in a given volume were made by the German chemist Joseph Loschmidt in 1865. Loschmidt used the results of kinetic theory and some rough estimates to do his calculation. The size of the atoms and the distance between them in the gaseous state are related both to the contraction of gas upon liquefaction and to the mean free path traveled by molecules in a gas. The mean free path, in turn, can be found from the thermal conductivity and diffusion rates in the gas. Loschmidt calculated the size of the atom and the spacing between atoms by finding a solution common to these relationships. His result for Avogadro's number is remarkably close to the present accepted value of 6.022×10^{23}. The precise definition of Avogadro's number is the number of atoms in 12 grams of the carbon isotope ^{12}C. Loschmidt's result for the diameter of an atom was approximately 10^{-8} centimetres.

Loschmidt's determinations

Much later, in 1908, the French physicist Jean Perrin used Brownian motion to determine Avogadro's number.

Brownian motion, first observed in 1827 by the Scottish botanist Robert Brown, is the continuous movement of tiny particles suspended in water. Their movement is caused by the thermal motion of water molecules bumping into the particles. Perrin's argument for determining Avogadro's number makes an analogy between particles in the liquid and molecules in the atmosphere. The thinning of air at high altitudes depends on the balance between the gravitational force pulling the molecules down and their thermal motion forcing them up. The relationship between the weight of the particles and the height of the atmosphere would be the same for Brownian particles suspended in water. Perrin counted particles of gum mastic at different heights in his water sample and inferred the mass of atoms from the rate of decrease. He then divided the result into the molar weight of atoms to determine Avogadro's number. After Perrin, few scientists could disbelieve the existence of atoms.

Electric properties of atoms. While atomic theory was set back by the failure of scientists to accept simple physical ideas like the diatomic atom and the kinetic theory of gases, it was also delayed by the preoccupation of physicists with mechanics for almost 200 years, from Newton to the 20th century. Nevertheless, several 19th-century investigators, working in the relatively ignored fields of electricity, magnetism, and optics, provided important clues about the interior of the atom. The studies in electrodynamics made by the British physicist Michael Faraday and those of Maxwell indicated for the first time that something existed apart from palpable matter, and data obtained by Gustav Robert Kirchhoff of Germany about elemental spectral lines raised questions that would only be answered in the 20th century by quantum mechanics.

Until Faraday's electrolysis experiments, scientists had had no conception of the nature of the forces binding atoms together in a molecule. Faraday concluded that electrical forces existed inside the molecule after he had produced an electric current and a chemical reaction in a solution with the electrodes of a voltaic cell. No matter what solution or electrode material he used, a fixed quantity of current sent through an electrolyte always caused a specific amount of material to form on an electrode of the electrolytic cell. Faraday concluded that each ion of a given chemical compound has exactly the same charge. Later, he discovered that the ionic charges are integral multiples of a single unit of charge, never fractions.

Faraday's discoveries

On the practical level, Faraday did for charge what Dalton had done for the chemical combination of atomic masses. That is to say, Faraday demonstrated that it takes a definite amount of charge to convert an ion of an element into an atom of the element and that the amount of charge depends on the element used. The unit of charge that releases a gram atomic weight of a simple ion is called the faraday in his honour. For example, one faraday of charge passing through water releases one gram of hydrogen and eight grams of oxygen. In this manner, Faraday gave scientists a rather precise value for the ratios of the masses of atoms to the electric charges of ions. The ratio of the mass of the hydrogen atom to the charge of the electron was found to be 1.035×10^{-8} kilogram per coulomb. Faraday did not know the size of his electrolytic unit of charge in units such as coulombs any more than Dalton knew the magnitude of his unit of atomic weight in grams. Nevertheless, scientists could determine the ratio of these units easily.

More significantly, Faraday's work was the first to imply the electrical nature of matter and the existence of subatomic particles and a fundamental unit of charge. Faraday wrote: "The atoms of matter are in some way endowed or associated with electrical powers, to which they owe their most striking qualities, and amongst them their mutual chemical affinity." Faraday did not, however, conclude that atoms cause electricity.

Light and spectral lines. In 1865 Maxwell unified the laws of electricity and magnetism in his publication "A Dynamical Theory of the Electromagnetic Field." In this paper, he concluded that light is an electromagnetic wave. His theory was confirmed by the German physicist Heinrich Hertz, who produced radio waves with sparks in 1887. With light understood as an electromagnetic wave, Maxwell's theory could be applied to the emission of light from atoms. The theory failed, however, to describe spectral lines and the fact that atoms do not lose all their energy when they radiate light. The problem was not with Maxwell's theory of light itself but rather with its description of the oscillating electron currents generating light. Only quantum mechanics could explain this behaviour (see below *The laws of quantum mechanics*).

By far the richest clues about the structure of the atom came from spectral lines. Mounting a particularly fine prism on a telescope, the German physicist and optician Joseph von Fraunhofer had discovered between 1814 and 1824 hundreds of dark lines in the spectrum of the Sun. He labeled the most prominent of these lines with the letters *A* through *G*. Together, they are now called Fraunhofer lines. A generation later, Kirchhoff heated different elements to incandescence in order to study the different coloured vapours emitted. Observing the vapours through a spectroscope, he discovered that each element has a unique and characteristic pattern of spectral lines. Each element produces the same set of identifying lines, even when it is combined chemically with other elements. In 1859 Kirchhoff and the German chemist Robert Wilhelm Bunsen discovered two new elements—cesium and rubidium—by first observing their spectral lines.

Spectral lines and atomic structure

Johann Jakob Balmer, a Swiss secondary-school teacher with a penchant for numerology, studied hydrogen's spectral lines and found a constant relationship between the wavelengths of the element's four visible lines (see Figure 2). In 1885 he published a generalized mathematical formula for all of the lines of hydrogen. The Swedish physicist Johannes Rydberg extended Balmer's work in 1890 and found a general rule applicable to many elements. Soon more series were discovered elsewhere in the spectrum of hydrogen and in the spectra of other elements as well. Stated in terms of the frequency of the light rather than its wavelength, the formula may be expressed:

$$\nu = R\,(1/n^2 - 1/m^2).$$

Here, n and m are integers and R is a constant. In the Balmer lines, m is equal to 2 and n takes on the values 3, 4, 5, and 6.

Adapted from W. Finkelnburg, *Structure of Matter* (1964); Springer-Verlag, Heidelberg

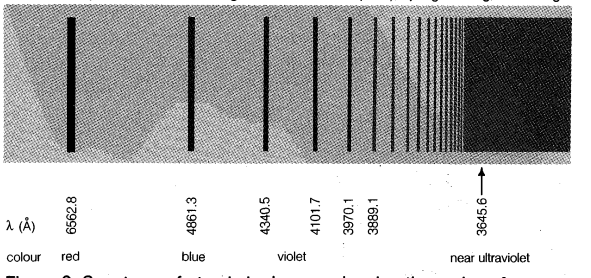


Figure 2: Spectrum of atomic hydrogen, showing the series of Balmer lines.

Discovery of electrons. During the 1880s and '90s, scientists searched cathode rays for the carrier of the electrical properties in matter. Their work culminated in J.J. Thomson's discovery of the electron in 1897. The existence of the electron showed that the 2,000-year-old conception of the atom as a homogeneous particle was wrong and that in fact the atom has a complex structure.

Cathode-ray studies began in 1854 when Heinrich Geissler, a glassblower and technical assistant to the German physicist Julius Plücker, improved the vacuum tube. Plücker discovered cathode rays in 1858 by sealing two electrodes inside the tube, evacuating the air, and forcing electric current between the electrodes. He found a green glow on the wall of his glass tube and attributed it to rays emanating from the cathode. In 1869, with better vacuums, Plücker's pupil Johann W. Hittorf saw a shadow cast by an object placed in front of the cathode. The shadow proved that the cathode rays originated from the cathode. The British physicist and chemist Sir William Crookes investigated cathode rays in 1879 and found that they were bent by a magnetic field; the direction of deflection

Crookes's study of cathode rays

suggested that they were negatively charged particles. As the luminescence did not depend on what gas had been in the vacuum or what metal the electrodes were made of, he surmised that the rays were a property of the electric current itself. As a result of Crookes's work, cathode rays were widely studied, and the tubes came to be called Crookes tubes.

Although Crookes believed that the particles were electrified charged particles, his work did not settle the issue of whether cathode rays were particles or radiation similar to light. By the late 1880s the controversy over the nature of cathode rays had divided the physics community into two camps. Most French and British physicists, influenced by Crookes, thought that cathode rays were electrically charged particles because they were affected by magnets. Most German physicists, on the other hand, believed that the rays were waves because they traveled in straight lines and were unaffected by gravity. A crucial test of the nature of the cathode rays was how they would be affected by electric fields. Heinrich Hertz, the aforementioned German physicist, reported that the cathode rays were not deflected when they passed between two oppositely charged plates in an 1892 experiment. The English physicist Thomson thought Hertz's vacuum might have been faulty and that residual gas might have reduced the effect of the electric field on the cathode rays.

Thomson repeated Hertz's experiment with a better vacuum in 1897. He directed the cathode rays between two parallel aluminum plates to the end of a tube where they were observed as luminescence on the glass. When the top aluminum plate was negative, the rays moved down; when the upper plate was positive, the rays moved up. The deflection was proportional to the difference in potential between the plates. With both magnetic and electric deflections observed, it was clear that cathode rays were negatively charged particles. Thomson's discovery established the particulate nature of electricity. Accordingly, he called his particles electrons.

Determining the properties of electrons

From the magnitude of the electrical and magnetic deflections, Thomson could calculate the ratio of mass to charge for the electrons. This ratio was known for atoms from electrochemical studies. Measuring and comparing it to the number for an atom, he discovered that the mass of the electron was very small, merely $^1/_{1836}$ that of a hydrogen ion. When scientists realized that an electron was virtually 1,000 times lighter than the smallest atom, they understood how cathode rays could penetrate metal sheets and how electric current could flow through copper wires. In deriving the mass-to-charge ratio, Thomson had calculated the electron's velocity. It was $^1/_{10}$ the speed of light, thus amounting to roughly 30,000 kilometres per second (18,000 miles per second). Thomson emphasized that "we have in the cathode rays matter in a new state, a state in which the subdivision of matter is carried very much further than in the ordinary gaseous state; a state in which all matter, that is, matter derived from different sources such as hydrogen, oxygen, etc., is of one and the same kind; this matter being the substance from which all the chemical elements are built up." Thus, the electron was the first subatomic particle identified, the smallest and the fastest bit of matter known at the time.

Millikan's electron charge experiment

In 1910 and 1911 the American physicist Robert Andrews Millikan greatly improved a method employed by Thomson for measuring the electron charge directly. Millikan produced microscopic oil droplets and observed them falling in the space between two electrically charged plates. Some of the droplets became charged and could be suspended by a delicate adjustment of the electric field. Millikan knew the weight of the droplets from their rate of fall when the electric field was turned off. From the balance of the gravitational and electrical forces, he could determine the charge on the droplets. He could find charges only in integral multiples of a quantity that in contemporary units is 1.602×10^{-19} coulomb. Millikan's electron charge experiment was the first to detect and measure the effect of an individual subatomic particle. Besides confirming the particulate nature of electricity, his experiment also supported previous determinations of Avogadro's number. Avogadro's number times the unit of charge gives Faraday's constant, the amount of charge required to electrolyze one mole of a chemical ion.

Identification of positive ions. In addition to electrons, positively charged particles also emanate from the anode in an energized Crookes tube. The German physicist Wilhelm Wien analyzed these positive rays in 1898 and found that the particles have a mass-to-charge ratio more than 1,000 times larger than that of the electron. Because the ratio of the particles is also comparable to the mass-to-charge-ratio of the residual atoms in the discharge tubes, scientists suspected that the rays were actually ions from the gases in the tube.

In 1913 Thomson refined Wien's apparatus to separate different ions and measure their mass-to-charge ratio on photographic plates. He sorted out the many ions in various charge states produced in a discharge tube. When he conducted his atomic mass experiments with neon gas, he found that a beam of neon atoms subjected to electric and magnetic forces split into two parabolas instead of one on a photographic plate. Chemists had assumed the atomic weight of neon was 20.2, but the traces on Thomson's photographic plate suggested atomic weights of 20.0 and 22.0, with the former parabola much stronger than the latter. He concluded that neon consisted of two stable isotopes: primarily neon-20, with a small percentage of neon-22. Eventually, a third isotope, neon-21, was discovered in very small quantities. It is now known that 1,000 neon atoms will contain 909 of neon-20, 88 of neon-22, and 3 of neon-21. Dalton's assumptions that all atoms of an element have an identical mass and that the atomic weight of an element is its mass were thus disproved. Today, the atomic weight of an element is recognized as the weighted average of the masses of its isotopes.

Development of the mass spectrograph

Francis William Aston, an English physicist, improved Thomson's technique when he developed the mass spectrograph in 1919. This device spread out the beam of positive ions into a "mass spectrum" of lines similar to the way light is separated into a spectrum. Aston analyzed about 50 elements over the next six years and discovered that most have isotopes.

Discovery of radioactivity. Like Thomson's discovery of the electron, the discovery of radioactivity in uranium by the French physicist Henri Becquerel in 1896 forced scientists to radically change their ideas about atomic structure. Radioactivity demonstrated that the atom was neither indivisible nor immutable. Instead of serving merely as an inert matrix for electrons, the atom could change form and emit an enormous amount of energy. Furthermore, radioactivity itself became an important tool for revealing the interior of the atom.

The German physicist Wilhelm Conrad Röntgen had discovered X rays in 1895, and Becquerel thought they might be related to fluorescence and phosphorescence, processes in which substances absorb and emit energy as light. In the course of his investigations, Becquerel stored some photographic plates and uranium salts in a desk drawer. Expecting to find the plates only lightly fogged, he developed them and was surprised to find sharp images of the salts. He then began experiments that showed that uranium salts emit a penetrating radiation independent of external influences. Becquerel also demonstrated that the radiation could discharge electrified bodies. In this case, discharge means the removal of electric charge, and it is now understood that the radiation ionizing molecules of air allows the air to conduct an electric current. Early studies of radioactivity relied on measuring ionization power or on observing the effects of radiation on photographic plates.

In 1898 the French physicists Pierre and Marie Curie discovered the strongly radioactive elements polonium and radium, which occur naturally in uranium minerals. Marie coined the term radioactivity for the spontaneous emission of ionizing, penetrating rays by certain atoms (see below).

Different forms of ionizing radiation

Experiments conducted by the British physicist Ernest Rutherford in 1899 showed that radioactive substances emit more than one kind of ray. It was determined that part of the radiation is 100 times more penetrating than the rest and can pass through aluminum foil $^1/_{50}$ of a

millimetre thick. Rutherford named the less penetrating emanations alpha rays and the more powerful ones beta rays, after the first two letters of the Greek alphabet. Investigators who, in 1899, found that beta rays were deflected by a magnetic field concluded that they are negatively charged particles similar to cathode rays. In 1903 Rutherford found that alpha rays were deflected slightly in the opposite direction, showing that they are massive, positively charged particles. Much later, Rutherford proved that alpha rays are nuclei of helium atoms by collecting the rays in an evacuated tube and detecting the buildup of helium gas over several days. A third kind of radiation was identified by the French chemist Paul Villard in 1900. Designated as the gamma ray, it is not deflected by magnets and is much more penetrating than alpha particles. Gamma rays were later shown to be a form of electromagnetic radiation, like light or X rays, but with much shorter wavelengths. Because of these shorter wavelengths, gamma rays have higher frequencies and are even more penetrating than X rays. In 1902, while studying the radioactivity of thorium, Rutherford and the English chemist Frederick Soddy discovered that radioactivity was associated with changes inside the atom that transformed thorium into a different element. They found that thorium continually generates a chemically different substance that is intensely radioactive. The radioactivity eventually makes the new element disappear. Watching the process, Rutherford and Soddy formulated the exponential decay law, which states that a fixed fraction of the element will decay in each unit of time. For example, half of the thorium product decays in four days, half the remaining sample in the next four days, and so on.

Law of exponential decay

Until the 20th century, physicists had studied such subjects as mechanics, heat, and electromagnetism that they could understand by applying common sense or by extrapolating from everyday experiences. The discovery of the electron and radioactivity, however, showed that classical Newtonian mechanics could not explain phenomena at atomic and subatomic levels. As the primacy of classical mechanics crumbled during the early 20th century, quantum mechanics was developed to replace it. Since then, experiments and theories have led physicists into a world that is often extremely abstract and seemingly contradictory.

MODELS OF ATOMIC STRUCTURE

Thomson's discovery of the negatively charged electron had raised theoretical problems for physicists as early as 1897, because atoms as a whole are electrically neutral. Where was the neutralizing positive charge and what held it in place? Between 1903 and 1907 Thomson tried to solve the mystery by adapting an atomic model that had been first proposed by Lord Kelvin in 1902. According to this theoretical system, often referred to as the "plum pudding" model, the atom is a sphere of uniformly distributed positive charge about one angstrom in diameter. Electrons are embedded in a regular pattern like raisins in a plum pudding to neutralize the positive charge. The advantage of the Thomson atom was that it was inherently stable: if the electrons were displaced, they would attempt to return to their original positions. In another contemporary model, the atom resembled the solar system or the planet Saturn, with rings of electrons surrounding a concentrated positive charge. The Japanese physicist Hantaro Nagaoka, in particular, developed the "Saturnian" system in 1904. The atom, as postulated in this model, was inherently unstable because, by radiating continuously, the electron would gradually lose energy and spiral into the nucleus. No electron could thus remain in any particular orbit indefinitely.

The so-called plum pudding model

Rutherford's nuclear model. Rutherford overturned Thomson's model in 1911 with his well-known gold foil experiment in which he demonstrated that the atom has a tiny, massive nucleus. Five years earlier Rutherford had noticed that alpha particles, beamed through a hole onto a photographic plate, would make a sharp-edged picture, while alpha particles beamed through a sheet of mica only 20 micrometres (or about 0.002 centimetre) thick would make an impression with blurry edges. For some particles, the blurring corresponded to a two-degree deflection. Remembering those results, Rutherford had his postdoctoral fellow, Hans Geiger, and an undergraduate student, Ernest Marsden, refine the experiment. The young physicists beamed alpha particles through gold foil and detected them as flashes of light or scintillations on a screen. The gold foil was only 0.00004 centimetre thick. Most of the alpha particles went straight through the foil, but some were deflected by the foil and hit a spot on a screen placed off to one side. Geiger and Marsden found that about one in 20,000 alpha particles had been deflected 45° or more. Rutherford asked why so many alpha particles passed through the gold foil while a few were deflected so greatly. "It was almost as incredible as if you fired a 15-inch shell at a piece of tissue paper, and it came back to hit you," Rutherford said later. "On consideration, I realized that this scattering backwards must be the result of a single collision, and when I made calculations I saw that it was impossible to get anything of that order of magnitude unless you took a system in which the greater part of the mass of the atom was concentrated in a minute nucleus. It was then that I had the idea of an atom with a minute massive centre carrying a charge."

Many physicists distrusted Rutherford's nuclear model because it was difficult to reconcile with the chemical behaviour of atoms. The model suggested that the charge on the nucleus was the most important characteristic of the atom, determining its structure. On the other hand, Mendeleyev's periodic table of the elements had been organized according to the atomic masses of the elements, implying that the mass was responsible for the structure and chemical behaviour of atoms.

Primacy of the positive charge on the nucleus

Moseley's X-ray studies. Henry Gwyn Jeffreys Moseley, a young English physicist killed in World War I, confirmed that the positive charge on the nucleus revealed more about the fundamental structure of the atom than Mendeleyev's atomic mass. Moseley studied the spectral lines emitted by heavy elements in the X-ray region of the electromagnetic spectrum. He built on the work done by several other British physicists—Charles Glover Barkla, who had studied X rays produced by the impact of electrons on metal plates, and Sir William Bragg and his son Lawrence, who had developed a precise method of using crystals to reflect X rays and measure their wavelength by diffraction. Moseley used a crystal of potassium ferrocyanide as a diffraction grating to examine the spectra of X rays produced by different metals. He arranged his crystal so that he could control and vary the angle between the crystal face and the X-ray beam. The X rays from each element were reflected at a unique set of angles. By measuring the angle, Moseley was able to obtain the wavelength of the X ray hitting the crystal.

Moseley found that the X rays radiated by each element have a characteristic frequency that differs according to a regular pattern. The difference in frequency is not governed by Mendeleyev's change in mass, however, but rather by the change in charge on the nucleus. He called this the atomic number. In his first experiments, conducted in 1913, Moseley used the K series of X rays (X radiation associated with the K energy state of an atom) and studied the elements up to zinc. The following year he extended his work up to gold in the periodic table, using the L series of X rays (X radiation associated with the L atomic-energy state). Moseley was conducting his research at the same time that the Danish theorist Niels Bohr was developing his quantum shell model of the atom (see below). The two conferred and shared data as their work progressed, and Moseley framed his equation in terms of Bohr's theory. Moseley presented formulas for the X-ray frequencies that were closely related to Bohr's formulas for the spectral lines in a hydrogen atom. Moseley showed that the frequency of a line in the X-ray spectrum is proportional to the square of the charge on the nucleus. The constant of proportionality depends on whether the X ray is in the K or L series. This is the same relationship that Bohr used in his formula applied to the Lyman and Balmer series of spectral lines. The regularity of the differences in X-ray frequencies allowed Moseley to order the elements by atomic number from aluminum to gold. He observed

Collaboration with Niels Bohr

that, in some cases, the order by atomic weights was incorrect. For example, cobalt has a larger atomic mass than nickel, but Moseley found that it has atomic number 27, while nickel has 28. When Mendeleyev constructed the periodic table, he based his system on the atomic masses of the elements and had to put cobalt and nickel out of order to make the chemical properties fit better. In a few places where Moseley found more than one integer between elements, he predicted correctly that a new element would be discovered. Because there is just one element for each atomic number, scientists could be confident for the first time of the completeness of the periodic table; no unexpected new elements would be discovered.

Bohr's shell model. In 1913 Bohr proposed his quantized shell model of the atom to explain how electrons can have stable orbits around the nucleus. The motion of the electrons in the Rutherford model was unstable because, according to classical mechanics and electromagnetic theory, any charged particle moving on a curved path emits electromagnetic radiation; thus, the electrons would lose energy and spiral into the nucleus. To remedy the stability problem, Bohr modified the Rutherford model by requiring that the electrons move in orbits of fixed size and energy. The energy of an electron depends on the size of the orbit and is lower for smaller orbits. Radiation can occur only when the electron jumps from one orbit to another. The atom will be completely stable in the state with the smallest orbit, since there is no orbit of lower energy into which the electron can jump.

Electron orbits of fixed size and shape

Bohr's starting point was to realize that classical mechanics by itself could never explain the atom's stability. A stable atom has a certain size so that any equation describing it must contain some fundamental constant or combination of constants with a dimension of length. The classical fundamental constants—namely, the charges and the masses of the electron and the nucleus—cannot be combined to make a length. Bohr noticed, however, that the quantum constant formulated by the German physicist Max Planck (see below) has dimensions which, when combined with the mass and charge of the electron, produce a measure of length. Numerically, the measure is close to the known size of atoms. This encouraged Bohr to use Planck's constant in searching for a theory of the atom.

Planck's constant

Planck had introduced his constant in 1900 in a formula explaining the light radiation emitted from heated bodies. According to classical theory, comparable amounts of light energy should be produced at all frequencies. This is not only contrary to observation but also implies the absurd result that the total energy radiated by a heated body should be infinite. Planck postulated that energy can only be emitted or absorbed in discrete amounts, which he called quanta (the Latin word for "how much"). The energy quantum is related to the frequency of the light by a new fundamental constant, h. When a body is heated, its radiant energy in a particular frequency range is, according to classical theory, proportional to the temperature of the body. With Planck's hypothesis, however, the radiation can occur only in quantum amounts of energy. If the radiant energy is less than the quantum of energy, the amount of light in that frequency range will be reduced. Planck's formula correctly describes radiation from heated bodies. Planck's constant has the dimensions of action, which may be expressed as units of energy multiplied by time, units of momentum multiplied by length, or units of angular momentum. For example, Planck's constant can be written as $h = 6.6 \times 10^{-34}$ joule seconds or 6.6×10^{-34} kilogram-metre/second-metres.

Using Planck's constant, Bohr obtained an accurate formula for the energy levels of the hydrogen atom (see Figure 3). He postulated that the angular momentum of the electron is quantized—*i.e.*, it can have only discrete values. He assumed that otherwise electrons obey the laws of classical mechanics by traveling around the nucleus in circular orbits. Because of the quantization, the electron orbits have fixed sizes and energies. The orbits are labeled by an integer, the quantum number n.

With his model, Bohr explained how electrons could jump from one orbit to another only by emitting or absorbing energy in fixed quanta. For example, if an electron jumps one orbit closer to the nucleus, it must emit energy equal to the difference of the energies of the two orbits. Conversely, when the electron jumps to a larger orbit, it must absorb a quantum of light equal in energy to the difference in orbits.

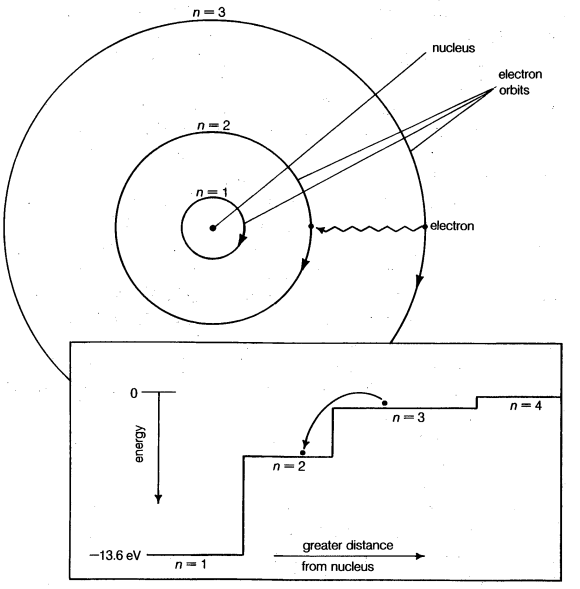


Figure 3: *The Bohr atom.*
The electron travels in circular orbits around the nucleus. The orbits have quantized sizes and energies. Energy is emitted from the atom when the electron jumps from one orbit to another closer to the nucleus. Shown here is the first Balmer transition, in which an electron jumps from orbit $n = 3$ to orbit $n = 2$, producing a photon of red light with an energy of 1.89 eV and a wavelength of 656×10^{-9} m.

The inherent stability of atoms

Bohr's model accounts for the stability of atoms because the electron cannot lose more energy than it has in the smallest orbit, the one with $n = 1$. The model also explains the Balmer formula for the spectral lines of hydrogen. The frequency of the light is related to its energy by Einstein's formula $E = hv$. The light energy is calculated from the difference in energies between the two orbits. The Balmer formula can be expressed as the difference of two terms, each term giving the energy of an orbit. Bohr's model not only explains the form of the Balmer formula but also accurately gives the value of the constant of proportionality R.

The usefulness of Bohr's theory extends beyond the hydrogen atom. Bohr himself noted that the formula also applies to the singly ionized helium atom, which, like hydrogen, has a single electron. The nucleus of the helium atom has twice the charge of the hydrogen nucleus, however. In Bohr's formula the charge of the electron is raised to the fourth power. Two of those powers stem from the charge on the nucleus; the other two come from the charge on the electron itself. Bohr modified his formula for the hydrogen atom to fit the helium atom by doubling the charge on the nucleus. Moseley applied Bohr's formula with an arbitrary atomic charge Z to explain the K- and L-series X-ray spectra of heavier atoms. The German physicists James Franck and Gustav Hertz confirmed the existence of quantum states in atoms in experiments reported in 1914. They made atoms absorb energy by bombarding them with electrons. The atoms would only absorb discrete amounts of energy from the electron beam. When the energy of an electron was below the threshold for producing an excited state, the atom would not absorb any energy.

Limitations of Bohr's model

Bohr's theory had major drawbacks, however. Except for the spectra of X rays in the K and L series, it could not explain properties of atoms having more than one electron. The binding energy of the helium atom, which has two electrons, was not understood until the development of quantum mechanics. Several features of the spectrum were inexplicable even in the hydrogen atom. High-resolution spectroscopy shows that the individual spectral lines

of hydrogen are divided into several closely spaced, fine lines. In a magnetic field the lines split even further. The German physicist Arnold Sommerfeld modified Bohr's theory by quantizing the shapes and orientations of orbits to introduce additional energy levels corresponding to the fine spectral lines. The quantization of the orientation of the angular-momentum vector was confirmed in an experiment in 1922 by other German physicists, Otto Stern and Walter Gerlach. They passed a beam of silver atoms through a nonhomogeneous magnetic field, one that is stronger on one side than on the other. The field deflected the atoms according to the orientation of their magnetic moments. (The magnetic moment of an object such as an atom or a compass needle is the measure of its interaction with a magnetic field. The moment points in some direction and is associated in classical physics with orbital currents and the angular momentum of charges.) In their experiment, Stern and Gerlach found only two deflections, not the continuous distribution of deflections that would have been seen if the magnetic moment had been oriented in any direction. Thus, it was determined that the magnetic moment and the angular momentum of an atom can have only two orientations. The discrete orientations of the orbits explain some of the magnetic field effects—namely, the so-called normal Zeeman effect, which is the splitting of a spectral line into three sublines. These sublines correspond to quantum jumps in which the angular momentum along the magnetic field is increased by one unit. decreased by one unit, or left unchanged.

Normal Zeeman effect

An additional quantum number was needed to complete the description of electrons in an atom. In 1925 Samuel A. Goudsmit and George E. Uhlenbeck, two graduate students in physics at the University of Leiden, in The Netherlands, added a quantum number to account for the fact that some spectral lines are divided into more sublines than can be explained with the original quantum numbers. Goudsmit and Uhlenbeck postulated that an electron has an internal spinning motion and that the corresponding angular momentum is one-half of the orbital angular momentum quantum. An electron has a magnetic moment, and its energy depends on whether the spin is aligned with or against the magnetic field. Independently, the Austrian-born physicist Wolfgang Pauli also suggested adding a two-valued quantum number for electrons, but for different reasons. He needed this additional quantum number to formulate his exclusion principle, which serves as the atomic basis of the periodic table and the chemical behaviour of the elements. According to the exclusion principle, one electron at most can occupy an orbital, taking into account all the quantum numbers. Pauli was led to this principle by the observation that an alkali in a magnetic field has a number of orbitals in the shell equal to the number of electrons that must be added to make the next noble gas. These numbers are twice the number of orbits available if the angular momentum and its orientation are considered alone.

In spite of these modifications, Bohr's model seemed to be a dead end by the early 1920s. It did not explain most fine spectral lines or the anomalous Zeeman effect, which is a complicated type of spectral line splitting that sometimes involves up to 15 sublines. (Its name notwithstanding, the anomalous Zeeman effect is more common than the aforementioned normal Zeeman effect.) Efforts to generalize the model to multielectron atoms had proved futile, and physicists despaired of ever explaining them.

The laws of quantum mechanics. Within a few short years scientists developed a consistent theory of the atom that explained its fundamental structure and its interactions. Crucial to the development of the theory was new evidence indicating that light and matter have both wave and particle characteristics at the atomic and subatomic levels. Theoreticians had objected to the fact that Bohr had used an ad hoc hybrid of classical Newtonian dynamics for the orbits and some quantum postulates for limiting the motion. The new theory ignored the fact that electrons are particles and treated them as waves. By 1926, physicists had developed the laws of quantum mechanics, also called wave mechanics, to explain atomic and subatomic phenomena.

The duality between the wave and particle nature of light was highlighted by the American physicist Arthur H. Compton in an X-ray scattering experiment conducted in 1922. Compton showed that X rays scatter from electrons exactly like particles. The X rays have discrete amounts of momentum, which is a property of particles. When X rays are scattered, their momentum is partially transferred to the electrons. The recoil electron takes some energy from an X ray, and as a result the X ray frequency is shifted. Both the discrete amount of momentum and the frequency shift of the light scattering are completely at variance with classical electromagnetic theory.

Louis-Victor de Broglie, a French physicist, had proposed in his 1923 doctoral thesis that all matter and radiations have both particle- and wavelike characteristics. Until the emergence of the quantum theory, physicists had assumed that matter was distinct from energy and followed different laws: energy radiations were waves and matter was particulate. Planck's theory was the first to propose that radiation has characteristics of both waves and particles. Believing in the symmetry of nature, Broglie ended the wave-particle dichotomy by applying Einstein's mass-energy formula. Using the old-fashioned word corpuscles for particles, Broglie wrote, "For both matter and radiations, light in particular, it is necessary to introduce the corpuscle concept and the wave concept at the same time. In other words, the existence of corpuscles accompanied by waves has to be assumed in all cases." Broglie's conception was an inspired one, but it had no experimental or theoretical foundation. The Austrian physicist Erwin Schrödinger had to supply the theory.

Broglie's wave-particle concept

Schrödinger's wave equation. In 1926 Schrödinger produced the mathematical wave equation that established quantum mechanics in widely applicable form. To understand how a wave equation is used, it is helpful to think of an analogy with the vibrations of a bell, violin string, or drumhead. These vibrations are governed by a wave equation, since the motion can propagate as a wave from one side of the object to the other. Certain vibrations in these objects are simple modes that are easily excited and have definite frequencies. For example, the motion of the lowest vibrational mode in a drumhead is in phase all over the drumhead with a pattern that is uniform around it; the highest amplitude of the vibratory motion is in the middle of the drumhead. In more complicated, higher frequency modes, the motion on different parts of the vibrating drumhead are out of phase, with inward motion on one part at the same time that there is outward motion on another.

Schrödinger postulated that the electrons in an atom should be treated like the waves on the drumhead. The different energy levels of atoms are identified with the simple vibrational modes of the wave equation. The equation is solved to find these modes, and then the energy of an electron is obtained from the frequency of the mode and Einstein's formula. Schrödinger's wave equation gives the same energies as Bohr's original formula but with a much more precise description of an electron in an atom. The lowest energy level of the hydrogen atom, called the ground state, is analogous to the motion in the lowest vibrational mode of the drumhead. In the atom the electron wave is uniform in all directions from the nucleus, is peaked at the centre of the atom, and has the same phase everywhere. Higher energy levels in the atom have waves that are peaked at greater distances from the nucleus. Like the vibrations in the drumhead, the waves have peaks and nodes that may form a complex shape. The different shapes of the wave pattern explain the quantum numbers of the energy levels, including the quantum numbers for angular momentum and its orientation.

The year before Schrödinger produced his wave theory, the German physicist Werner Heisenberg published a mathematically equivalent system using matrix algebra to describe energy levels and their transitions. Today, physicists use Schrödinger's system because it can be visualized more easily.

In 1929 the Norwegian physicist Egil Hylleraas applied the Schrödinger equation to the helium atom with two electrons. He obtained only an approximate solution, but

his energy calculation was quite accurate. With Hylleraas' explanation of the two-electron atom, physicists realized that the Schrödinger equation could be a powerful mathematical tool for describing nature on the atomic level, even if exact solutions could not be obtained.

The existence of antiparticles. The English physicist P.A.M. Dirac introduced a new equation for the electron in 1928. The Schrödinger equation does not satisfy the principles of relativity, and so it can be used to describe only those phenomena in which the particles move much more slowly than the velocity of light. In order to satisfy the conditions of relativity, Dirac was forced to postulate not one but four distinct wave functions for the electron. Two of these correspond to the two spin orientations. The remaining components allowed additional states of the electron that had not yet been observed. Dirac interpreted them as antiparticles with a charge opposite to that of electrons. The discovery of the positron in 1932 by the American physicist Carl David Anderson proved the existence of antiparticles and was a triumph for Dirac's theory.

The discovery of the positron

After his discovery, subatomic particles could no longer be considered immutable. Given enough energy, electrons and positrons can be created from a few particles in a vacuum tube. They also can annihilate each other and disappear into some other form of energy. The history of subatomic physics from this point has been much the story of finding new kinds of particles that can be created in vacuums.

ADVANCES IN NUCLEAR AND SUBATOMIC PHYSICS

The 1920s witnessed further advances in nuclear physics with Rutherford's discovery of induced radioactivity. Bombardment of light nuclei by alpha particles produced new radioactive nuclei. In 1928 the Russian-born American physicist George Gamow explained the lifetimes in alpha radioactivity using the Schrödinger equation.

Discovery of neutrons. The constitution of the nucleus was poorly understood at the time because the only known particles were the electron and the proton. It had been established that nuclei are typically about twice as heavy as can be accounted for by protons alone and thus have to contain more than just such particles. A consistent theory was impossible until the English physicist James Chadwick discovered the neutron in 1932. He found that alpha particles reacted with beryllium nuclei, ejecting neutral particles with nearly the same mass as protons. Almost all nuclear phenomena can be understood in terms of a nucleus composed of neutrons and protons. Surprisingly, the neutrons and protons in the nucleus behave to a large extent as though they were in independent wave functions, just like the electrons in an atom. Each neutron or proton is described by a wave pattern with peaks and nodes and angular momentum quantum numbers. The theory of the nucleus based on these independent wave functions is called the shell model. It was introduced in 1948 by Maria Goeppert Mayer of the United States and J. Hans D. Jensen of West Germany, and it developed in succeeding decades into a comprehensive theory of the nucleus.

Shell model of the nucleus

The interactions of neutrons with nuclei had been studied during the mid-1930s by the Italian-born American physicist Enrico Fermi and others. Nuclei readily capture neutrons, which, unlike protons or alpha particles, are not repelled from the nucleus by a positive charge. When a neutron is captured, the new nucleus has one higher unit of atomic mass. If a nearby isotope of that atomic mass is more stable, the new nucleus will be radioactive, convert the neutron to a proton, and assume the more stable form.

Nuclear fission was discovered by the German chemists Otto Hahn and Fritz Strassmann in 1938. In fission, a uranium nucleus captures a neutron and gains enough energy to trigger the inherent instability of the nucleus, which splits into two lighter nuclei of roughly equal size. The fission process releases more neutrons, which can be used to produce further fissions (see below *Nuclear fission*). The first nuclear reactor, a device designed to permit controlled fission chain reactions, was constructed at the University of Chicago under Fermi's direction, and the first self-sustaining chain reaction was achieved in this reactor in 1942. In 1945 American scientists produced the first atomic bomb, which used uncontrolled fission reactions in either uranium or the artificial element plutonium.

Quantum field theory. Dirac not only proposed the relativistic equation for the electron but also initiated the relativistic treatment of interactions between particles known as quantum field theory. An important aspect of quantum field theory is that interactions can extend only over a given distance if there is a particle to carry the force. The electromagnetic force, which operates over a long distance, is carried by a particle called the photon, the light quantum. In 1934 the Japanese physicist Yukawa Hideki proposed that there should be a particle that carries the nuclear force as well. Because the force is short-range, the particle should be massive. Massive particles were indeed found in cosmic rays, but these did not have the correct interaction properties. They were later dubbed muons. Evidence for Yukawa's particle, known as the pion, was found in cosmic-ray tracks in 1947 by the British physicist Cecil Frank Powell. The existence of the pion was confirmed when the particle was created in a particle accelerator in 1948.

Pions

Since then, the number of subatomic particles discovered has grown enormously. Most of them have been created and studied by means of accelerators that produce high-energy collisions between particles. The new particles, formed by the collision process, live only a short time before decaying into more stable particles. Particularly noteworthy among the many particles discovered since 1960 are those responsible for the interaction in beta radioactivity. In quantum field theory, beta radioactivity is a manifestation of an interaction called the weak force. The particles that transmit this force are known as W and Z particles. These messenger particles, or bosons, were discovered in 1983 during experiments at the European Organization for Nuclear Research (CERN).

Hadrons and quarks. Many of the newly discovered particles did not at first seem to have any specific role in subatomic physics. Particles of a class known as hadrons, which includes protons and neutrons, interact strongly with one another. They show patterns analogous to the patterns of energy levels in atoms. Just as the existence of energy levels in hydrogen is explained by the presence of the electron, the different hadrons are considered to be energy levels of a more fundamental particle inside them. The specific patterns of the hadrons have been explained by postulating a new fundamental constituent, the quark. Three quarks are thought to combine to form a proton, a neutron, or any of the massive hadrons known as baryons. A quark combines with an antiquark to form mesons such as the pion.

Quarks have never been observed, and physicists do not expect to find one. Presumably the forces between quarks are so strong that they cannot be separated from other quarks in a hadron. There are several indirect confirmations of the existence of quarks. In experiments conducted with high-energy electron accelerators in the 1960s, electrons that had been deflected by large angles in scattering from hadrons were observed. As in Rutherford's experiment, the large-angle deflection implies that there are very small objects inside the hadron surrounded by large electric fields. The small objects are presumed to be quarks. Physicists developed a quantum field theory known as quantum chromodynamics (QCD) during the mid-1970s to accommodate quarks and their peculiar properties. This theory explains qualitatively the confinement of quarks to hadrons. Physicists believe that the theory should explain all aspects of hadrons, but mathematical complications unfortunately prevent rigorous calculations.

Quantum chromodynamics

Atomic structure and interactions

ELECTRONS

As noted above, the electron was the first subatomic particle discovered. Its interactions determine atomic structure, the chemical behaviour of atoms in molecules, and the properties of larger aggregations of atoms such as bulk solids. There are four kinds of forces in nature, and the electron is subject to three of them—gravity, electromagnetism, and weak interaction. Only the electromagnetic

force is significant in determining the properties of atoms and their chemistry. Within the framework of the equations used to describe the motion of the electron, the only two numerical properties that need to be specified are the electron's mass and its charge. These are given in Table I along with other basic atomic constants.

There are four levels of complexity in the equations used to describe the properties of electrons. At the simplest level, classical equations such as Newton's equation are applied. The motion of the electron beam in a television tube is adequately described by classical physics.

The equations of classical physics, however, become invalid when one attempts to describe the motion within distances smaller than the de Broglie wavelength of the electron (see MECHANICS: *Quantum mechanics*). Atomic properties fall into this small-distance regime, and quantum equations must be used. The simplest quantum equation is the above-mentioned Schrödinger equation, which is valid when the velocity of the electron is small compared to the speed of light. The electron is described by a function that obeys a wave equation. The function may be visualized as a cloud; where the function is large, the cloud is dense and the electron's presence is strongly felt.

Atomic and chemical structure is well described by the Schrödinger wave functions when two additional nonclassical properties of the electron are included. The first property is spin, which is like an internal rotational motion of the electron. The magnitude of an electron's spin is fixed, but its orientation in space can vary. Spin appears in the Schrödinger equation as an attribute of the electron. There are, in fact, two separate wave functions associated with the probability that spin will be pointed in a particular direction or in the opposite direction. Other spin orientations are obtained by suitably combining the two functions to make intermediate directions. The other nonclassical property involved is the Pauli exclusion principle, which states in its general form that the wave function of identical particles must reverse sign when the coordinates of the particles are interchanged. As a consequence of the Pauli principle, two electrons cannot have the same wave function. This principle is extremely important in determining the structure of atoms, molecules, and bulk matter.

The third level of complexity in the description of the electron is Dirac's equation (see above), which is a quantum equation postulated to satisfy the requirements of Einstein's relativity. Any particle governed by the Dirac equation is called a fermion; the electron is the most familiar example. The mathematics of the Dirac equation require that its particles have two spin orientations and obey the generalized Pauli principle. Thus, all the properties needed for determining atomic structure are built automatically into the theory. Furthermore, the Dirac equation predicts that for each kind of fermion there is an oppositely charged antiparticle with the same mass. The electron's antiparticle is the positron. The electron can be made to disappear by combining with a positron. When the two particles annihilate each other, their rest energy is converted to gamma rays or some other form of energy. Electrons can also be created from other forms of energy, but always in association with positrons. According to the Dirac equation, charged fermions such as the electron have a magnetic moment pointing along the direction of spin. The magnetic moment of electrons is very close to the value predicted by the Dirac equation. The magnetism of permanent magnets arises from the combined effect of individual electrons having spins aligned in the same direction.

The electron as a fermion

The most sophisticated level of complexity in describing electrons is the theory of quantum electrodynamics. Using the Dirac and other equations, this theory builds a wave function not only for observable electrons but also for particles and quanta that may be created from a vacuum. The predictions of quantum electrodynamics deviate only slightly from the Dirac equation. For example, the magnetic moment of electrons is predicted to deviate by 0.1 percent from the Dirac value due to vacuum modifications. These deviations have been measured accurately and agree perfectly with quantum electrodynamics as far as experiments can determine.

Quantum electrodynamics

ELECTRONIC STRUCTURE OF ATOMS

Schrödinger's theory of atoms. The theory of the electronic structure of atoms is based for the most part on the Schrödinger equation, which in actuality is not a single equation. Each different physical system is described by its own equation, a partial differential equation that has as many variables as there are coordinates of interest in the application. For example, the Schrödinger equation for the hydrogen atom has three coordinates—the x, y, and z coordinates of the separation between the electron and the hydrogen nucleus. Similarly, there are six variables in the Schrödinger equation of the helium atom, since the positions of two electrons are required.

The Schrödinger equation has many solutions, each of which describes a possible state of the atom. Each state is specified by a function that depends on the coordinates of the particle(s) and by its associated energy. The function, called the wave function, is the mathematical representation of what is descriptively called the cloud. The wave function describes, among other things, the probability of an electron being at any given coordinate position. Electrons are more likely to be found in regions of the atom where the wave function is large. More precisely, the probability is proportional to the square of the wave function. At positions where the wave function goes through zero, the probability of finding electrons vanishes.

Normally, an atom is found in its ground state, *i.e.*, the state with the least energy or the most bound state. States with higher energy are called excited states. Atomic spectra are the light quanta emitted when an atom makes a transition from an excited state to one of lower energy.

The Schrödinger equation can only be solved exactly in special circumstances. Schrödinger himself found the solution for the hydrogen atom. It is important for physicists and chemists to understand the properties of the hydrogen atom obtained from the Schrödinger equation because these same properties appear in the behaviour of electrons in more complex atoms.

For atoms with more than one electron, no exact solutions of the Schrödinger equation are known. Nevertheless, it is possible to obtain very accurate numerical solutions using approximation methods. An approximation scheme introduced by the English physicist Douglas R. Hartree is the basis for most calculations and for the prevailing physical understanding of the wave mechanics of atoms. In this method it is assumed that the electrons move independently. The electrons are allowed to interact only through the average electric field made by combining the charge of the nucleus with the charge distribution of the other electrons. A Schrödinger equation for individual electrons moving in an average electric field must then be solved. Each electron has its own wave function, which is called an orbital.

The Hartree method

A technical difficulty with this method is that the electric field is not known ahead of time because it depends on the charge distribution of the electrons. In turn, the charge distribution can be determined only from the wave functions, which require prior knowledge of the field. The difficulty is overcome by solving the Schrödinger equation in successive approximations. First, one makes a guess for the field, finds the wave functions, and then uses the derived charge distribution to make a better approximation for the field. This iterative process is continued until the final charge and electric field distribution agree with the input to the Schrödinger equation. The Hartree method (sometimes called the Hartree–Fok method to give credit to V. Fok, a Soviet physicist who generalized Hartree's scheme) is widely used to describe electrons in atoms, molecules, and solids.

The hydrogen atom. Bohr's model of the hydrogen atom is rudimentary but nevertheless remarkably accurate in its predicted energy levels. The derivation of the energy formula is as follows. Bohr begins with the classical equation relating the velocity of a charged particle in a circular orbit, v, to the radius of the orbit r and the electric force constants. The equation, obtained by balancing the centrifugal force and the electric force, is

$$mv^2/r = e^2/4\pi\varepsilon_0 r^2.$$

Next, Bohr postulates that angular momentum is quantized in integer multiples of the reduced Planck's constant, $\hbar$. The angular momentum is then given by the formula

$$m_e v r = n\hbar,$$

where m_e is the mass of the electron, and n is an integer, which can have values 1, 2, 3, and so on. The two equations are combined to solve for the unknown quantities v and r. The resulting orbits have discrete radii, depending on n. The energies of the electrons in these orbits are given by Bohr's formula,

$$E_n = \frac{m_e e^4}{(4\pi\varepsilon_0)^2\, 2\hbar^2}\,\frac{1}{n^2}.$$

When the Schrödinger equation is solved for the hydrogen atom, one finds that the possible energies of the electron are the same as those in Bohr's model (see Figure 4). The wave functions associated with these energies are, however, quite different from the circular orbits hypothesized by Bohr. The lowest orbital is a spherically symmetric function. It falls off with distance from the nucleus r according to

$$\exp(-r/a_0).$$

Here, a_0 is a length constant, which happens to coincide with the radius of the smallest Bohr orbit. Known as the Bohr radius, it is given by the formula

$$a_0 = \frac{4\pi\varepsilon_0\hbar^2}{m_e e^2}.$$

The numerical value of a_0 is 0.53 angstrom; the meaning and values of the other quantities in this equation are given in Table 1. Since the wave function depends only on distance and not on angle, the electron cloud surrounding the nucleus has a spherical shape. Its radius is roughly one angstrom. The electron probability is highest at the nucleus and in its immediate vicinity. The probability falls off smoothly with distance from the nucleus, vanishing completely only at infinite distance. The wave function falls to inconsequentially small values at moderate distances, however. For example, the probability for the particle to be farther than $10a_0$ from the nucleus is only 1 in 30,000.

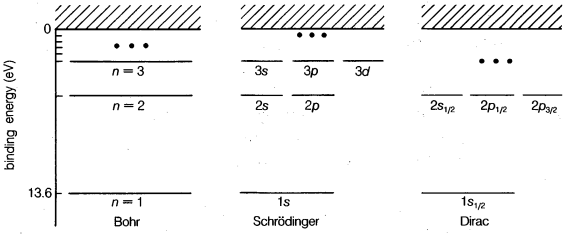


Figure 4: Energy levels of the hydrogen atom, according to Bohr's model and quantum mechanics using the Schrödinger equation and the Dirac equation.

Higher states of the hydrogen atom have more complex patterns in the wave function. All of the bound states of the hydrogen atom may be described by wave functions having the form

$$(\text{polynomial in } x, y, \text{ and } z)\exp(-r/n\,a_0),$$

where n is a positive integer called the principal quantum number. The energy of the state is given by Bohr's formula with the same integer value of n. Several different states may have the same energy, in which case they are said to be degenerate. Shells are sets of degenerate states. The number of distinct orbitals in a shell is its degeneracy or its multiplicity. For example, there are four orbitals with $n = 2$ at the energy E_2 in the Bohr formula, forming a shell. One of the orbitals in this shell, like the ground state, has a spherically symmetric wave function. Spherically symmetric orbitals are called s states; the nomenclature for orbitals of different symmetry is provided in Table 2. The excited s state differs from the ground state in that it is smaller at the centre, and it goes through zero at some intermediate distance from the nucleus. The excited s state has another peak beyond this point, and the overall probability extends farther out than in the ground state. The other $n = 2$ orbitals form a threefold shell that is designated the p shell. The polynomials in their wave functions are simply the monomials x, y, or z. The orbital having the monomial x has a vanishing probability on the plane with $x = 0$, which is therefore a nodal plane. Similarly, the other p states have nodal planes oriented in other directions but all going through the centre. The probability distributions of the electron for these orbitals are sketched in Figure 5. One could imagine a state with a nodal plane in some other direction as well. Such a state, however, could be made by adding the wave functions of the three original p orbitals, with some constant multiplying factors. According to the rules of quantum mechanics, this state would not be distinct from the original ones that produced it. The atom would have some probability of being in one of the three original states, depending on the multiplying factors.

Degenerate states

Table 2: Spectroscopic Designation of Orbitals

angular momentum quantum number l	degeneracy	designation
0	1	s
1	3	p
2	5	d
3	7	f

States are distinguished from each other by numerical labels called quantum numbers. The principal quantum number n has already been mentioned. Three additional quantum numbers are associated with the electron wave functions of the hydrogen atom. Two of these depend on the polynomial function of x, y, and z in the wave function. Conventionally, these two quantum numbers are labeled l, representing orbital angular momentum, and m, representing the orientation of the angular momentum with respect to some axis. The states of different l are often designated by letters, as given in Table 2. There are several general rules for determining the values of l and m allowed in a wave function. The orbital angular momentum takes on integer values starting from $l = 0$. Spherically symmetric wave functions have $l = 0$. Positive values of l apply to states whose wave function varies with angle. The more lobes there are in the angular pattern of the wave function, the higher is the value of l. The m quantum number, also called the magnetic quantum number, is restricted to a range of integers depending on l. The m quantum number ranges from $-l$ and increases in integer steps to $+l$. Distinct states exist for all values of m in this range. Thus, wave functions with a given orbital angular momentum l always have a degeneracy equal to $2l + 1$. The quantum number l and the associated degeneracy $2l + 1$ of the m states is a general feature of

Quantum numbers

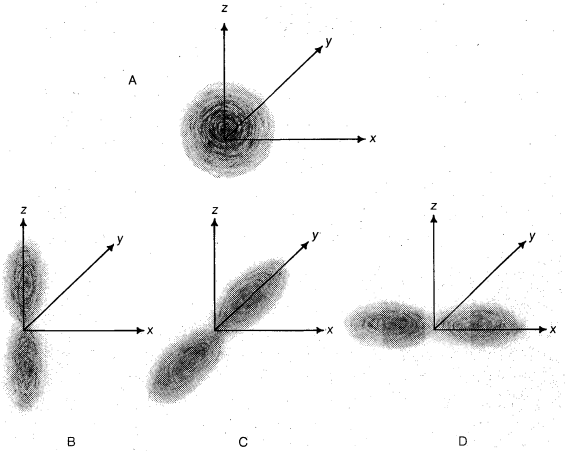


Figure 5: *Electron densities in wave functions of the Schrödinger equation.*
(A) The lowest s orbital, recognizable by its spherical symmetry and the absence of any nodes. (B,C,D) The three p orbitals.

wave functions obtained in physical systems with a force directed toward a centre. For example, the ground state of the hydrogen atom is spherically symmetric with $l = 0$ and only one distinct orientation. The p-wave states in the $n = 2$ shell have orbital angular momentum $l = 1$. By the rules of the orientational degeneracy, there are three distinct p states, labeled $m = -1$, 0, and $+1$. In higher shells one finds $l = 2$ states, which have two angular nodal surfaces and a degeneracy of 5.

The relationship between the principal quantum number n and the orbital angular momentum l has not been discussed up to this point. For each value of n, states of the hydrogen atom exist with l ranging from 0 to $n - 1$. The fact that all l states for a given n are degenerate is a special circumstance of the Schrödinger equation for the hydrogen atom. Each of the l's has a multiplet of m states, described in the previous paragraph.

The Schrödinger equation and Bohr's formula are quite accurate, but small deviations in energy can be observed in the hydrogen spectrum, and the actual multiplicity of states is twice that predicted by the Schrödinger equation. These flaws can be corrected by introducing the electron spin, with either the Schrödinger equation or the more precise Dirac equation. The electron spin behaves exactly like an angular momentum, but with a value ½. There are two m states, with values $-½$ and $+½$, which provide the needed doubling in the multiplicity of states. The interaction of the spin changes the energies of the states, so they are no longer independent of l. The coupling of the spin to the orbital angular momentum is treated by defining a new angular momentum quantum number j. For a given l, j can have the value $l + ½$ or $l - ½$, depending on whether the orientations of the spin and orbital angular momentum are parallel or antiparallel. However, in an $l = 0$ state such as the ground state, only $j = ½$ is allowed. This method of combining spin and orbital angular momentum is known as spin-orbit coupling. In the Dirac equation the $n = 2$ states are shifted in energy by the spin-orbit coupling, with the $j = ½$ states slightly lower than the $j = 3/2$ state. The predicted splitting agrees very well with the observed fine structure in the spectrum.

Spin-orbit coupling

Further refinements to the theory of the hydrogen atom have been made with quantum field theory. The theory predicts a very small energy difference between the two multiplets with $j = ½$. This tiny splitting has been observed, and its magnitude agrees with theory.

One final ingredient completes the description of the hydrogen atom—the interaction of its electron with the magnetic moment of its nucleus. The proton nucleus of ordinary hydrogen has a spin of ½ and a magnetic moment. The energies of the atom depend on the relative orientation of the electron and the proton's spin. This can be characterized by a new quantum number, the total spin F. It can have a value 0 or 1, depending on whether the two spins are parallel or antiparallel. The energy difference between these two states is very small, so that transitions are in the microwave region of the electromagnetic spectrum. The frequency of the radiation is 1.4×10^9 hertz, which is a well-known feature of the radiation from interstellar hydrogen gas observed in radio astronomy.

Multielectron atoms. The properties of multielectron atoms are governed to a large extent by simple principles that determine the electron wave function. Electrons are placed in orbitals similar to the wave functions of hydrogen, starting with the lowest energy $n = 1$ orbital. The Pauli principle requires that each electron go into a different orbital. (It should be remembered that the two spin orientations in a wave function count as different orbitals.) With this procedure, each element acquires a unique structure of electron orbitals that gives it its characteristic atomic and chemical properties. A simple example is the helium atom. Helium has two electrons, both of which go into the $n = 1$ orbitals according to the spin degeneracy rule. All $n = 1$ electrons are tightly bound, and so the helium atom is not readily ionized; neither does it form chemical bonds by sharing its electrons with other atoms. The next element, lithium, has three electrons; only two can go in the $n = 1$ orbitals, so the third must be placed in an $n = 2$ orbital. This shell is much higher in energy; thus, the last electron is loosely bound. The ionization energy, which is the energy required to remove an electron from an atom, shows clearly the effects of the electron shells. Figure 6, displaying the ionization energies of the elements, shows that helium is the highest of any element. There is a big jump from helium to lithium, which needs only 5 eV of energy to ionize.

Ionization energy

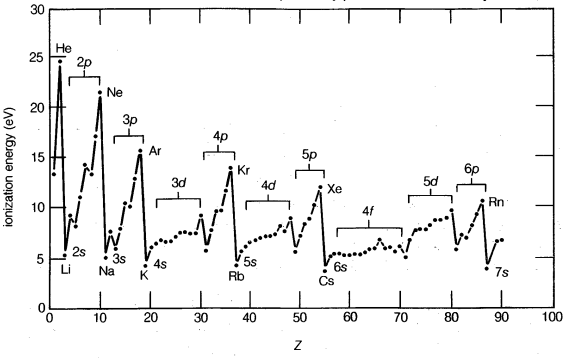


Figure 6: *Ionization energies of the elements.* The shell of the least bound electron is designated by the principal quantum number n, followed by the letter for the angular momentum.

The shell behaviour of complex atoms is derived from the approximate solution of the Schrödinger equation, since exact solutions to the multielectron equation do not exist. The Hartree method (see above), treating each electron with its own wave function, gives an adequate approximation for most purposes. In Hartree's theoretical scheme, the Pauli principle is imposed by requiring all wave functions of the electrons to be distinct. An important part of the Hartree theory is the behaviour of the self-consistent force field governing the electrons. At small distances from the nucleus, the Hartree field is virtually the same as the field of the nucleus. The contribution from the electrons is small because they are spread out in all directions. At large distances the nucleus appears to be surrounded by a cloud of electrons screening its positive charge with their negative charges. The last electron, if it is far away, "sees" a field resembling that of a single positive charge. The electron wave functions in this potential are qualitatively very similar to the hydrogen wave functions, retaining strong shell effects. The range of the wave functions and their energies, however, depend on both the atomic number and the presence of electrons in other shells. The $n = 1$ electrons are closest to the nucleus and are exposed to nearly its full electric field. The highest shell electrons are farthest away on average and feel a field similar to that of a hydrogen atom.

The size of atoms does not vary much from element to element, growing only slightly from light to heavy elements. This is due to a compensation between two opposing tendencies. The outer electrons of heavy atoms are in higher shells, the probability distributions of which are farther from the nucleus. On the other hand, the size of each shell is dependent on nuclear charge; the higher charge in heavier nuclei pulls a given shell in closer.

Atomic spectroscopy and lasers. The energy levels of atoms are studied by measuring the wavelengths of light emitted by the atoms in excited states. Atoms become excited when they absorb energy. In a gas, excited atoms are produced with heat or electric currents. The light emitted is analyzed in a spectrograph, which disperses the light in different directions according to its wavelength or frequency. Transitions from higher to lower energy levels appear as bright lines in the spectrum (Figure 7). This technique is known as emission spectroscopy. If white light is passed through a gas of the element being studied, light will be absorbed for those frequencies that allow a transition of the atom from the ground state to an excited state. The absorption creates a dark line in the spectrum of the white light. This is called absorption spectroscopy. The characteristics of the lines can be measured with great accuracy using spectrographs together with other optical

Emission and absorption spectroscopy

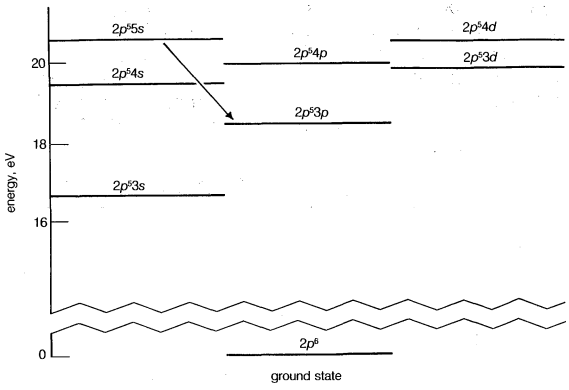


Figure 7: *Energy levels of the neon atom.*
The orbitals occupied by the six outer electrons are indicated on the levels. The transition responsible for the red light from common lasers (wavelength of 6330 Å) is shown by the arrow.

instruments called interferometers. Until lasers were developed, however, it was impossible to achieve the resolution desired because the frequency of the light emitted is affected by the motion of the atoms and collisions between them. The frequency shift from atomic motion, known as the Doppler effect, limits the resolution to about 1 part in 10,000. However, with intense light at a fixed frequency available from lasers, one can preferentially excite atoms moving at some particular velocity. Modern instrumentation makes it possible to observe structure in the lines to 1 part in 10,000,000. This accuracy allows the effects of nuclear spin, size, and shape to be measured directly with atomic lines.

The principle of the laser is based on the transitions between energy levels in an atom, or molecule. Normally, more atoms are in lower states than in upper states. Energy would be absorbed from a beam of radiation because most of the transitions would go from a lower level to an upper one. If the number of atoms in the upper and lower states is equal, there will be the same number of transitions in both directions. Therefore, the intensity of the atom's light beam would not change going through the medium. Conditions for the laser are obtained if more atoms are in the upper state than in the lower state; this is called a population inversion. As a result, a laser has more transitions going downward than upward, and a beam going through the medium is amplified. Passing the amplified beam back and forth between mirrors produces the very high intensity laser light.

Characteristic X rays. The inner shells of an atom are responsible for the characteristic X rays of an element. These are X rays of definite wavelengths emitted when atoms are bombarded with energetic electrons or other forms of radiation. The radiation knocks an inner electron out of its orbital, leaving a vacancy. An electron from a higher shell then jumps into the empty orbital, emitting a photon in the process. Because of the strong binding of inner shells in heavy atoms, such photons are produced with energies ranging from 100 to 100,000 eV in the X-ray region of the electromagnetic spectrum. The approximate energies of these transitions are given by the Bohr formula with a modified value of the nuclear charge. More accurate values for the energies can be calculated on the basis of the Hartree theory.

The ionization energy of atoms depends on the energy of the electron orbitals in the highest occupied shells. The variation of these energies with atomic number displays a periodic behaviour, as seen in Figure 6. Particularly noticeable is the increase in ionization energy as the *p* shells of the atom are filled. This is precisely what one expects to find: as the charge of the nucleus increases, the binding energy of the orbitals in a given shell increases. When a shell is filled, the next electron must go into a higher, less bound shell, producing a break in the ionization-energy plot. The fact that the behaviour is periodic shows that the inner cores of atoms act on the outer electrons in much the same way for different atomic numbers. The nuclear charge is screened by the inner electrons so that the outer electrons are effectively governed by the field from a lower atomic number.

The electron affinity of atoms exhibits a similar periodicity. Electron affinity is the binding energy of the extra electron when an atom forms a negative ion. Such atoms as helium that have all their electrons in tightly bound filled shells do not attract additional electrons strongly enough to bind them. On the other hand, such atoms as fluorine that have a single vacancy in a tightly bound shell readily attract and capture an extra electron into that orbital. The systematics of the electron affinities are shown in Figure 8. Obviously, the behaviour of the loosely bound electrons and shell effects are crucial to an understanding of the chemical properties of elements. The basic period of eight in the periodic table is associated with the filling of the *s* and *p* orbitals of a given *n* shell (see Figure 9). For each spin orientation, there are one *s* orbital and three *p* orbitals, for a total degeneracy of eight. In the hydrogen atom, the *s* and *p* shells are degenerate, and in complex atoms they remain close in energy. Higher angular momentum shells, however, have considerably less binding than their *s* and *p* partners in the same *n* shell. As a result, they are filled out of sequence in the periodic table. For example, the 3*d* shell is filled after the 4*s* shell.

Electron affinity

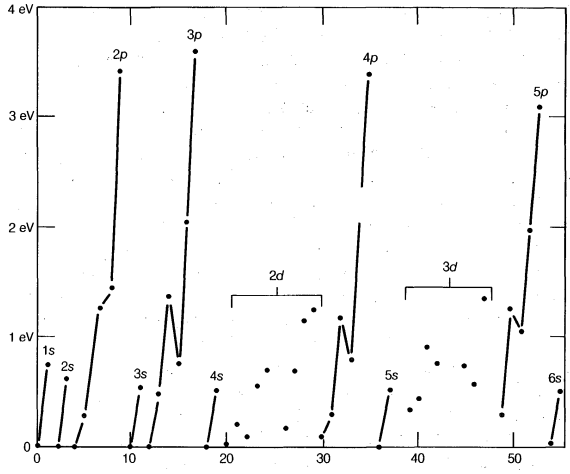


Figure 8: *Electron affinities of the elements.*
The shell of the most bound vacancy is indicated by its spectroscopic designation.

The diminished binding of the higher 1 shells can be understood simply with a classical picture of the electron's motion. The higher angular momentum shells correspond to orbits that are more nearly circular, whereas the *s*- and *p*-wave functions correspond to elliptical orbits in which the electron approaches the centre of the atom more closely. Because the centre has the strong attraction from the unscreened nucleus, these low 1 orbitals will have a greater binding. For a given energy, elongated orbits extend farther out than the circular orbits. This greater

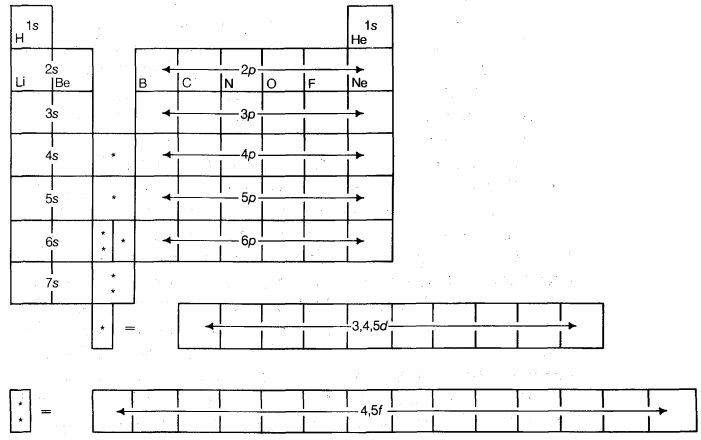


Figure 9: Periodic table of the elements showing the valence shells.

extension corresponds to the fact that in s and p orbits the electron probability is significant to larger distances. Most of chemistry has to do with these outer electrons. The $l = 3$ shell (f shell) is well shielded by other electrons, and the elements that differ only in the number of f-shell electrons have very similar chemical properties.

Exotic atoms. Atoms can be formed with other charged particles serving as the negatively charged electrons or the positively charged nucleus. Mu mesons, pi mesons, or antiprotons can be substituted for electrons, while a positron can replace the nucleus. These exotic atoms exhibit energy levels and transitions just like ordinary atoms. The transitions between high n orbits follow the Bohr formula, with an appropriate change in the mass m in that formula. The energy levels with low n are affected by the finite size of the particles and by other forces that may act besides the electric force. Physicists use these atoms to investigate the forces and the sizes of the particles involved.

Interatomic forces and chemical bonds. The forces between atoms are complex and varied. In close proximity, all atoms repel each other. At intermediate distances, the forces of chemical binding are prominent. Finally, at large distances, all atoms are attracted to each other by a much weaker force. These forces are all explained by the quantum theory of electrons in atoms. In the presence of other atoms, the electrons become rearranged. If the new arrangement has less energy, it is the preferred state of the atoms, and there is an attractive force to reach this state. Conversely, if the rearrangement increases the energy, the resulting force between the atoms is repulsive. The rearrangement of electrons may take place in different ways, and different names are given for the resulting chemical bonds (see below).

Repulsion between atoms

The universal repulsion between atoms at short distances is responsible for atoms taking up a definite amount of space and for the incompressibility of solid materials. This repulsion has two causes. First, as the atoms are pushed into each other, the shielding provided by the electrons around each nucleus is less effective and the electric force between the two nuclei repels them. A second repulsive influence arises from the Pauli principle. Not only is it forbidden to have two electrons with the same wave function but the wave functions must be different enough to construct an antisymmetric total wave function. When two atoms are close together, the electrons in the overlapping region spend part of their time in their original orbitals and part of their time in orbitals of higher energy in order to satisfy the Pauli principle.

In chemical bonds the rearrangement of electrons leaves atoms in a state of lower energy, stabilizing them at fixed distances from each other in definite geometric configurations. The amount of energy associated with chemical bonds ranges from 1 to 10 eV. A number of different kinds of bonds are recognized in chemistry. If the rearrangement causes a net shift of the electron's probability from one atom to the other, the bond is called heteropolar. An extreme example of this would be an ionic bond, in which an electron has moved from an orbital in one atom to that in the other. This electron transfer gives one atom a net positive charge and the other a net negative charge. The two resulting ions are then held together by the electric force of attraction between unlike charges. An example of ionic bonding occurs in alkali halide crystals such as sodium chloride (common table salt). The alkali metals are easily ionized, and the halogens have a strong affinity for additional electrons, as seen in Figures 6 and 8. The original concept of chemical binding was based on the attraction of opposite charges.

Strong chemical bonding of the homopolar, or covalent, type also occurs. In this case there is no net transfer of electrons from one atom to another. The rearrangement of the electron distribution is more subtle. In a covalent bond between two atoms, the electron density shifts from the outer surfaces of the atoms to the region between the two atomic centres. In general this does not happen, because the Pauli principle affects the wave functions by spreading the electrons apart. The Pauli principle, however, may be satisfied for a pair of electrons by giving them opposite spin orientations, in which case the spatial wave functions are combined to bring the two electrons closer together. A covalent bond can be formed between two atoms if there is an unpaired electron on each atom. In the bonded state the electrons will have their spins oppositely aligned. An atom can form a number of covalent bonds equal to the number of its unpaired electrons, which in turn depends on the shell degeneracy. For example, the nitrogen atom has five electrons distributed over four orbitals in the $n = 2$ s and p shells. Two electrons must occupy the same orbital as a pair, but the remaining three electrons can go unpaired in separate orbitals, giving nitrogen a valence of 3.

The geometry of chemical bonding, with definite angles between the bonds in triatomic and more complicated molecules, depends on the geometry of orbitals with lobes extending at various angles. If atoms bonded with pure p orbitals, the bonds would extend out at 90° angles from each other like the wave functions depicted in Figure 5. Actual chemical bonds, however, are made primarily by combining s and p orbitals together in the wave functions. This so-called hybridization produces wave functions with different geometries. In the carbon atom, for example, the hybridization yields four bonds directed toward the corners of a tetrahedron.

Hybridization

The ionic and covalent bonds are two extremes in a range of bonding behaviour. The bonding wave functions do not need to be localized on two neighbouring atoms. In so-called delocalized bonding the electron wave functions extend over several atoms.

All of these properties can be predicted quite well using the Hartree theory. The positions of the atomic centres are first roughly estimated. The Schrödinger equation is then solved for each electron, with the requirement that the total forces from all the atoms be consistent with the total charge density from the electrons and nuclei. After the wave functions of all the electrons are found, the energy of the molecule can be calculated. This is repeated for other positions of the atomic centres until the minimum energy is found. The resulting arrangement of atomic centres and electron wave functions provides the configuration of the stable molecule. Depending on the behaviour of the Hartree wave functions, one can find various kinds of chemical bonds. The Hartree theory is quite reliable in predicting the geometry of molecules and the character of the bonds, but it is less successful in predicting accurate bond energies.

Some forces between atoms at larger distances are too weak to form chemical bonds. These forces have two origins. One long-range force is associated with the electric fields that extend outside of any molecule with heteropolar electron distributions. An example is the water molecule, which has a net shift of electron charges from the hydrogen atoms to those of oxygen. The electric field attracts both positively and negatively charged ions, depending on the orientation of the water molecule. This explains why liquid water is such a good solvent for polar and ionicly bonded molecules. Another long-range force, called the van der Waals force, is a weak attraction between all atoms. It provides the cohesive force in nonpolar liquids such as liquid air or gasoline. These liquids have a low boiling point because the bond energies of the van der Waals force are very low (only a few tenths of an electron volt). The force arises from a subtle effect in quantum mechanics—namely, the existence of fluctuating electric fields outside an atom even though the electron wave function surrounds the nucleus and neutralizes its charge. The fluctuating fields are associated with the possible positions of the electron if it is frozen at some particular spot in its orbital.

Van der Waals force

Observing atoms. Individual atoms are far too small to be seen in any light microscope. The inherent limitation on such a microscope is the wavelength of light, which determines the minimum size of an image. In comparison, the distance between atoms in a solid is less than 1,000th of the wavelength of light. Therefore, the minimum size of any image would contain too many atoms for the eye to differentiate. There are, however, several direct methods for observing individual atoms and their arrangement in molecules and solids. In a transmission

Direct observation methods

electron microscope, electrons are transmitted through the material to be studied. Images are formed because the electrons are absorbed differently by different atoms. With this technique, one can just barely see the individual atoms of the heaviest elements when they are separated by several atomic lengths (see Figure 10). Another technique involves the use of a field-ion microscope. In this case, a wire with a very fine tip is placed in a near-vacuum and given a positive electric charge. Individual atoms on the surface of the tip ionize residual gas molecules. The ions are then accelerated by the charge in straight lines away from the tip. They form spots on photographic plates at different positions corresponding to the location of atoms on the tip (see Figure 11). A third imaging technique is called scanning tunneling electron microscopy. Here, a wire with an extremely fine point is positioned within 10^{-9} metres of the surface to be studied. Electrons jump from the surface to the wire through individual atoms on the surface. The probe wire is moved back and forth over the surface in a raster pattern, and variations in the electron current indicate the positions of the atoms.

Albert V. Crewe, Enrico Fermi Institute, University of Chicago

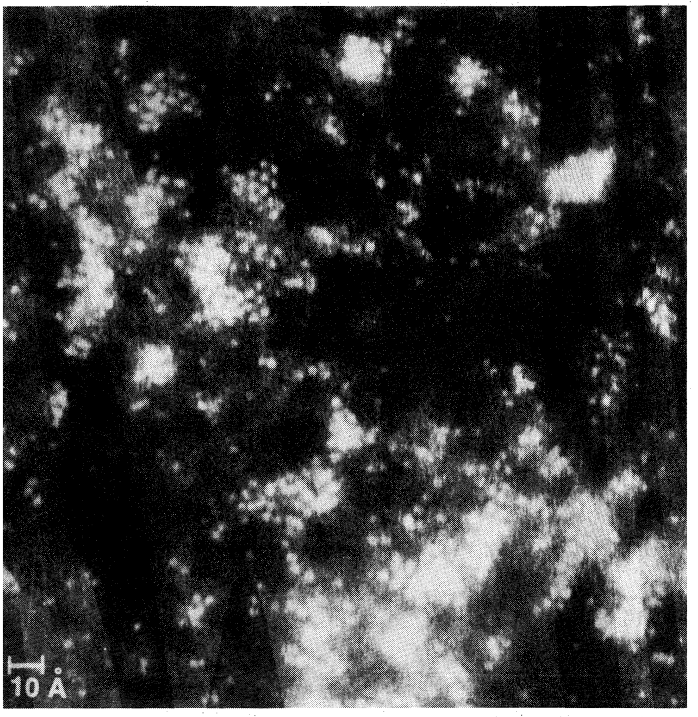


Figure 10: Transmission electron micrograph of single uranium atoms and microcrystals obtained from a solution of uranyl acetate.

Diffraction techniques

The positions of atoms in solids and molecules are commonly determined by wave diffraction. Diffraction refers to the highly directional patterns formed by waves when they travel through regularly ordered mediums. Diffraction techniques use a crystal or a large number of atoms in a regular arrangement. The pattern of the waves reflected from the crystal depends on the relative positions of the atoms in the crystal. The information contained in the wave pattern is far from complete, however, so there may be ambiguities in reconstructing the atomic arrangement. The most common technique of this kind is X-ray diffraction. X rays are similar to visible light, but they have much shorter wavelengths. Radiation of this type that can be used for diffraction has a wavelength of 10^{-10} metre or smaller, which is even smaller than atoms. As the X-ray waves are most affected by the electrons in atoms, this technique locates atoms primarily by their electron cloud. Another diffraction technique uses neutrons from a nuclear reactor. Neutrons, like all particles, have wave properties and can be diffracted from a crystal. Neutrons of a particular wavelength are beamed on a crystal, and diffraction sends them off in various directions. The geometric arrangement of the atoms being studied is inferred from the diffraction pattern. Neutron waves interact most strongly with atomic nuclei, and so this technique complements X-ray diffraction by mapping the nuclear centres of atoms.

Bulk matter. The behaviour of bulk matter depends on the relative magnitude of the binding forces between atoms or molecules and the thermal energy of motion. The thermal energy of motion at room temperature is about 0.025 eV. If the binding energy is less than about 10 times the thermal energy, the substance will be in a gaseous state under normal conditions. This typically applies to small molecules such as air molecules that interact only by the weak van der Waals force. With larger ratios of binding energies to thermal energies, the substance condenses into a liquid or a solid. Solids are classified into several types, depending on the kind of bonding between their atoms. Four main types of ordered solids are ionic crystals, molecular crystals, covalent crystals, and metals. Ionic crystals are held together entirely by the electrostatic forces between ions. Common table salt is an example. A molecular crystal is composed of distinct molecules held together by weaker forces such as the van der Waals force. In covalent crystals the chemical bonds extend from atom to atom over the entire crystal. Diamond, a form of carbon, is an example of a covalently bonded solid; as in simple organic compounds, each carbon atom has four bonds pointed to the corners of a tetrahedron. In metals the concept of chemical valence breaks down. The atoms are closely packed together, and the number of neighbours for a given atom is determined more by the geometry of touching spheres than by the number of valence electrons.

Classification of solids on the basis of bonding

The Hartree theory provides a common framework for describing the varying properties of these different kinds of solids. The electron orbitals are allowed to spread out over the entire solid. Effectively, one deals with an infinite number of electrons and orbitals, which can be handled with mathematical techniques developed by the Swiss-born U.S. physicist Felix Bloch and others. The orbitals display shell-like features even in a very large system. There are groups of orbitals with energies close together. Known as bands, they are separated by gaps in energy. The characteristics of any such solid depend very much on the extent to which the orbitals are filled in the highest band. This filling level is called the Fermi energy. If

Band structure of solids

By courtesy of the Department of Physics and Astronomy, Michigan State University

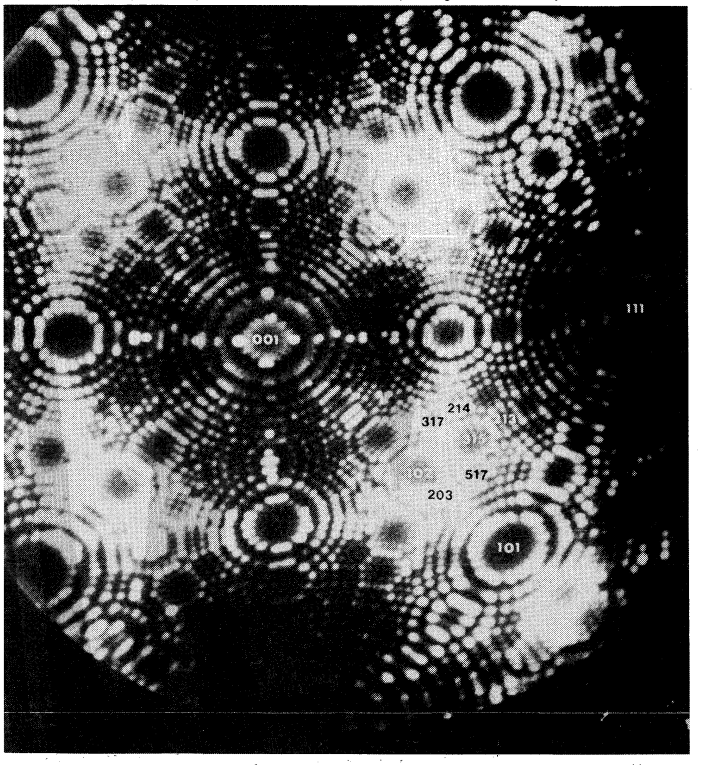


Figure 11: Field-ion micrograph of atoms on the surface of a platinum needle.

the orbitals are completely filled up to the band gap, the states of the electrons are essentially fixed and the material will be an insulator. On the other hand, if the band is only partially filled, electrons can jump from one state to another with only an infinitesimal change in energy. In this case, the electrons can move about, and the material is a conductor. The sharp distinction between insulators and conductors becomes blurred in materials called semiconductors, which have small energy gaps. In silicon, for instance, the energy gap is only 1.4 eV, and it is possible to make silicon conduct by exciting electrons to the higher band. In some materials there is no energy gap between the occupied and empty orbitals, but the number of orbitals at the Fermi energy is so small that conductivity is poor. Graphite, another form of carbon, has a band structure of this type.

THE NUCLEUS

Nucleons. The protons and neutrons that make up the atomic nucleus are the lightest members of the baryon family of subatomic particles. The proton has a mass of 1.673×10^{-27} kilogram, which is 1,836 times larger than the mass of the electron. The neutron has approximately the same mass as the proton—1.675×10^{-27} kilogram. Subatomic masses are commonly expressed in units of rest energy, which is related to mass by Einstein's formula $E = mc^2$. The proton's rest energy is 938.3 MeV and that of the neutron is 939.6 MeV. All baryons except the proton are unstable and can decay into lighter baryons. The neutron, having more rest energy than the proton, can decay into a proton with the emission of an electron and an anti-neutrino. Its lifetime outside the nucleus is 10 minutes. In the nucleus the binding energy is larger than the difference in rest energies between the neutron and the proton. The binding forces stabilize the nucleus against the decay of its neutrons.

Properties of protons and neutrons

Other properties of nucleons are their charge, spin, and magnetic moment. The charge on a proton is equal and opposite to the electron's charge. The neutron is uncharged. Like electrons and other elementary fermions, nucleons have a spin quantum number of $1/2$ and are described by a wave function that is doubled in multiplicity by two spin orientations. Nucleons and electrons also have a magnetic moment associated with their spin. However, the numerical value of a nucleon's magnetic moment, unlike an electron's, is quite different from that predicted by the Dirac equation; in fact, the proton's moment is 2.8 times the value predicted. Furthermore, the neutron has a somewhat smaller moment opposite to that of the proton, though according to the Dirac equation it should have none at all.

While electrons are point particles, nucleons have a finite extension. The distribution of the positive charge in a proton can be measured in electron scattering experiments. The charge is distributed in a smooth cloud extending to a distance of about 1×10^{-15} metre, or 1 femtometre, from the centre. The neutron is neutral overall, but electron scattering measurements show that it also has an internal structure with a slight positive charge in the core surrounded by a shell of negative charge.

Internuclear forces. The most important characteristic that distinguishes nucleons from electrons and other leptons has to do with the forces acting on them. Nucleons are subject to the three forces of nature affecting leptons: gravitation, electromagnetism, and the weak force. However, nucleons also exert a very powerful force on each other—namely, the nuclear, or strong, force. This force has an extremely short range so that its effects are not felt outside an atom. Indeed, even within an atom, the strong force is insignificant just a few fermis (one fermi = 10^{-15} metre) from the surface of the nucleus. By contrast, the effects of the electromagnetic and gravitational forces can be felt at all distances.

Strong force

The character of the strong force is rather complicated and only partially understood. At very small distances between nucleons, less than one femtometre, the force is repulsive and tends to keep the nucleons apart. At intermediate distances, about one to two femtometres, the force is strongly attractive and can bind the nucleons by energy amounting to many millions of electron volts. Beyond a few femtometres, the strong force becomes inconsequential in comparison to the electromagnetic force. A graph of the behaviour of the strong force is shown in Figure 12.

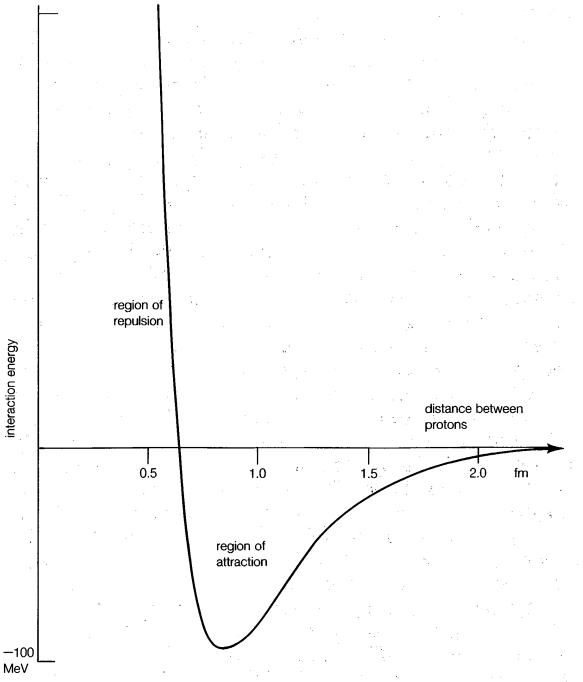


Figure 12: *Proton force.*
The interaction energy of two protons is shown as a function of the distance between them. At small separations, the energy is positive and the force is repulsive. At intermediate distances, the nuclear interaction is negative and the force is attractive. At distances larger than shown on the graph, the nuclear force becomes negligible and the repulsive electric force between the two charges is all that is left. The interaction depicted is for protons with antiparallel spins.

The effects of the strong force are seen in the interactions of free nucleons and the binding energy of nuclei. Nucleon interactions can be studied by bombarding hydrogen with a beam of protons or neutrons. The beam particles passing within a certain distance of a proton from the hydrogen target will be deflected by the strong force. The strength of the interaction between beam and target nucleons is characterized by a quantity called the cross section, which has units of area. It is the apparent area of the target particle that intercepts the beam particle and results in an interaction or a deflection. The cross section of high-energy nucleons scattering or otherwise interacting with each other is about four square femtometres. If the particles are thought of as spheres that interact when they come into contact, the radius of the individual nucleon sphere would be about 0.7 femtometre. At low energies the cross sections become larger, partly because there is more time for the forces to act with slower moving particles. Quantum wave effects also become more important at low energy; they allow the influence of the force to be felt at larger distances and help increase the cross section.

The cross section

Most of the existing detailed knowledge about nuclear forces comes from matching observed cross sections, including their dependence on energy and scattering angle, with quantum mechanical calculations based on models of the strong force. Yukawa's meson exchange picture (see above) explains the major features of the strong force beyond one femtometre. In quantum field theory, two particles may exert a force on each other at a distance only by exchanging a third particle. The particles exchanged in the strong force are mesons. The pion, the lightest of the mesons, is responsible for the longest range portion of the strong force. When the pion is exchanged, it can transfer a charge from one nucleon to another. In this process, called an exchange interaction, a proton is converted to a neutron and a neutron becomes a proton. At shorter distances, heavier mesons become important, and several pions may be exchanged at the same time. The main

Exchange of mesons

attraction between nucleons in bound nuclei is caused by the exchange of multiple pions and higher mesons. The exchange of mesons also can change the internal structure of the nucleon, transforming it into some other baryon. Even when there is not enough energy to produce the excited baryon, its effects are felt in the strong force.

The strong force also depends on the spin orientation of the nucleons. This dependence adds to the complexity of the force. For example, the force between protons and neutrons is most attractive when the spins are pointing in the same direction. On the other hand, the Pauli principle does not allow two protons with parallel spins to be located close to each other, and the interaction with the opposite spin alignment is weaker. Part of the spin dependence is known as the tensor force. This force reorients the spin and exerts a torque on the nucleons. The tensor force helps explain a property of the deuteron, which is a neutron and proton bound together. Under the influence of ordinary forces, the wave function would be spherically symmetric. Due to the tensor force, however, the deuteron is elongated in shape. The elongation is measured as its quadrupole moment. Another significant part of the spin-dependent force is the spin-orbit force. This force acts when the nucleons are moving around each other and have orbital angular momentum. It tends to align the orbital angular momentum with the spin orientation.

Tensor and spin-orbit forces

An important consequence of the strong force is the binding of nuclei. The simplest example is the deuteron, with a binding energy of 2.2 MeV. The binding energy is the sum of an interaction energy and a kinetic energy of the particles. By itself, the interaction energy is much more than 2 MeV, but a large kinetic energy nearly cancels it. In heavier nuclei the interaction energy is higher because there are more neighbouring nucleons; consequently, the binding is greater. Protons also are attracted to each other by the strong force, but their interaction is not sufficient to bind them together. Thus, there is no isotope of helium consisting only of two protons.

Shape and size of nuclei. Most nuclei are spherical, and their size is determined by the number of nucleons present. The density of nucleons is roughly constant in the interior of a nucleus, and so the volume of the sphere is proportional to the atomic mass number. In this respect, the nucleus resembles a liquid drop, the volume of which is proportional to the amount of matter it contains. The density of nucleons in the interior of a nucleus is about 3×10^{17} kilograms per cubic metre; each nucleon takes up about six cubic femtometres of volume. The radii of nuclei found in nature range from two to eight femtometres.

The constant interior density of nuclei, called the saturation property, is quite different from the distribution of electrons in atoms. The density of electrons varies considerably with their distance from the centre of the atom; electrons are highly concentrated in the inner shells and more dispersed in the outer ones. The saturation of nuclear matter, on the other hand, arises from a delicate balance between the attractive and repulsive components of the nuclear force. Aspects of the interaction contributing to the saturation are: (1) the short-range repulsion that prevents the nucleons from occupying the same volume; (2) the charged pion exchange that is less effective at high density; (3) the tensor force; and (4) the interactions that make internal excitations of the nucleons.

Saturation property

The most precise information on nuclear sizes comes from electron-scattering measurements. Because of their quantum-wave properties, electrons scattering from nuclei exhibit diffraction, which is preferential scattering at specific angles. Detailed pictures of the size and shape of the nucleus can be inferred from these diffraction patterns. Figure 13 shows an example of inferred charge density on a slice through the centre of a nucleus of lead-208. The charge is roughly uniform inside a sphere and has a diffuse edge on the surface. The density falls from its inside value to zero over a distance of roughly two femtometres. One does not see the individual particles in the nucleus because their wave functions are diffuse clouds. There are, however, perceptible variations in density in the interior. These oscillations show that the wave functions of the nucleons inside a nucleus have a shell behaviour just like

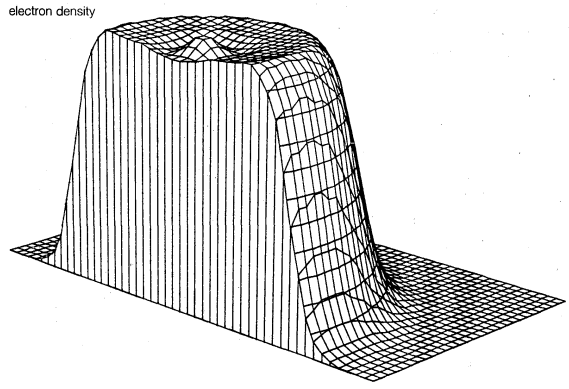


Figure 13: Profile of the charge distribution within the nucleus of lead-208 inferred from electron-scattering experiments.

the electrons in atoms. The peaks occur at positions where the proton wave functions are maximal, and the valleys occur where the wave functions pass through zero.

Mass and binding energy. The mass of a nucleus is typically about 1 percent less than the total mass of the nucleons composing it. This mass deficiency is related to the binding energy of nuclei by Einstein's formula $E = mc^2$. The binding energy determines which nuclei are stable and how much energy is released in a nuclear reaction. The picture of the nucleus as a charged liquid drop describes the overall trends of the binding energies. Each particle in the interior of a classical liquid is bound by the same amount. Very roughly, nuclear binding energies behave accordingly, with a typical binding energy of 8 MeV per nucleon. In detail, the trend of nuclear binding energies may be seen in Figure 14, which graphs the binding energy per nucleon as a function of the atomic mass number. The graph shows the near constancy of the binding energy from nuclei as light as helium-4 up to very heavy nuclei. By contrast, the binding energy of atoms varies greatly with atomic number because their inner shells are so tightly bound.

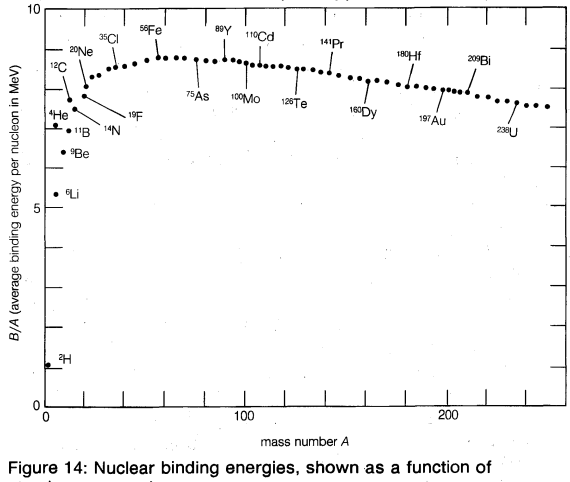


Figure 14: Nuclear binding energies, shown as a function of atomic mass number.

Particles at the surface of a liquid have fewer neighbours to interact with, so there is surface tension and a deficiency of energy from the surface. In the same way, lighter nuclei have a relatively large surface-to-volume ratio and are not as strongly bound. This may be seen in Figure 14 in the behaviour of the curve for smaller atomic mass numbers.

Repulsion between the charges of protons reduces the binding energy of a nucleus. The negative contribution from this repulsion is called the Coulomb energy. The effect is largest for the highest atomic numbers, which are the most highly charged nuclei. The surface energy favouring large nuclei and the Coulomb energy favouring low atomic numbers oppose each other; consequently, the optimum mass number for a nucleus is in the middle

Coulomb energy

region, around mass 60 and the iron nucleus. Nuclear energy can be released by any process that transforms nuclei to others closer to the middle masses. Thus, nuclear energy can be released by fusing light nuclei or by splitting heavy nuclei in the fission process. Energy production in stars results primarily from the fusion of hydrogen to make helium. Hotter stars can fuse heavier elements, but all nuclear energy is exhausted when iron is attained as a fusion product. Terrestrial elements are believed to have been formed by this process of stellar nucleosynthesis. Elements heavier than iron may have been formed by very transient processes in supernova explosions. (For a more detailed discussion of stellar nucleosynthesis, see STARS AND STAR CLUSTERS.)

Nuclear binding energy also depends on the relative numbers of neutrons and protons in a nucleus. In light nuclei the most binding is obtained with an equal number of neutrons and protons, which optimizes the neutron-proton attraction. In heavier nuclei the Coulomb repulsion between protons gives more binding to nuclei with an excess of neutrons. The most stable heavy nuclei have 50 percent more neutrons than protons.

Nuclear spin and magnetic moment. Nuclei, like atoms or other particles, may have an internal angular momentum called spin. Because of the characteristics of the forces involved, the spin of a nucleus is zero if the nucleus has an even number of neutrons and protons. Nuclei with an odd number of nucleons always have a half-integer spin. In many cases the magnitude of the spin can be determined from the nuclear shell model (see below). A nucleus with a spin also has a magnetic moment aligned along the spin axis. The magnetic moment of a nucleus is very weak compared to that of an electron, but it can still produce an observable effect in atomic properties. The magnetic field from a nucleus disturbs the degeneracy of atomic levels and causes the atomic spectral lines to split. This splitting is called hyperfine structure, as distinguished from the fine-structure splitting caused by the magnetic moment of electrons.

Magnetic resonance

An important application of nuclear magnetism is a method of analysis called magnetic resonance. In this technique, a material is placed in a strong magnetic field that tends to align the nuclear spins. Partially aligned spins will precess about the magnetic field direction, the way a gyroscope precesses when it is subjected to an outside torque. With the application of an extra time-varying magnetic field, the nuclei can be induced to precess in phase with each other. The resulting precessional motion of the spins causes the magnetization of the material to vary at a definite frequency, which is detected as a radio-frequency signal from the material. Magnetic resonance has several practical applications. Based on the fact that the precession rate is proportional to the field, it can be used to measure magnetic fields precisely. Notable is its use in a medical diagnostic procedure known as magnetic resonance imaging (MRI; see RADIATION: *Imaging techniques*). The nucleus most commonly employed in magnetic resonance is the proton in the hydrogen atom. Because the hydrogen nucleus has the largest magnetic moment for its spin, it provides the strongest resonance signals. Other nuclei used include those of phosphorus and fluorine.

Energy levels. Nuclei have energy levels just as atoms do. A typical excitation energy of a nuclear level is 1 MeV, which is 1,000,000 times larger than atomic energy levels. As in atomic transitions, a quantum of light may be emitted when the nucleus changes states. These quanta are the gamma rays that are often by-products of radioactivity.

Each energy level is characterized by its own values of spin, magnetic moment, and quadrupole moment. In addition, the quantum numbers of isotopic spin and parity are useful for describing nuclear levels. The isotopic spin quantum number characterizes the symmetry of the wave function when neutrons are interchanged with protons. Levels can exist in different nuclei with wave functions that are identical except for the interchange of neutrons and protons. Except for the Coulomb energy, these sets of wave functions have the same energies; the sets are called isospin multiplets (Figure 15). The parity quantum number can take on only two values, even or odd, depending on whether the wave function reverses sign when the coordinates of the particles are reversed. The parity and spin quantum numbers are useful in understanding the rates of transitions between energy levels; selection rules govern changes in spin and parity for the favoured transitions.

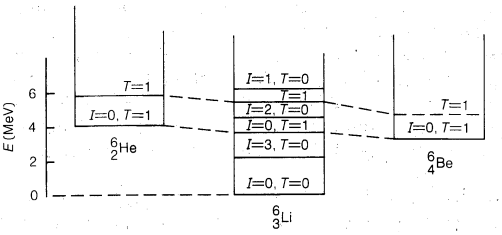


Figure 15: The energy levels of three nuclei with $A = 6$, relative to the ground state of ^{6}Li after subtraction of the calculated electrostatic energy. The isospin $T = 1$ levels exist in all three nuclei.

To excite nuclei out of their ground states, energy must be provided from some external source. A common way to do this is with nuclear reactions. A nucleus is bombarded with energetic particles of one sort or another. The particles interact, giving up part of their energy to induce an excited state of the nucleus. Under special circumstances it is possible to excite a nucleus with the gamma ray produced by another nucleus. In a nuclear process known as the Mössbauer effect, the nucleus of some isotope absorbs a gamma ray that has been emitted by another nucleus of the same isotope. Because the transitions take place between the same two levels, the gamma ray has exactly the right energy to be absorbed. The effect is difficult to produce because part of the transition energy normally goes to nuclear recoil during emission and absorption and is unavailable to the gamma ray.

Mössbauer effect

Among the many energy levels to which a nucleus can be excited, a few correspond to simple modes of motion for the nucleons. Deformed nuclei (see below) can be given energy in the form of rotational motion. Nuclei also can vibrate, and corresponding vibrational states may be found among the energy levels. Giant dipole resonance is an example of a nuclear vibration exhibited by all nuclei. In this mode of motion neutrons and protons in the nucleus oscillate back and forth, with the neutrons moving against the protons, as shown in Figure 16. Because there is a large oscillating electric field associated with this motion, gamma rays are absorbed and emitted readily. Another kind of vibration is called the quadrupole vibration. In this mode the nucleus oscillates in shape, changing between

From G.F. Bertsch and R.A. Broglia, "Giant Resonances in Hot Nuclei," *Physics Today* (August 1986)

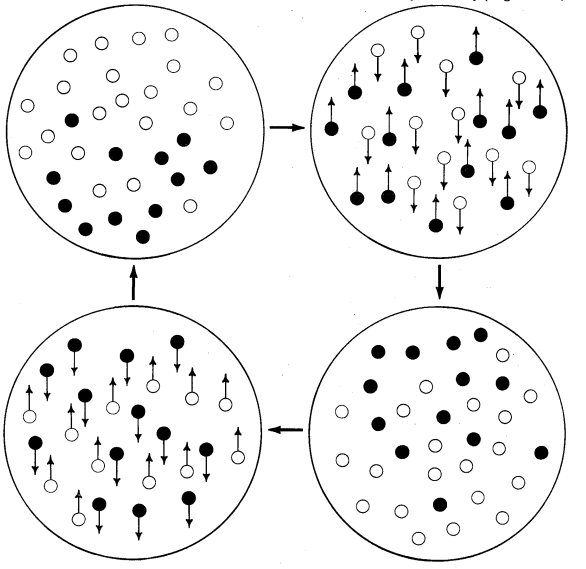


Figure 16: *Nuclear vibratory motion in the giant dipole resonance.*
Protons and neutrons, indicated by circles and solid dots, move in opposite directions.

prolate and oblate spheroids. This mode is easily excited by the inelastic scattering of particles from the nucleus. Its properties are deduced from the diffraction pattern observed in the scattering. In yet another kind of vibration, called the breathing mode, the nucleus expands and contracts. The frequency of the breathing mode depends on the compressibility of the nuclear medium and is used to establish its value.

Nuclear shell model. A comprehensive explanation of many nuclear properties is provided by a theory known as the shell model. It is based on the Hartree approximation to the Schrödinger equation. As noted earlier, this approximation assumes that all the particles have independent wave functions governed by a common force field. Of course, the forces in the nucleus are different from those in the atom as a whole. In the nuclear Hartree theory the forces are too strong to include directly, varying violently from repulsive to attractive. Instead, true interaction is replaced by smoother forces that are adjusted to reproduce the saturation property of nuclei.

The resulting force field pulls the nucleons in when they come to the surface, but it cancels out when nucleons are in the interior. The nucleons behave as though they were in a potential well with a flat bottom. The nuclear shells are the sets of wave functions that go with this flat-bottomed potential well. Some of the properties of these shells are similar to those of atomic shells, but there are also important differences. Provided the well is spherically symmetric, the shells have an angular momentum quantum number, as do the atomic shells. The lowest shell is the *s* shell, which can hold a neutron and proton of each spin orientation. Thus, the two protons and two neutrons of helium-4 fill the *s* shell and give this light nucleus its unusual stability. The energy levels of higher shells are shown in Figure 17. Differences from the atomic levels in hydrogen are readily apparent. First, there is no degeneracy between the *s* and *p* shells, so that the atomic periodicity of 8 does not occur in nuclear physics. Second, the spin-orbit force makes shells that have nucleon spin parallel to orbital angular momentum lower in energy than the shell with antiparallel spin. For example, in the *p* shell, there are two sub-shells, the $p_{3/2}$ and the $p_{1/2}$. The $p_{3/2}$, with orbital angular momentum 1 parallel to spin angular momentum 1/2, is lower in energy than the $p_{1/2}$, which has an antiparallel spin and orbital angular momentum. Combining orbital and spin angular momentum in this way is called *j-j* coupling. The number of protons or neutrons that can go in a *j* shell is its multiplicity equal to $2j + 1$. From this filling rule and the shell ordering in Figure 17, one can determine the shell structure of a nucleus for any given proton and neutron number.

Properties of nuclear shells

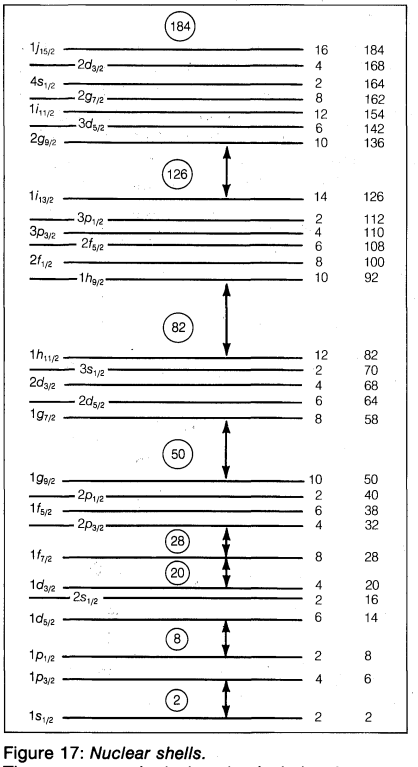


Figure 17: *Nuclear shells.*
The spectroscopic designation includes the quantum number *j*, which is a half integer. The degeneracy of the shells and the cumulative number of particles contained are listed on the right. Especially stable nuclei are found at the magic numbers, shown in circles.

An important consequence of the *j-j* coupling is the energy gap between shells that occurs at proton or neutron numbers 28, 50, 82, and 126. These so-called magic numbers are the nuclear equivalent of the atomic numbers of the noble (or inert) gases 2, 10, 18, 36, 54, 86. Nuclei with magic numbers of protons and neutrons are unusually stable and have especially large gaps between their ground and excited states.

Magic numbers

The ground state of nuclei with an odd number of neutrons or protons has a spin that is often given by a simple rule—namely, that the angular momentum of all the nucleons except the last are aligned to cancel. The last nucleon has the angular momentum of its shell, which gives the spin of the nucleus. For example, the nucleus of lithium-7 consists of three protons and four neutrons. According to the rule, the spin of lithium-7 is determined by the shell of the third proton. This is the $p_{3/2}$ shell, and indeed the spin of the nucleus is 3/2. Another property that can be inferred from the shell model is the parity of the nucleus. It can be determined from the shell model simply by combining the parities of the individual shell wave functions for nucleons.

Nuclear excited states are described in the shell model either by changing the orbitals that nucleons occupy in a shell or by moving nucleons from a lower shell to a higher one. Wave functions constructed in this way accurately describe the excited-state properties of light nuclei and nuclei near magic numbers. Outside of these limited regions of nuclei, the shell model allows an enormous number of states. The numerical complexity of dealing with so many states, however, prevents precise calculations from being carried out. A simplified treatment of the wave function explaining many properties of nuclear vibrations is based on a generalization of the Hartree method to allow the potential field to vary in time. The nucleons move in the time-varying field, and the frequency of motion can be calculated by requiring the field generated by the nucleons to be consistent with the field in which they move.

Deformed nuclei. Not all nuclei are spherical in shape. In certain regions of the periodic table, many nuclei are spheroidal. Typically, these nuclei have a prolate distortion, meaning that they have an elongated shape like a lemon. For example, the nuclei near mass number 160 are prolate spheroids, with the longer axis about 30 percent larger than the other two axes. The existence of these deformations was first detected in atomic spectra: the quadrupole moment distorts the electric field around the nucleus, producing characteristic features in the hyperfine structure of the atomic spectrum. The energy levels of these nuclei also exhibit characteristic features. A body capable of rotation and governed by the laws of quantum mechanics has energy levels with spacings that increase uniformly with higher angular momentum. Deformed nuclei exhibit this type of spectrum (see Figure 18).

Elongation of nuclei

Deformations can be explained by the Hartree theory of the nuclear wave function. The orbitals describing the individual nucleons are rarely spherically symmetric; many are, in fact, quite elongated. When a shell is filled, the various asymmetries in the different orbitals balance out so that the entire shell has a spherical mass distribution. If a shell is only partly filled, the overall shape deviates more or less from spherical symmetry, depending on which particular orbitals are occupied. In the case of an atom, the repulsion of the electrons tends to make the overall shape spherical: when one electron is in an orbital elongated in a certain direction, the energy is minimized if the other electrons are in orbitals elongated in other directions. The

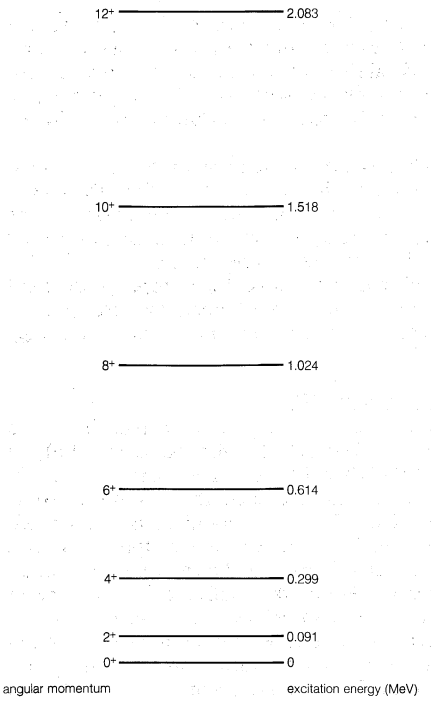


Figure 18: Energy-level spectrum of the deformed nucleus erbium-164.

overall density ends up being close to spherical, even when the individual orbitals are not. In a nucleus there is a net attraction between nucleons so that, if a proton's orbital is elongated in some particular direction, the neutron would tend to be in a similarly directed orbital. Many nucleons can act together and deform the entire nucleus. When the Hartree equations are solved for nuclei with partially filled shells, the resulting potential well is elongated and the nucleons occupy states elongated in the same direction.

Nuclear reactions. Nuclear reactions mediated by the strong force can occur when a nucleus is bombarded by some hadronic projectile, such as a proton, a pion, or another nucleus. For a reaction to occur, the projectile must come within a few femtometres of the nucleus, preventing positively charged particles from producing reactions at low energy. Since the nucleus is positively charged, a similarly charged projectile will not be able to come within range unless it has enough energy to overcome the force of the electric repulsion. For example, alpha particles will not come close enough to interact with an iron nucleus unless their energy exceeds about 10 MeV. For less highly charged nuclei, the numbers are smaller; a proton of 1 MeV can interact with a lithium nucleus. The first artificial nuclear reactions were induced by using a voltage source of 1,000,000 volts to accelerate protons to an energy of 1 MeV and bombard a lithium target. At high bombarding energies the electrostatic force is less important, and the cross section for producing a reaction is approximately equal to the area of a circle with the radius of the nucleus.

Neutrons can induce reactions at very low energies, even thermal energies, because they are not repelled by the positive charge of the nucleus. The cross section for reactions with thermal neutrons can be extremely large because the wave function of the neutron is spread out at low energy. Radioactive nuclei are commonly produced by bombarding nuclei with neutrons in nuclear reactors.

Hadronic reactions

Reactions induced by hadronic projectiles may be divided into two main categories: direct reactions and more complex reactions. In direct reactions the projectile retains most of its energy and momentum. A small amount of energy is transferred to the target, producing transitions to particular excited states. These direct reactions occur when the projectile passes near the target nucleus without penetrating it. The angular distribution of scattered projectile particles shows diffraction behaviour depending on the properties of the energy level excited. An example of a direct reaction is the following:

$$p + {}^{40}\mathrm{Ca} \longrightarrow d + {}^{39}\mathrm{Ca}.$$

Here, a proton, p, passes by the calcium-40 nucleus, picking up a neutron to form a deuteron, d. The deuteron comes off in approximately the same direction as the motion of the incident proton but with diffractional variations in intensity at different angles. From the diffraction pattern, the orbital angular momentum of the neutron in the target may be inferred. Measurements such as these were important in establishing the validity of the shell model.

If the projectile goes into the target nucleus, it is likely to be absorbed completely, sharing its energy with the other nucleons in the nucleus. The resulting highly excited nucleus, called a compound nucleus, can decay in many different ways. The main constraints in determining the possible reactions are that the total number of neutrons and protons, the total charge, and the total energy must remain the same. These are the conservation laws for baryon number, charge, and energy. In principle, it is possible to change a neutron into a proton or vice versa in a hadronic reaction, but the probability is so small that this cannot be observed in the laboratory.

An example of a reaction following this more complex path is

$$a + {}^{14}\mathrm{N} \longrightarrow {}^{17}\mathrm{O} + p$$

This reaction was observed by Rutherford, who used alpha particles from a radioactive source. This provided enough energy to allow the alpha particle, a, to touch and fuse with the nitrogen-14 nucleus. There is a high probability that the resulting compound nucleus will decay by emitting a proton, but it can decay by other modes as well. The following example is a reaction that takes place when a hydrogen-containing medium is bombarded with low-energy neutrons, n:

$$n + {}^{1}\mathrm{H} \longrightarrow {}^{2}\mathrm{H} + \gamma$$

In this case, the energy released is equal to the binding energy of hydrogen-2, which is given off in the gamma ray, γ.

Many different kinds of reactions are possible in collisions induced by accelerated particles. Accelerators can give the projectiles enough energy to disrupt the nucleus completely and create new particles in the process. For example, a cyclotron can boost carbon nuclei to an energy of 500 MeV, enough to dissociate all the nucleons in a target nucleus of oxygen. In practice, such a reaction would involve many different processes, and the end result would be the fusion and dissociation of the nuclei into fragments of various sizes (see Figure 19). The annihilation of a pion in a nucleus is another example of a hadronic reaction. In this process, the rest energy of the pion would be released, dissociating the nucleus and giving kinetic energy to nucleons or other fragments.

Lepton-induced reactions

Nuclear reactions also can be induced by leptons such as electrons and neutrinos. The probability for an electron-induced reaction is smaller than for a hadronic reaction because of the difference in forces. The main process is inelastic scattering, in which the nucleus is excited to a higher energy level. If the inelastic scattering transfers a large amount of energy to the nucleus (*e.g.*, 20 MeV or larger), the nucleus becomes disrupted and ejects particles.

The weak interaction also causes reactions, but these are very difficult to observe. One part of the weak interaction changes a neutron into a proton or vice versa and brings about a corresponding conversion between an electron and a neutrino. The direction of the change is such as to conserve the total charge. An example of a neutrino-induced reaction is the reaction by which neutrinos were detected from a supernova in 1987. In this case, the neutrinos—or more precisely antineutrinos, $\bar{\nu}$—from the cosmic object interacted with protons of hydrogen atoms in a large tank of water. Positrons, e^+, were created according to the reaction

$$\bar{\nu} + p \longrightarrow n + e^{+}$$

* * *

The positrons were detected by the light they emitted as they traveled through the water. (G.F.B./S.McG.)

Isotopes

Every chemical element has one or more isotopes. All the isotopes of an element exhibit nearly identical chemical behaviour but have different masses. The differences in mass can be explained in terms of atomic structure.

As explained earlier, an atom is first identified and labeled according to the number of protons in its nucleus. This atomic number is ordinarily given the symbol Z. The great importance of the atomic number derives from the observation that all atoms with the same atomic number have nearly, if not precisely, identical chemical properties. A large collection of atoms with the same atomic number constitutes a sample of an element. A bar of pure uranium, for instance, would consist entirely of atoms with atomic number 92. The periodic table of the elements (see above) assigns one place to every atomic number, and each of these places is labeled with the common name of the element, as, for example, calcium, radon, or uranium.

Not all the atoms of an element need have the same number of neutrons in their nuclei. In fact, it is precisely the variation in the number of neutrons in the nuclei of atoms that gives rise to isotopes. Hydrogen is a case in point. It has the atomic number 1. Three nuclei with one proton are known that contain 0, 1, and 2 neutrons, respectively. The three share the place in the periodic table assigned to atomic number 1 and hence are called isotopes (from the Greek *isos,* meaning "same," and *topos,* signifying "place") of hydrogen.

Many important properties of an isotope depend on its mass. The total number of neutrons and protons (symbol A), or mass number, of the nucleus gives approximately the mass measured on the so-called atomic-mass-unit (amu) scale. The numerical difference between the actual measured mass of an isotope and A is called the mass defect (symbol Δ; see Table 3).

The specification of Z, A, and the chemical symbol (a one- or two-letter abbreviation of the element's name, say Sy) in the form ${}^{A}_{Z}\mathrm{Sy}$ identifies an isotope adequately for most purposes. Thus in the standard notation, ${}^{1}_{1}\mathrm{H}$ refers to the simplest isotope of hydrogen and ${}^{235}_{92}\mathrm{U}$ to an isotope of uranium widely used for nuclear power generation and nuclear weapons fabrication. (Authors who do not wish to use symbols sometimes write out the element name and mass number—hydrogen-1 and uranium-235 in the examples above.)

The term nuclide is used to describe particular isotopes, notably in cases where the nuclear rather than the chemical properties of an atom are to be emphasized. The lexicon of isotopes includes three other frequently used terms: isotones for isotopes of different elements with the same number of neutrons; isobars for isotopes of different elements with the same mass number; and isomers for isotopes identical in all respects except for the total energy content of the nuclei.

Isotones, isobars, and isomers

THE DISCOVERY OF ISOTOPES

Evidence for the existence of isotopes emerged from two independent lines of research, the first being the study of radioactivity. By 1910 it had become clear that certain processes associated with radioactivity, discovered some years before by Henri Becquerel, could transform one element into another. In particular, ores of the radioactive elements uranium and thorium had been found to contain small quantities of several radioactive substances never before observed. These substances were thought to be elements and accordingly received special names. Uranium ores, for example, yielded "ionium," and thorium ores gave "mesothorium." Painstaking work completed soon afterward revealed, however, that ionium, once mixed with ordinary thorium, could no longer be retrieved by chemical means alone. Similarly, mesothorium was shown to be chemically indistinguishable from radium. As chemists used the criterion of chemical indistinguishability as part of the definition of an element, they were forced to conclude that ionium and mesothorium were not new elements after all, but rather new forms of old ones. Generalizing from these and other data, Frederick Soddy in 1910 observed that "elements of different atomic weights may possess identical (chemical) properties" and so belong in the same place in the periodic table. With considerable prescience, he extended the scope of his conclusion to include not only radioactive species but stable elements as well. A few years later, Soddy published a comparison of the atomic weights of the stable element lead as measured in ores rich in uranium and thorium, respectively. He expected a difference because uranium and thorium

By courtesy of the Department of Physics and Astronomy, Michigan State University

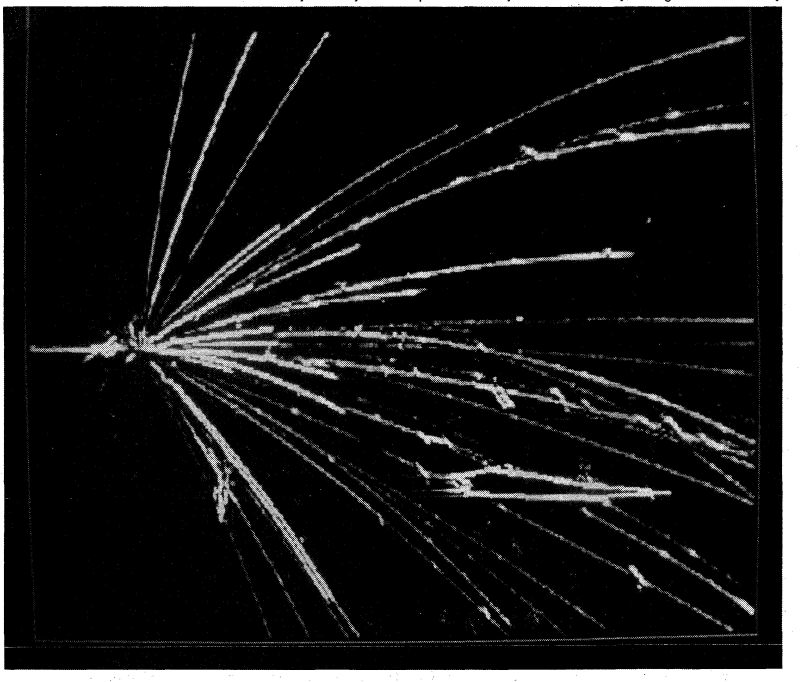


Figure 19: Particle tracks from the collision of an accelerated nucleus of a niobium atom with another niobium nucleus. The single line on the left is the track of the incoming projectile nucleus, and the other tracks are fragments from the collision.

decay into different isotopes of lead. The lead from the uranium-rich ore had an average atomic weight of 206.08 compared to 207.69 for the lead from the thorium-rich ore, thus verifying Soddy's conclusion.

The unambiguous confirmation of isotopes in stable elements not associated directly with either uranium or thorium followed a few years later with the development of the mass spectrograph by Francis William Aston. His work grew out of the study of positive rays (sometimes called canal rays), first discovered in 1886 by Eugen Goldstein and soon thereafter recognized as beams of positive ions. As a student in the laboratory of J.J. Thomson, Aston had learned that the gaseous element neon produced two positive rays. The ions in the heavier ray had masses about two units, or 10 percent, greater than the ions in the lighter ray. To prove that the lighter neon had a mass very close to 20 and that the heavier ray was indeed neon and not a spurious signal of some kind, Aston had to construct an instrument that was considerably more precise than any other of the time. By 1919 he had done so and convincingly argued for the existence of neon-20 and neon-22. Information from his and other laboratories accumulated rapidly in the ensuing years, and by 1935 the principal isotopes and their relative proportions were known for all but a handful of elements.

NUCLEAR STABILITY

Isotopes are said to be stable if, when left alone, they show no perceptible tendency to change spontaneously. Under the proper conditions, however, say in a nuclear reactor or particle accelerator or in the interior of a star, even stable isotopes may be transformed, one into another. The ease or difficulty with which these nuclear transformations occur varies considerably and reflects differing degrees of stability in the isotopes. Accordingly, it is important and useful to measure stability in more quantitative terms.

A uniform scale of nuclear stability, one that applies to stable and unstable isotopes alike, is based on a comparison of measured isotope masses with the masses of their constituent electrons, protons, and neutrons. For this purpose, electrons and protons are paired together as hydrogen atoms. The actual masses of all the stable isotopes differ appreciably from the sums of their individual particle masses. For example, the isotope $^{12}_{6}C$, which has a particularly stable nucleus, has an atomic mass defined to be exactly 12 amu. The total separate masses of 6 electrons and 6 protons, treated as 6 hydrogen atoms and 6 neutrons, add up to 12.09894 amu. The difference, Δm, between the actual mass of the assembled isotope and the masses of the particles gives a measure of the stability of the isotope: the larger and more negative the value of Δm, the greater the stability of the isotope. The difference in mass is often expressed as energy by using Einstein's equation in the form $E = \Delta mc^2$. Here, c is the speed of light. The quantity of energy calculated in this way is called the nuclear binding energy (E_B).

Calculating nuclear binding energies of nuclides

A single mathematical equation accurately reproduces the nuclear binding energies of more than 1,000 nuclides. It can be written in the form

$$E_B(\text{MeV}) = c_1 A[1 - k\{(N-Z)/A\}^2] - c_2 A^{2/3}[1 - k\{(N-Z)/A\}^2] - c_3 Z^2 A^{-1/3} + c_4 Z^2 A^{-1} + \delta.$$

The terms $c_1 = 15.677$, $c_2 = 18.56$, $c_3 = 0.717$, $c_4 = 1.211$, and $k = 1.79$, while δ may take any of several values (see below). The numerical values of these terms do not come from theory but from a selection process that ensures the best possible agreement with experimental data. On the other hand, theory helps justify, at least qualitatively, the mathematical form of each term. Modeled on an analogy to a liquid drop (see above), the first term represents the favourable contribution to the binding of the nucleus made by short-range, attractive nuclear forces between neutrons and protons. The second term corrects the first by allowing for the expectation that nucleons at the surface of the nucleus, unlike those in the interior, do not experience forces of nuclear attraction equally from all sides. Both the first and second terms have a second empirical component of the form $k[(N-Z)/A]^2$, which is referred to as the symmetry energy. It vanishes (neither helps nor hinders binding) when N is equal to Z (when the nucleus is "symmetric"), but then works increasingly to destabilize the nucleus as N and Z grow apart. The third term symbolizes the coulombic, or electrostatic, energy of repulsion of the protons; its derivation assumes a uniform distribution of charge within the nucleus. The fourth term makes a small correction to the third. This correction is necessitated by the observation that the nuclear charge distribution becomes somewhat more spread out near the surface of the nucleus. The last term, the so-called pairing energy, takes on any one of three values depending on whether N and Z are both even ($\delta = 11/A^{1/2}$), their sum is odd ($\delta = 0$), or both are odd ($\delta = -11/A^{1/2}$). More detailed treatments sometimes give other values for δ as well.

Symmetry energy

The largest observed deviations from the equation occur at certain favoured, or magic, numbers of neutrons or protons (2, 8, 20, 28, 50, 82, and 126). Magic nuclei are more stable than the binding energy equation would predict. The isotope of helium with 2 neutrons and 2 protons is said to be doubly magic. Various models of nuclear shell structure help to explain its stability.

Division of the binding energy E_B by A, the mass number, yields the binding energy per nucleon. This important quantity reaches a maximum value for nuclei in the vicinity of $^{56}_{26}Fe$. When two deuterium atoms fuse to form helium, the binding energy per nucleon increases and energy is released. Similarly, when the nucleus of an atom of ^{235}U fissions into two smaller nuclei, the binding energy per nucleon again increases with a concomitant release of energy.

RADIOACTIVE ISOTOPES

Only a small fraction of the isotopes are known to be stable indefinitely. All the others disintegrate spontaneously with the release of energy by processes broadly designated as radioactive decay. Each "parent" radioactive isotope eventually decays into one or at most a few stable isotope "daughters" specific to that parent. The radioactive parent tritium (^3H, or hydrogen-3), for example, always turns into the daughter helium-3 (^3He) by emitting an electron.

Radioactive decay

Under ordinary conditions, the disintegration of each radioactive isotope proceeds at a well-defined and characteristic rate. Thus, without replenishment, any radioactive isotope will ultimately vanish. Some isotopes, however, decay so slowly that they persist on Earth today even after the passage of more than 4,500,000,000 years since the last significant injection of freshly synthesized atoms from some nearby star. Examples of such long-lived radioisotopes include potassium-40, rubidium-87, neodymium-144, uranium-235, uranium-238, and thorium-232.

In this context, the widespread occurrence of radioisotopes that decay more rapidly, such as radon-222 and carbon-14, may at first seem puzzling. The explanation of the apparent paradox is that nuclides in this category are continually replenished by specialized nuclear processes: by the slow decay of uranium in the Earth in the case of radon and by the interactions of cosmic rays with the atmosphere in the case of carbon-14. Nuclear testing and the release of material from nuclear reactors also introduce radioactive isotopes into the environment.

ELEMENTAL AND ISOTOPIC ABUNDANCES

The composition of any object can be given as a set of elemental and isotopic abundances. One may speak, for example, of the composition of the ocean, the solar system, or indeed the Galaxy in terms of their respective elemental and isotopic abundances. Formally, the phrase elemental abundances usually connotes the amounts of the elements in an object expressed relative to one particular element (or isotope of it) selected as the standard for comparison. Isotopic abundances refer to the relative proportions of the stable isotopes of each element. They are most often quoted as atom percentages, as in Table 3.

Since the late 1930s, geochemists, astrophysicists, and nuclear physicists have joined together to try to explain the observed pattern of elemental and isotopic abundances. A more or less consistent picture has emerged. Hydrogen and

much helium are thought to have formed at the time of the "big bang"—the primordial explosion from which the universe is believed to have originated. The rest of the elements come, directly or indirectly, from stars. Cosmic rays (see below) produce a sizable proportion of the elements with mass numbers between 5 and 10; these elements are relatively rare. A substantial body of evidence shows that stars synthesize the heavier elements by nuclear processes collectively termed nucleosynthesis. In the first instance, then, nucleosynthesis determines the pattern of elemental abundances everywhere. The pattern is not immutable, for once matter escapes from stars it may undergo various processes of physical and chemical separation. A newly formed small planet, for example, may not exert enough gravitational attraction to capture the light gases hydrogen and helium. On the other hand, the processes that change elemental abundances normally alter isotopic abundances to a much lesser degree. Thus, virtually all terrestrial and meteoritic iron analyzed to date consists of 5.8 percent ^{54}Fe, 91.8 percent ^{56}Fe, 2.15 percent ^{57}Fe, and 0.29 percent ^{58}Fe. Table 3 lists the isotopic abundances of the stable elements and of a few radioactive elements as well. The relative constancy of the isotopic abundances makes it possible to tabulate meaningful average atomic weights for the elements. The availability of atomic weights is very important to chemists.

While there is general agreement on how the elements formed, the interpretation of elemental and isotopic abundances in specific bodies continues to occupy the attention of scientists. They obtain their raw data from several sources. Most knowledge concerning abundances comes from the study of the Earth, meteorites, and the Sun.

Solar system abundances of isotopes

Currently accepted estimates of solar-system (as opposed to terrestrial) abundances are pieced together mainly from two sources: (1) Chemical analyses of Type I carbonaceous chondrites, a special kind of meteorite, provide information about all but the most volatile elements—*i.e.*, those that existed as gases that the parent body of the meteorite could not trap in representative amounts. (2) Spectroscopic analysis of light from the Sun furnishes information about the volatile elements deficient in meteorites.

To the extent that the Sun resembles other stars, the elemental and isotopic abundances of the solar system have universal significance. Figure 20 shows the solar-system pattern. It has several notable features. First, the lighter isotopes, those of hydrogen and helium, constitute more than 98 percent of the mass; heavier isotopes make up scarcely 2 percent. Second, apart from the exceptions discussed below, as A or Z increases through the periodic table of the elements, abundances generally decrease. For example, the solar system as a whole contains about 1,000,000 times more carbon, nitrogen, and oxygen than the much heavier elements platinum and gold, though the proportions of the latter may vary widely from object to object. The decrease in abundance with increasing mass reflects in part the successive nature of nucleosynthesis. In nucleosynthesis, a nuclide of lower mass often serves as the seed or target for the production of a nuclide of

Table 3: Abundances of the Isotopes

element	Z	symbol	A	abundance	Δ (MeV)
Hydrogen	1	H	1	99.985	7.289
			2	0.0151*	13.136
Helium	2	He	3	1.38×10^{-4}*	14.931
			4	99.99986	2.425
Lithium	3	Li	6	7.5†	14.086
			7	92.5†	14.907
Beryllium	4	Be	9	100	11.348
Boron	5	B	10	19.9	12.051
			11	80.1	8.668
Carbon	6	C	12	98.90	0
			13	1.10	3.125
Nitrogen	7	N	14	99.634*	2.863
			15	0.367	0.102
Oxygen	8	O	16	99.762	−4.737
			17	0.038	−0.809
			18	0.200	−0.782
Fluorine	9	F	19	100	−1.487
Neon	10	Ne	20	90.51	−7.046
			21	0.27	−5.734
			22	9.22*	−8.027
Sodium	11	Na	23	100	−9.531
Magnesium	12	Mg	24	78.99	−13.933
			25	10.00	−13.192
			26	11.01	−16.214
Aluminum	13	Al	27	100	−17.197
Silicon	14	Si	28	92.23	−21.492
			29	4.67	−21.895
			30	3.10	−24.433
Phosphorus	15	P	31	100	−24.441
Sulfur	16	S	32	95.03	−26.016
			33	0.75	−26.587
			34	4.22	−29.932
			36	0.02	−30.664
Chlorine	17	Cl	35	75.77	−29.014
			37	24.23	−31.762
Argon	18	Ar	36	0.337	−30.231
			38	0.063	−34.715
			40	99.600	−35.040
Potassium	19	K	39	93.2581	−33.807
			40	0.0117‡	−33.535
			41	6.7302	−35.560
Calcium	20	Ca	40	96.941	−34.847
			42	0.647	−38.548
			43	0.135	−38.409
			44	2.087	−41.470
			46	0.0043	−43.138
			48	0.187	−44.216
Scandium	21	Sc	45	100	−41.070
Titanium	22	Ti	46	8.0	−44.126
			47	7.3	−44.932
			48	73.8	−48.487
			49	5.5	−48.558
			50	5.4	−51.426
Vanadium	23	V	50	0.250‡	−49.220
			51	99.750	−52.200
Chromium	24	Cr	50	4.35	−50.258
			52	83.79	−55.415
			53	9.50	−55.283
			54	2.37	−56.931
Manganese	25	Mn	55	100	−57.709
Iron	26	Fe	54	5.8	−56.251
			56	91.72	−60.604
			57	2.2	−60.179
			58	0.28	−62.152
Cobalt	27	Co	59	100	−62.226
Nickel	28	Ni	58	68.27	−60.225
			60	26.10	−64.471
			61	1.13	−64.220
			62	3.59	−66.746
			64	0.91	−67.098
Copper	29	Cu	63	69.17	−65.579
			65	30.83	−67.261
Zinc	30	Zn	64	48.6	−66.002
			66	27.9	−68.899
			67	4.1	−67.879
			68	18.8	−70.006
			70	0.6	−69.560
Gallium	31	Ga	69	60.1	−69.323
			71	39.9	−70.141
Germanium	32	Ge	70	20.5	−70.561
			72	27.4	−72.584
			73	7.8	−71.295
			74	36.6	−73.423
			76	7.8	−73.215
Arsenic	33	As	75	100	−73.035
Selenium	34	Se	74	0.9	−72.215
			76	9.0	−75.254
			77	7.6	−74.602
			78	23.7	−77.028
			80	49.8	−77.762
			82	9.2‡	−77.596
Bromine	35	Br	79	50.69	−76.070
			81	49.315	−77.977
Krypton	36	Kr	78	0.35	−74.152
			80	2.25	−77.892
			82	11.6	−80.591
			83	11.5	−79.983
			84	57.0	−82.431
			86	17.3	−83.263
Rubidium	37	Rb	85	72.17	−82.164
			87	27.85‡	−84.592
Strontium	38	Sr	84	0.56	−80.640
			86	9.86	−84.518
			87	7.0§	−84.875
			88	82.58	−87.916
Yttrium	39	Y	89	100	−87.702
Zirconium	40	Zr	90	51.45	−88.770
			91	11.22	−87.893
			92	17.15	−88.457
			94	17.38	−87.268
			96	2.80	−85.442
Niobium	41	Nb	93	100	−87.210
Molybdenum	42	Mo	92	14.84	−86.808
			94	9.25	−88.413
			95	15.92	−87.709
			96	16.68	−88.792
			97	9.55	−87.542
			98	24.13	−88.113
			100	9.63	−86.186
Ruthenium	44	Ru	96	5.52	−86.072
			98	1.89	−88.226
			99	12.7	−87.618
			100	12.6	−89.220
			101	17.0	−87.951
			102	31.6	−89.099
			104	18.7	−88.098
Rhodium	45	Rh	103	100	−88.027
Palladium	46	Pd	102	1.02	−87.902
			104	11.14	−89.397
			105	22.33	−88.419
			106	27.33	−89.910
			108	26.7	−89.523
			110	11.8	−88.335
Silver	47	Ag	107	51.839	−88.407
			109	48.17	−88.722
Cadmium	48	Cd	106	1.25	−87.133
			108	0.89	−89.261
			110	12.49	−90.351
			111	12.81	−89.255
			112	24.13	−90.582
			113	12.22‡	−89.051
			114	28.73	−90.023
			116	7.50	−88.721
Indium	49	In	113	4.3	−89.367
			115	95.7‡	−89.534

*In atmosphere; other sources variable. †Commercial sources of Li may be depleted in ^{6}Li and enriched in ^{7}Li. ‡Long-lived radioactive nuclide.
§Variations due to decay of long-lived, radioactive parent.
Source: E. Browne and R.B. Firestone, *Table of Radioactive Isotopes* (1986). John Wiley and Sons, New York.

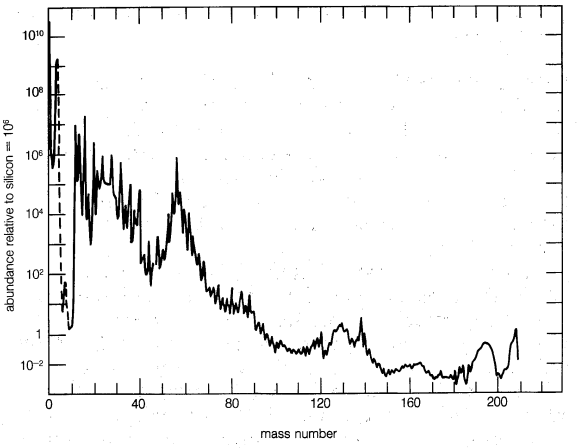


Figure 20: Solar system abundances.
From A.G.W. Cameron in C.A. Barnes, D.D. Clayton, and D.N. Schramm (eds.), *Essays in Nuclear Astrophysics* (1982); Cambridge University Press

higher mass. As the conversion of the lower mass target to the higher mass product is usually far from complete, abundances tend to decrease as mass increases. A third feature of interest is that stable isotopes with even numbers of protons and neutrons occur more often than do isotopes with odd ones (the so-called odd-even effect). Out of the almost 300 stable nuclides known, only five have odd numbers of both protons and neutrons; more than half have even values of Z and A-Z. Fourth, among the isotopes with even Z and A-Z certain species stand out by virtue of their considerable nuclear stability and comparatively high abundances. Nuclides that have equal and even numbers of neutrons and protons, the "alpha-particle" nuclides, fall into this category, which includes carbon-12, magnesium-24, and argon-36. Finally, peaks in the abundance distribution occur near the special values of Z and A-Z defined above as magic numbers. The high abundances manifest the extra nuclear stability that the magic numbers confer. Elements with enhanced abundances include nickel ($Z = 28$), tin ($Z = 50$), and lead ($Z = 82$).

The odd-even effect

The study of cosmic rays and of the light emitted by stars yields information about elemental and isotopic abundances outside the solar system. Cosmic rays are ions with high energy that are given off by stars. The Sun produces cosmic rays, too, but of much lower average energy than those reaching the solar system from outside. The abundance pattern in cosmic rays resembles that of the solar system in many ways. This fact suggests that solar and overall galactic abundances may be similar. Two explanations have been advanced to account for why solar and cosmic-ray abundances do not agree in all respects. The first is that cosmic rays undergo nuclear reactions as they pass through interstellar matter—*i.e.*, collisions that transform their nuclei. The second is that material from unusual stars with exotic compositions may be more prominent in either the solar system or in cosmic rays.

Table 3: Abundances of the Isotopes (continued)

element	Z	symbol	A	abundance	Δ (MeV)
Tin	50	Sn	112	0.97	−88.655
			114	0.65	−90.557
			115	0.36	−90.032
			116	14.53	−91.523
			117	7.69	−90.396
			118	24.22	−91.652
			119	8.58	−90.066
			120	32.59	−91.102
			122	4.63	−89.945
			124	5.79	−88.237
Antimony	51	Sb	121	57.4	−89.591
			123	42.8	−89.223
Tellurium	52	Te	120	0.096	−89.380
			122	2.60	−90.309
			123	0.908‡	−89.172
			124	4.817	−90.525
			125	7.14	−89.025
			126	18.95	−90.067
			128	31.69	−88.993
			130	33.80‡	−87.348
Iodine	53	I	127	100	−88.984
Xenon	54	Xe	124	0.10	−87.660
			126	0.090	−89.163
			128	1.91	−89.861
			129	26.4	−88.697
			130	4.1	−89.881
			131	21.2	−88.426
			132	26.9	−89.290
			134	10.4	−88.126
			136	8.9	−86.432
Cesium	55	Cs	133	100	−88.094
Barium	56	Ba	130	0.106	−87.300
			132	0.10	−88.454
			134	2.42	−88.973
			135	6.592	−87.874
			136	7.854	−88.910
			137	11.23	−87.737
			138	71.70	−88.277
Lanthanum	57	La	138	0.089‡	−86.531
			139	99.91	−87.238
Cerium	58	Ce	136	0.19	−86.500
			138	0.25	−87.575
			140	88.48	−88.089
			142	11.08	−84.542
Praseodymium	59	Pr	141	100	−86.027
Neodymium	60	Nd	142	27.13	−85.959
			143	12.19	−84.012
			144	23.80‡	−83.757
			145	8.31	−81.441
			146	17.2	−80.935
			148	5.76	−77.418
			150	5.64	−73.694
Samarium	62	Sm	144	3.1	−81.974
			147	15.0‡	−79.276
			148	11.3‡	−79.346
			149	13.8	−77.147
			150	7.4	−77.061
			152	26.7	−74.773
			154	22.7	−72.466
Europium	63	Eu	151	47.9	−74.663
			153	52.3	−73.379
Gadolinium	64	Gd	152	0.20‡	−74.719
			154	2.18	−73.718
			155	14.81	−72.082
			156	20.47	−72.547
			157	15.65	−70.835
			158	24.84	−70.702
			160	21.8	−67.943
Terbium	65	Tb	159	100	−69.544
Dysprosium	66	Dy	156	0.06	−70.537
			158	0.10	−70.419
			160	2.35	−69.683
			161	18.9	−68.065
			162	25.5	−68.190
			163	24.9	−66.390
			164	28.2	−65.977
Holmium	67	Ho	165	100	−64.908
Erbium	68	Er	162	0.14	−66.347
			164	1.61	−65.952
			166	33.6	−64.935
			167	22.95	−63.299
			168	26.8	−62.999
			170	14.9	−60.118
Thulium	69	Tm	169	100	−61.282
Ytterbium	70	Yb	168	0.13	−61.578
			170	3.05	−60.772
			171	14.3	−59.315
			172	21.9	−59.264
			173	16.12	−57.560
			174	31.8	−56.953
			176	12.7	−53.502
Lutetium	71	Lu	175	97.41	−55.173
			176	2.59‡	−53.395
Hafnium	72	Hf	174	0.162‡	−55.849
			176	5.206	−54.581
			177	18.606	−52.893
			178	27.297	−52.447
			179	13.629	−50.476
			180	35.100	−49.793
Tantalum	73	Ta	180	0.012	−48.860
			181	99.988	−48.445
Tungsten	74	W	180	0.13	−49.648
			182	26.3	−48.250
			183	14.3	−46.370
			184	30.67	−45.710
			186	28.6	−42.517
Rhenium	75	Re	185	37.40	−43.826
			187	62.60‡	−41.224
Osmium	76	Os	184	0.02	−44.257
			186	1.58‡	−43.007
			187	1.6	−41.227
			188	13.3	−41.145
			189	16.1	−38.995
			190	26.4	−38.717
			192	41.0	−35.893
Iridium	77	Ir	191	37.3	−36.716
			193	62.7	−34.543
Platinum	78	Pt	190	<0.02‡	−37.338
			192	0.79	−36.311
			194	32.9	−34.787
			195	33.9	−32.821
			196	25.3	−32.671
			198	7.2	−29.930
Gold	79	Au	197	100	−31.165
Mercury	80	Hg	196	0.14	−31.851
			198	10.02	−30.979
			199	16.84	−29.571
			200	23.13	−29.529
			201	13.22	−27.687
			202	29.80	−27.370
			204	6.85	−24.716
Thallium	81	Tl	203	29.524	−25.784
			205	70.477	−23.846
Lead	82	Pb	204	1.4	−25.132
			206	24.1	−23.809
			207	22.1	−22.476
			208	52.4	−21.772
Bismuth	83	Bi	209	100	−18.282
Thorium	90	Th	232	100‡	−35.447
Uranium	92	U	234	0.0055‡	38.142
			235	0.7200‡	40.915
			238	99.275‡	47.306

*In atmosphere; other sources variable. †Commercial sources of Li may be depleted in ^{6}Li and enriched in ^{7}Li. ‡Long-lived radioactive nuclide.
§Variations due to decay of long-lived, radioactive parent.
Source: E. Browne and R.B. Firestone, *Table of Radioactive Isotopes* (1986). John Wiley and Sons, New York.

The determination of elemental and isotopic abundances in stars of the Milky Way system and of more distant galaxies poses formidable experimental difficulties. Research in the field is active and reveals trends in composition among stars that are consistent with nucleosynthetic theory. The "metallicity"—or proportion of heavy elements—in stars, for instance, seems to increase with stellar age. In addition, many stars with compositions far different from that of the solar system are known. Their existence has led some investigators to doubt whether the concept of cosmic, as opposed to solar-system, abundances is meaningful. For the present it is perhaps enough to quote the American astrophysicist James W. Truran: "The local pattern of abundances is generally representative. The gross abundance features throughout our galaxy, in other galaxies, and even apparently in quasars are generally similar to those of solar system matter, testifying to the fact that the underlying stellar systems share the same nucleosynthetic processes."

VARIATIONS IN ISOTOPIC ABUNDANCES

Although isotopic abundances are fairly constant throughout the solar system, variations do occur. Variations in stable isotopic abundances are usually less than 1 percent, but they can be larger. Whatever their size, they provide geologists and astronomers with valuable clues to the histories of the objects under study. Several different processes can cause abundances to vary, among them radioactive decay and mass fractionation.

Radioactive decay. This process transmutes an isotope of one element into an isotope of another; *e.g.*, potassium-40 (^{40}K) to argon-40 (^{40}Ar) or uranium-235 (^{235}U) to lead-207 (^{207}Pb). As a consequence, the isotopic composition of the daughter element produced by the radioactive decay—argon or lead in the cases cited—may vary significantly from sample to sample. The variations become especially pronounced when the material under study forms with only a small amount of the daughter element present initially. The isotopic composition of argon in the earth's atmosphere is a case in point.

Transmutation of isotopes

Compared to stellar or solar-system abundances, atmospheric argon contains a much higher proportion of ^{40}Ar and much less ^{36}Ar and ^{38}Ar. The excess ^{40}Ar in the atmosphere evidently leaked out of crustal rocks and other potassium-bearing materials where it was produced by the decay of ^{40}K. Because the Earth trapped a relatively small amount of cosmically normal argon during its accretion, the ^{40}Ar generated since then by radioactive decay dominates the isotopic pattern in the atmosphere.

Mass fractionation. Physical and/or chemical processes affect differently the isotopes of an element. When the effect is systematic, increasing or decreasing steadily as mass number increases, the new pattern of isotopic abundances is said to be mass fractionated with respect to some standard pattern. For small fractionations—a few percent or less—the normal isotopic ratio M_h/M_l changes by an amount proportional to $\Delta m = M_h - M_l$, where M_l is the mass of the lighter isotope. For oxygen subjected to mass fractionation the percentage change of the ratio ^{18}O/^{16}O should be twice that in the ratio ^{17}O/^{16}O. Sometimes a set of samples will form from a single reservoir but with each one having experienced a different degree of mass fractionation. A graph of one isotopic ratio M_h/M_l against a second, $M_{h'}/M_l$ will then yield a straight line of slope $(M_h - M_l)/(M_{h'} - M_l)$. Such plots find important use in deciding whether groups of objects originated from a common source and how those groups evolved. When the oxygen isotope abundances of samples from the Earth and the Moon are considered in this way, the results suggest that both the planet and its satellite are members of a family of objects distinct from the families to which most meteorites belong.

Other causes of isotopic abundance variations. In addition to the known processes described, unknown causes also contribute to observed variations in isotopic abundances. To explain the variations scientists have speculated that certain stars, novas or supernovas, may have injected unusual packets of mass that somehow avoided thorough mixing with normal matter. Another proposal is that some samples underwent intense irradiation by high-energy particles.

PHYSICAL PROPERTIES ASSOCIATED WITH ISOTOPES

Broadly speaking, differences in the properties of isotopes can be attributed to either of two causes: differences in mass or differences in nuclear structure. Scientists usually refer to the former as isotope effects and to the latter by a variety of more specialized names. The isotopes of helium afford examples of both kinds. Mass effects are considered first.

Isotope effects

Helium has two stable isotopes, ^{3}He and ^{4}He, and exists in the gaseous state under normal conditions. At a given temperature and pressure, any volume of ^{4}He will weigh one-third more than the same volume of ^{3}He. More generally, for the same spatial distribution of atoms, the substance with the heavier isotope is expected to be denser. When deuterium, $^2_1$H, is substituted for hydrogen, $^1_1$H, to form heavy water, $^2_1H_2O$, its density is about 10 percent greater than that of normal H_2O.

A second difference related directly to mass concerns atomic velocities. Lighter species travel at higher average speeds. Atoms of ^{3}He, on the average, move 15 percent faster than those of gaseous ^{4}He at the same temperature. Many other properties that depend on atomic motion such as the thermal conductivity and viscosity of gases manifest predictable isotope effects.

Contrasts in the behaviour of the helium isotopes extend to the liquid and solid states. The contrasts are attributable to the effects of both mass and nuclear structure. Figure 21 shows which states or phases of helium are stable—

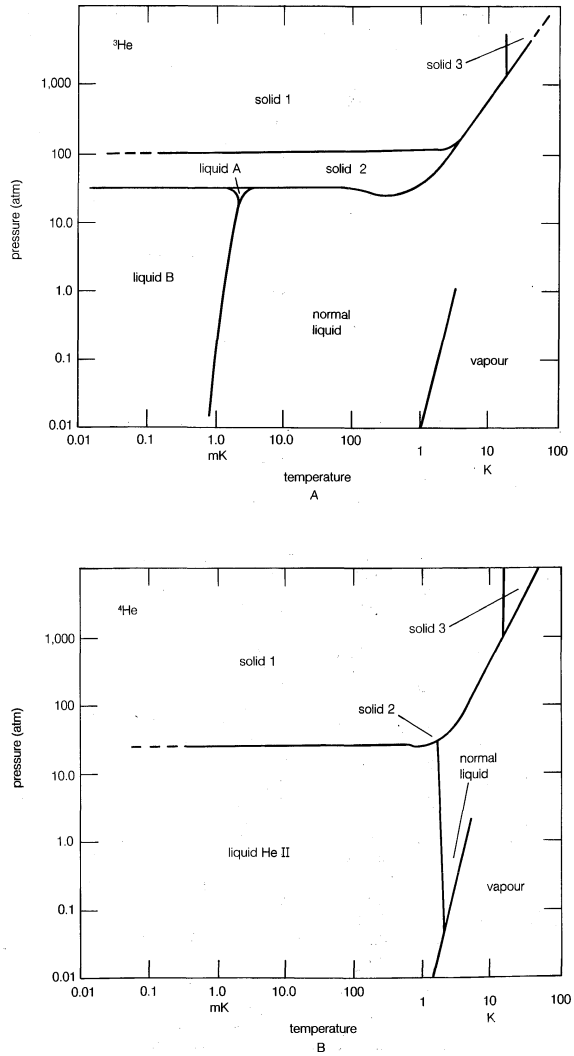


Figure 21: The phase diagrams of (A) helium-3 and (B) helium-4 show which states of these isotopes are stable (see text).

i.e., which actually occur at various temperatures and pressures. The lines on the diagrams delimit the ranges of stability of each phase. Although there are many similarities between the two diagrams, close examination reveals that they do not match up either quantitatively (in the positions of the lines) or qualitatively (in the types and numbers of phases at the lowest temperatures). It will be noted that ^{3}He forms three distinguishable liquid phases of which two are superfluids, while ^{4}He may exist only as two distinct liquids of which one is a superfluid. Unlike all other isotopes of the elements in the periodic table, neither ^{3}He nor ^{4}He solidifies under low pressures at a temperature near 0 Kelvin (K) (−273° C, or −459° F).

Effects of nuclear structure

Several other differences between isotopes depend on nuclear structure rather than on nuclear mass. First, radioactivity results from the interplay, distinctive for each nucleus, of nuclear and electrostatic forces between neutrons, protons, and electrons. Helium-6, for example, is radioactive, whereas helium-4 is stable. Second, nuclei may have differing quantities of angular momentum and may (or may not) act as tiny magnets, depending on the number and arrangement of their neutrons and protons. The magnetic properties of a nucleus play a crucial role in determining how the corresponding atom interacts with light. Finally, the spatial distribution of the protons in the nucleus affects in measurable ways the behaviour of the surrounding electrons. The addition of one neutron to an isotope allows the protons to spread out or to occupy a larger region of space. Electrons that have an appreciable probability of penetrating the nucleus are sensitive to this change. The addition of a neutron may also cause the nucleus to assume a nonspherical shape. A nonspherical distribution of charge in the nucleus affects the way in which the electrons behave, in particular the energies that they may emit or absorb in the form of light.

EFFECT OF ISOTOPES ON ATOMIC AND MOLECULAR SPECTRA

The study of how atoms and molecules interact with electromagnetic radiation, of which visible light is one form, goes by the name of spectroscopy. Spectroscopy has contributed much to the understanding of isotopes, and vice versa. To the extent that the characteristic spectrum of an atom or a molecule (*i.e.,* the light emitted or absorbed by it) is regarded as a physical property, the special relation between spectroscopy and isotopy warrants individual treatment here.

Atoms typically absorb or emit light exclusively at certain frequencies. Quantum mechanics explains this observation in a general way by associating with each atom (or molecule) well-defined states of energy. The atom may pass from one state to another only when energy is supplied (or removed) in the amount separating one state from another.

Precise measurements of the light emitted by isotopes of an element show small but significant differences termed shifts by spectroscopists. On the whole, these shifts are quite small. They originate in both mass and nuclear structure effects. The effects due to mass are largest for light isotopes. As nuclear mass increases, they decrease by an amount roughly proportional to $1/A^2$ and become insignificant in the heavier elements.

The effects due to nuclear structure relate primarily to the angular momentum, the magnetic moment, and the so-called electric quadrupole moment of the nucleus. The latter measures deviations from sphericity in the charge distribution. The magnetic moment and its attendant effects form the foundation of magnetic resonance spectroscopy, a field that has become very important in many branches of science.

When atoms join together in molecules, they can enter into characteristic vibrations and rotations. Just as an atom has a set of energy states associated primarily with the possible configurations of its electrons, so molecules have sets of energy states associated with their vibrations and rotations, as well as a set of electronic states. Light of the correct energy will induce changes from one vibrational (and/or rotational) state to another. Two ways in which isotopy relates to molecular vibrations, in particular, can be illustrated with the simplest of all molecules—diatomic molecules—which consist of only two atoms. Vibrational spectroscopy shows that isotopically heavier diatomic molecules have higher bond energies. (Bond energy is the amount of energy needed to separate the two atoms.) Quantum mechanical theory makes it possible to calculate from vibrational spectra just how much stronger the bond to the heavier isotope is. The differences between the chemical bond energies of isotopes help to explain why the isotopes do not behave identically in chemical reactions (see below). The second relation concerns the spacing between vibrational energy levels: the vibrational energy levels of an isotopically heavier molecule lie closer together. Consequently, it takes less energy to excite ^{18}O–^{18}O from one vibrational level to the next than it does ^{16}O–^{16}O. Spectroscopists made good use of this fact when they inferred from the spectra of isotopically mixed diatoms the existence of previously unknown isotopes. Oxygen-18 was discovered in this way.

Analysis of vibrational spectra

This second point, the distinguishability of the vibrational spectra of isotopically different molecules, is of great importance in the study of polyatomic molecules (molecules that contain three or more atoms). One key issue for chemists is the nature of the vibrations in polyatomic molecules: How do the nuclei of the atoms oscillate in relation to each other? The answer to this question bears strongly on what transient shapes the molecule may assume, how it will react with other molecules, and the rate at which it will do so. It is usually impossible to obtain this information from a study of the vibrational spectra of molecules made from atoms at natural abundance levels. Fortunately, the systematic substitution of heavier isotopes at known points in polyatomic molecules gives rise to new sets of vibrational spectra that clarify the nature of the atomic motions.

There is a second, fundamental reason for investigating the vibrational spectra of isotopically substituted, or "labeled," molecules. In interpreting spectra, spectroscopists rely on the mathematical results of quantum theory. Often, a close analysis of vibrational spectra of labeled molecules offers the best means for testing the soundness of the prevailing theoretical understanding of molecules.

CHEMICAL EFFECTS OF ISOTOPIC SUBSTITUTION

As isotopic abundances remain almost constant during most chemical processes, chemists do not normally distinguish the behaviour of one isotope from that of another. Indeed, in the limit of high temperatures, isotopes distribute themselves at random without preference for any particular chemical form. Nonetheless, under certain circumstances, nonrandom isotopic effects can become appreciable. Specifically, the lower the temperature and the lighter the isotope, the more noticeable the effects are likely to be. The reason is that the heavier isotopes tend to displace the lighter ones in those molecules where both heavy and light isotopes form the strongest chemical bond.

The exchange reaction $H_2 + D_2 \rightarrow 2HD$ provides an example of random behaviour at high temperature and isotope-specific behaviour at lower ones. If two volumes of gas consisting, respectively, of H_2 and D_2 only, are mixed, the hydrogen–hydrogen and deuterium–deuterium bonds will gradually break and new molecules will form until the vessel contains an appreciable quantity of H–D as well as of H_2 and D_2. At high temperatures the amount of HD observed at equilibrium approaches that predicted on the basis of probability (entropy) considerations alone—*i.e.,* a random distribution. How much would that be? A mathematical analysis shows that the concentrations of H_2, D_2, and HD should be equal to f_{H^2}, f_{D^2}, and $2f_Hf_D$, respectively, to a very good approximation. Here f represents the fraction of each kind of atom.

Experiment shows that as temperature increases, the concentrations of H_2, D_2 and HD approach the values expected. Although gratifying, the corroboration provides little information of chemical interest because the same results apply equally to the nitrogen isotopes ^{14}N and ^{15}N, to the chlorine isotopes ^{35}Cl and ^{37}Cl, and to many other pairs that differ greatly from hydrogen in their chemical behaviour. The variations from the random statistical distribution that occur at lower temperatures are more

interesting to a chemist because of what they reveal about the particular element.

At low temperatures the formation of D_2 (and H_2) is favoured at the expense of HD. A detailed theoretical treatment traces the cause of this favouritism to the comparative strength of the deuterium–deuterium bond. The result can be generalized: At equilibrium, the heavier isotope tends to concentrate wherever it forms the strongest chemical bond. For example, in the exchange reaction,

$$R-H+R'-D \rightarrow R-D+R'-H,$$

the hydrogen and deuterium switch partners. One may think of the hydrogen and deuterium as competing for the more attractive partner, supposed here to be R rather than R′. In accordance with the generalization above, the deuterium will tend to monopolize R with which, by hypothesis, it forms a stronger bond than it does with R′. Deuterium has a slight edge in the competition for R in spite of the fact that the hydrogen must also form a stronger bond with R than with R′.

Chemical equilibrium constants

Special quantities called chemical equilibrium constants express in quantitative terms the extent to which a chemical reaction favours products (the substances written to the right of the arrow) or reactants (the substances written to the left of the arrow). For reactions of the type cited, which chemists call exchange reactions, equilibrium constants are typically within a few percent of the values expected for a random distribution. The largest variations are observed for the low-Z elements, such as hydrogen; the variations are quite small for elements with higher atomic numbers, seldom exceeding 1 percent.

As implied above, the equilibrium constants for exchange reactions change slightly with temperature. This fact was put to use by the American chemist Harold C. Urey when he devised a method for inferring the temperature at which carbonates formed in the sea. He noted that, given a choice between water (H_2O) and carbonate (CO_3^{2-}, a principal constituent of seashells), the isotope ^{18}O shows a slight preference for the carbonate. The preference increases as temperature decreases. By measuring the $^{18}O/^{16}O$ ratio in a sample of carbonate and comparing it with the ratio in local seawater, it is possible to calculate a temperature at which the carbonate and the water equilibrated.

While isotopic substitutions usually change chemical equilibrium constants by small amounts, they can increase the rates of chemical reactions by a factor of 10 or more in the most extreme cases.

EFFECT OF ISOTOPIC SUBSTITUTION ON REACTION RATES

Chemical reactions take place when chemical bonds between atoms break or form. In the laboratory, chemical reactions proceed at well-defined rates. By introducing a heavy isotope into a reacting molecule, one may change the rate at which the molecule reacts. Two factors determine the size of the change.

The first factor is where the isotopic substitution is made in the reacting molecule. The largest effects, primary isotope effects, occur when one introduces a new isotope in the reaction "centre"—*i.e.*, the place in the molecule where chemical bonds are broken and/or formed during the reaction. If, on the other hand, the isotope is placed some distance from the reaction centre, it produces a much smaller, secondary isotope effect.

The second factor determining the size of the change in reaction rate is the relative, or percentage, difference in the masses of the original and substituted isotopes. The 300 percent difference in mass between ^3H (tritium) and ^1H can lead to more than 15-fold changes in reaction rates.

Both primary and secondary isotope effects decrease rapidly with increasing atomic number because the percentage difference in mass between isotopes tends to decrease. The substitution of deuterium for hydrogen, for example, may slow a reaction down by a factor of six. In contrast, the substitution of ^{18}O for ^{16}O would typically change a reaction rate by only a few percent. There is a much larger relative mass difference between hydrogen and deuterium than there is between ^{18}O and ^{16}O.

Primary isotope effects are often interpreted in terms of what is known as transition state theory. The theory postulates that to react, molecules must first reorganize themselves into a special, energy-rich configuration called a transition state. Other things being equal, the more energy required to form the transition state, the slower the reaction will be. A reaction in which a hydrogen atom shifts from one large molecule, symbolized as R–H, to another, symbolized as R′–H, furnishes an example:

The transition state theory

$$R-H+R' \rightarrow R\cdots H\cdots R' \rightarrow R+H-R'.$$

The middle structure with the dotted lines represents a transition state. The energy needed to form the transition state and hence the rate of reaction depends on the strength of the R–H bond among other factors. As deuterium would form a stronger bond to R than hydrogen, it follows that the substitution of deuterium for hydrogen would slow the reaction down. The amount by which the reaction slows down would depend heavily on just how much stronger the R–D bond is than the R–H bond.

ISOTOPE SEPARATION AND ENRICHMENT

Most elements are found as mixtures of several isotopes. For certain applications in industry, medicine, and science, samples enriched in one particular isotope are needed. Many methods have therefore been developed to separate the isotopes of an element from one another. Each method is based on some difference—sometimes a very slight one—between the physical or chemical properties of the isotopes of an element.

Mass spectrometry. Although the instrumentation normally serves analytical purposes, when suitably modified a mass spectrometer can also be used on a larger scale to prepare a purified sample of virtually any isotope. Uranium-235 for the first atomic bomb was separated with specially built mass spectrometers. Because of its high operational costs, this method is ordinarily restricted to the production of a few milligrams to a few grams of various stable isotopes for scientific investigation.

Distillation. The same factors that lead to the enrichment of alcohol in the vapour above a solution of water and alcohol permit the enrichment of isotopes. At temperatures below 220° C, for example, light water ($^1_1H_2O$) vaporizes to a slightly greater extent than heavy water ($^2_1H_2O$, or D_2O). The distillation of normal water, which contains both molecules, produces a vapour slightly enriched in $^1_1H_2O$. The residual liquid retains a correspondingly enhanced concentration of heavy water. It is usually, though not always, true that the molecule with the lighter isotope will be more volatile.

Chemical exchange reactions. Slight differences between the preferences of isotopes for one chemical form over another can serve as the basis for separation. The preparation of nitrogen enriched in ^{15}N by ion-exchange techniques illustrates this principle. Ammonia in water $NH_3(aq)$ will bind to a so-called ion-exchange resin (R–H). When poured over a vertical column of resin, a solution of ammonia reacts to form a well-defined horizontal band at the top of the column. The addition of a solution of lye (sodium hydroxide) will force the band of ammonia to move down the column. As the resin holds $^{15}NH_3$ slightly more tenaciously than $^{14}NH_3$, the $^{14}NH_3$ tends to concentrate at the leading, or bottom, edge of the band and the $^{15}NH_3$ at the trailing, or topmost, edge. Solutions depleted or enriched in ^{15}N are collected as they wash off the column.

Gaseous diffusion. Gases can diffuse through the small pores present in many materials. The diffusion proceeds in a random manner as gas molecules bounce unpredictably off the walls of the porous medium. The average time a molecule of gas takes to traverse such a barrier depends on its velocity and certain other factors. According to kinetic theory, at a given temperature a lighter molecule will have a larger average velocity than a heavier one. This result provides the basis for a separation method widely used to produce uranium enriched in the readily fissionable isotope ^{235}U, which is needed for nuclear reactors and nuclear weapons. (Natural uranium contains only about 0.7 percent ^{235}U, with the remainder of the isotopic mixture consisting almost entirely of ^{238}U.) In the separation process, natural uranium in the form of uranium hexafluoride

Uranium isotope separation and enrichment

(UF_6) gas is diffused from one compartment of a chamber to another through a porous barrier. Since the molecules of $^{235}UF_6$ travel at a higher velocity than those of $^{238}UF_6$, they pass into the second compartment more rapidly than the latter. Because the percentage of ^{235}U increases only slightly after traversal of the barrier, the process must be repeated hundreds of thousands of times to obtain the desired concentration of the isotope.

Gas centrifugation. When a mixture of gaseous molecules spins at high speed in a specially designed closed container, the heaviest species will concentrate near the outer walls and the lightest near the axis. The American physicist Jesse W. Beams used a gas centrifuge to separate isotopes, specifically the isotopes of chlorine, for the first time in 1936. Much subsequent work focused on the separation of $^{235}UF_6$ from $^{238}UF_6$ for which the gas centrifuge promised considerable savings in energy costs. Today, something less than 5 percent of the world's enriched uranium is produced by this method.

Photochemical enrichment methods. As discussed above, the frequencies of light absorbed by isotopes differ slightly. Once an atom has absorbed radiation and reached an excited state, its chemical properties may become quite different from what they were in the initial, or ground, state. Certain chemical and physical processes may proceed from an excited state that would not occur at all in the ground state. This observation is the nub of a method for isotope separation conceived some time ago but made practical only recently by the development of the laser. A laser produces light with such a narrow range of frequencies that it can excite one isotope without exciting neighbouring isotopes. Ordinary light sources produce much broader ranges of frequencies and therefore cannot excite isotopes selectively.

Several schemes of enrichment based on the use of lasers have been proposed because they promise to be relatively inexpensive compared to gas diffusion. Government-sponsored research in the United States has concentrated on a method that begins with metallic uranium. Upon heating in an oven, uranium vaporizes and escapes as a beam of particles through a small hole. Lasers tuned to the correct frequencies cause the ^{235}U atoms (but not the ^{238}U atoms) in the beam to be converted to a form (ions) in which they can be collected on a charged plate. (G.F.H.)

Radioactivity

As was noted above, an unstable nucleus will decompose spontaneously, or decay, into a more stable configuration but will do so only in a few specific ways by emitting certain particles or certain forms of electromagnetic energy. Radioactive decay is a property of several naturally occurring elements as well as of artificially produced isotopes of the elements. The rate at which a radioactive element decays is expressed in terms of its half-life; *i.e.,* the time required for one-half of any given quantity of the isotope to decay. Half-lives range from more than 1,000,000,000 years for some nuclei to less than 10^{-9} second (see below *Rates of radioactive transitions*). The product of a radioactive decay process—called the daughter of the parent isotope—may itself be unstable, in which case it, too, will decay. The process continues until a stable nuclide has been formed.

THE NATURE OF RADIOACTIVE EMISSIONS

The emissions of the most common forms of spontaneous radioactive decay are the alpha (α) particle, the beta (β) particle, the gamma (γ) ray, and the neutrino. The alpha particle is actually the nucleus of a helium-4 atom, with two positive charges ^4_2He. Such charged atoms are called ions. The neutral helium atom has two electrons outside its nucleus balancing these two charges. Beta particles may be negatively charged (beta minus, symbol e^-), also called a negatron, or positively charged (beta plus, symbol e^+), also called a positron. (The beta minus [β^-] particle is actually an electron created in the nucleus during beta decay without any relationship to the orbital electron cloud of the atom.) The positron is regarded as the antiparticle of the negatron because the two particles, when brought together, will mutually annihilate each other. Gamma rays are electromagnetic radiations like radio waves, light, and X rays. Beta radioactivity also produces the neutrino and antineutrino, particles that have no charge and no rest mass, symbolized by ν and $\bar{\nu}$, respectively.

Positive and negative beta particles

In the less common forms of radioactivity, fission fragments, neutrons, or protons may be emitted. Fission fragments are themselves complex nuclei with usually between one-third and two-thirds the charge Z and mass A of the parent nucleus. Neutrons and protons are, of course, the basic building blocks of complex nuclei, having approximately unit mass on the atomic scale and having zero charge or unit positive charge, respectively. The neutron cannot long exist in the free state. It is rapidly captured by nuclei in matter; otherwise, in free space it will undergo beta-minus decay to a proton, a negatron, and an antineutrino with a half-life of 12.8 minutes. The proton is the nucleus of ordinary hydrogen and is stable.

TYPES OF RADIOACTIVITY

The early work on natural radioactivity associated with uranium and thorium ores identified two distinct types of radioactivity: alpha and beta decay.

Alpha decay. In alpha decay, an energetic helium ion (alpha particle) is ejected, leaving a daughter nucleus of atomic number two less than the parent and of atomic mass number four less than the parent. An example is the decay (symbolized by an arrow) of the abundant isotope of uranium, ^{238}U, to a thorium daughter plus an alpha particle:

$$^{238}_{92}U \longrightarrow {}^{234}_{90}Th + {}^{4}_{2}He \qquad \begin{matrix} Q_a = 4.268 \text{ MeV} \\ t_{1/2} = 4.51 \times 10^9 \text{ years} \end{matrix}$$

Given for this and subsequent reactions are the energy released (Q) in millions of electron volts (MeV) and the half-life ($t_{1/2}$). It should be noted that in every reaction the charges, or number of protons, shown in subscript are in balance on both sides of the arrow, as are the atomic masses, shown in superscript.

Beta-minus decay. In beta-minus decay, or negatron emission, an energetic negative electron is emitted, producing a daughter nucleus of one higher atomic number and the same mass number. An example is the decay of the uranium daughter product thorium-234 into protactinium-234:

$$^{234}_{90}Th \longrightarrow {}^{234}_{91}Pa + e^- + \bar{\nu} \qquad \begin{matrix} Q_{\beta^+} = 0.263 \text{ MeV} \\ t_{1/2} = 24.1 \text{ days} \end{matrix}$$

In the above reaction for beta decay, $\bar{\nu}$ represents the antineutrino. (In the accepted relativistic theory a particle has a total mass m, given by Einstein's equation $E = mc^2$, in which c is the speed of light, and E is the total energy. Most particles have a characteristic rest mass, m_0, their mass when they are not moving. Neutrinos and antineutrinos have zero rest mass; hence, they must always have the speed of light.)

Gamma decay. A third type of radiation, gamma radiation, usually accompanies alpha or beta decay. Gamma rays are photons and are without rest mass or charge. Alpha or beta decay may simply proceed directly to the ground (lowest energy) state of the daughter nucleus without gamma emission, but the decay may also proceed wholly or partly to higher energy states (excited states) of the daughter. In the latter case, gamma emission may occur as the excited states transform to lower energy states of the same nucleus. (Alternatively to gamma emission, an excited nucleus may transform to a lower energy state by ejecting an electron from the cloud surrounding the nucleus. This orbital electron ejection is known as internal conversion and gives rise to an energetic electron and often an X ray as the atomic cloud fills in the empty orbital of the ejected electron. The ratio of internal conversion to the alternative gamma emission is called the internal-conversion coefficient.)

Gamma rays–photon emission

Isomeric transitions. There is a wide range of rates of half-lives for the gamma-emission process. Usually dipole transitions (see below *Gamma transition*), in which the

gamma ray carries off one $\hbar$ unit of angular momentum, are fast, less than nanoseconds (one nanosecond equals 10^{-9} second). The law of conservation of angular momentum requires that the sum of angular momenta of the radiation and daughter nucleus is equal to the angular momentum (spin) of the parent. If the spins of initial and final states differ by more than one, dipole radiation is forbidden, and gamma emission must proceed more slowly by a higher multipole (quadrupole, octupole, etc.) gamma transition. If the gamma-emission half-life exceeds about one nanosecond, the excited nucleus is said to be in a metastable, or isomeric, state (the names for a long-lived excited state), and it is customary to classify the decay as another type of radioactivity, an isomeric transition. An example of isomerism is found in the protactinium-234 nucleus of the uranium-238 decay chain:

somerism

$$^{234m}_{91}\text{Pa} \longrightarrow {}^{234}_{91}\text{Pa} + \gamma \qquad Q_\gamma = 0.0698 \text{ MeV}, \quad t_{1/2} = 1.17 \text{ min}$$

The letter *m* following the mass number stands for metastable and indicates a nuclear isomer.

Beta-plus decay. During the 1930s new types of radioactivity were found among the artificial products of nuclear reactions: beta-plus decay, or positron emission, and electron capture. In beta-plus decay an energetic positive electron is created and emitted, along with a neutrino, and the nucleus transforms to a daughter, lower by one in atomic number and the same in mass number. For instance, carbon-11 ($Z = 6$) decays to boron-11 ($Z = 5$), plus one positron and one neutrino:

$$^{11}_{6}\text{C} \longrightarrow {}^{11}_{5}\text{B} + e^+ + \nu \qquad Q_{\beta^+} = 0.97 \text{ MeV}, \quad t_{1/2} = 20.4 \text{ min}$$

Electron capture. Electron capture (EC) is a process in which decay follows the capture by the nucleus of an orbital electron. It is similar to positron decay in that the nucleus transforms to a daughter of one lower atomic number. It differs in that an orbital electron from the cloud is captured and annihilated by the nucleus with subsequent emission of an atomic X ray as the orbital vacancy is filled by an electron from the cloud about the nucleus. An example is the nucleus of beryllium-7 capturing one of its inner electrons to give lithium-7:

$$^{7}_{4}\text{Be} + e^- \longrightarrow {}^{7}_{3}\text{Li} + \nu \qquad Q_{\text{EC}} = 0.8616 \text{ MeV}, \quad t_{1/2} = 53 \text{ days}$$

The main features of radioactive decay of a nuclear species are often displayed in a decay scheme. Figure 22 shows the decay scheme of beryllium-7. Indicated are the half-life of the parent and that of the excited daughter state, as well as its energy 0.4774 MeV. The spins and parities of all three states are provided on the upper left-hand side of the level. The multipolarity of the gamma ray (magnetic dipole, M1, plus 0.005 percent electric quadrupole, E2) is indicated above the vertical arrow symbolizing the gamma transition. The slanted arrows symbolize the electron-capture decay with labels giving the percentage of decay directly to ground state (89.7 percent) and the percentage of EC decay going via the excited state (10.3 percent). The boldface numbers following the percentages are so-called log *ft* values, to be encountered below in connection with beta-decay rates. The overall energy release, Q_{EC}, is indicated below. The Q_{EC} is necessarily a calculated value because there is no general practical means of measuring the neutrino energies accompanying EC decay. With a few electron-capturing nuclides, it has been possible to measure directly the decay energy by measurement of a rare process called inner bremsstrahlung (braking radiation). In this process the energy release is shared between the neutrino and a gamma ray. The measured distribution of gamma-ray energies indicates the total energy release. Usually there is so much ordinary gamma radiation with radioactive decay that the inner bremsstrahlung is unobservable.

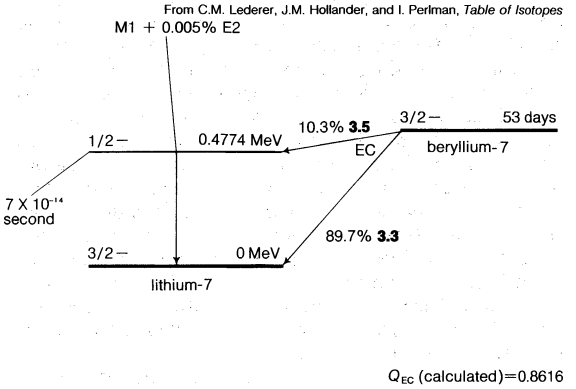


Figure 22: Radioactive decay of beryllium-7 to lithium-7 by electron capture (EC; see text).

Inner bremsstrahlung

Spontaneous fission. Yet another type of radioactivity is spontaneous fission. In this process the nucleus splits into two fragment nuclei of roughly half the mass of the parent. This process is only barely detectable in competition with the more prevalent alpha decay for uranium, but for some of the heaviest artificial nuclei, such as fermium-256, spontaneous fission becomes the predominant mode of radioactive decay. Kinetic-energy releases from 150 to 200 MeV may occur as the fragments are accelerated apart by the large electrical repulsion between their nuclear charges. The reaction is as follows:

$$^{256}_{100}\text{Fm} \longrightarrow {}^{140}_{54}\text{Xe} + {}^{112}_{46}\text{Pd} + 4n + \text{other fission products} \qquad Q = 150\text{–}200 \text{ MeV}, \quad t_{1/2} = 2.7 \text{ hours}$$

Only one of several product sets is shown. A few neutrons are always emitted in fission of this isotope, a feature essential to chain reactions. Spontaneous fission is not to be confused with induced fission, the process involved in nuclear reactors. It is a property of uranium-235, plutonium-239, and other isotopes to undergo fission after absorption of a slow neutron. Other than the requirement of a neutron capture to initiate it, induced fission is quite similar to spontaneous fission regarding total energy release, numbers of secondary neutrons, and so on (see below *Nuclear fission*).

Proton radioactivity. Proton radioactivity, discovered in 1970, is exhibited by an excited isomeric state of cobalt-53, ^{53m}Co, 1.5 percent of which emits protons:

$$^{53m}_{27}\text{Co} \xrightarrow{1.5\%} {}^{52}_{26}\text{Fe} + p \qquad Q_p = 1.57 \text{ MeV}$$
$$^{53m}_{27}\text{Co} \xrightarrow{98.5\%} {}^{53}_{26}\text{Fe} + e^+ + \nu \qquad t_{1/2} = 0.243 \text{ sec}$$

In addition to the above types of radioactivity, there is a special class of rare beta-decay processes that gives rise to heavy-particle emission. In these processes the beta decay partly goes to a high excited state of the daughter nucleus, and this state rapidly emits a heavy particle.

Special beta-decay processes. One such process is beta-delayed neutron emission, which is exemplified by the following reaction:

$$^{17}_{7}\text{N} \longrightarrow {}^{17*}_{8}\text{O} + e^- + \bar{\nu} \qquad Q_{\beta^-} = 8.68 \text{ MeV}$$
$$\longrightarrow {}^{16}_{8}\text{O} + n \qquad E_{max\ n} = 1.81 \text{ MeV}$$
$$t_{1/2} = 4.16 \text{ sec}$$

(Note: the asterisk denotes the short-lived intermediate excited states of oxygen-17, and $E_{max\ n}$ denotes the maximum energy observed for emitted neutrons.) There is a small production of delayed neutron emitters following nuclear fission, and these radioactivities are especially important in providing a reasonable response time to allow control of nuclear fission reactors by mechanically moved control rods.

Beta-delayed alpha emission

Among the positron emitters in the light-element region, a number beta decay partly to excited states that are unstable with respect to emission of an alpha particle. Thus, these species exhibit alpha radiation with the half-life of the beta emission. Both the positron decay from boron-8

and negatron decay from lithium-8 are beta-delayed alpha emission, because ground as well as excited states of beryllium-8 are unstable with respect to breakup into two alpha particles. Another example, sodium-20 (^{20}Na) to give successively neon-20 (^{20}Ne; the asterisk again indicating the short-lived intermediate state) and finally oxygen-16 is listed below:

$$^{20}_{11}\mathrm{Na} \longrightarrow {}^{20}_{10}\mathrm{Ne}^{*} + e^{+} + \nu \qquad Q_{\beta^+} = 13.0 \text{ MeV}$$
$$\hookrightarrow {}^{16}_{8}\mathrm{O} + \alpha \qquad E_{max\,\alpha} = 4.44 \text{ MeV}$$
$$t_{1/2} = 0.39 \text{ sec}$$

Beta-delayed proton emissions

In a few cases, positron decay leads to an excited nuclear state not able to bind a proton. In these cases, proton radiation appears with the half-life of the beta transition. The combination of high positron-decay energy and low proton-binding energy in the daughter ground state is required. In the example given below, tellurium-111 (^{111}Te) yields antimony-111 (^{111}Sb) and then tin-110 (^{110}Sn) successively:

$$^{111}_{52}\mathrm{Te} \longrightarrow {}^{111}_{51}\mathrm{Sb}^{*} + e^{+} + \nu \qquad Q_{\beta^+} \text{ uncertain}$$
$$\hookrightarrow {}^{110}_{50}\mathrm{Sn} + p \qquad E_{max\,p} = 3.7 \text{ MeV}$$
$$t_{1/2} = 19.5 \text{ sec}$$

Heavy-ion radioactivity. In 1980 A. Sandulescu, D.N. Poenaru, and W. Greiner described calculations indicating the possibility of a new type of decay of heavy nuclei intermediate between alpha decay and spontaneous fission. The first observation of heavy-ion radioactivity was that of a 30-MeV, carbon-14 emission from radium-223 by H.J. Rose and G.A. Jones in 1984. The ratio of carbon-14 decay to alpha decay is about 5×10^{-10}. Observations also have been made of carbon-14 from radium-222, radium-224, and radium-226, as well as neon-24 from thorium-230, protactinium-231, and uranium-232. Such heavy-ion radioactivity, like alpha decay and spontaneous fission, involves quantum-mechanical tunneling through the potential-energy barrier. Shell effects play a major role in this phenomenon, and in all cases observed to date the heavy partner of carbon-14 or neon-24 is close to doubly magic lead-208 (see below *Nuclear models*).

OCCURRENCE OF RADIOACTIVITY

Some species of radioactivity occur naturally on Earth. A few species have half-lives comparable to the age of the elements (about 6×10^9 years), so that they have not decayed away after their formation in stars. Notable among these are uranium-238, uranium-235, and thorium-232. Also, there is potassium-40, the chief source of irradiation of the body through its presence in potassium of tissue. Of lesser significance are the beta emitters vanadium-50, rubidium-87, indium-115, tellurium-123, lanthanum-138, lutetium-176, and rhenium-187, and the alpha emitters cerium-142, neodymium-144, samarium-147, gadolinium-152, dysprosium-156, hafnium-174, platinum-190, and lead-204. Besides these approximately 10^9-year species, there are the shorter-lived daughter activities fed by one or another of the above species; *e.g.*, by various nuclei of the elements between lead ($Z = 82$) and thorium ($Z = 90$).

Cosmic ray bombardment

Another category of natural radioactivity includes species produced in the upper atmosphere by cosmic ray bombardment. Notable are 5,720-year carbon-14 and 12.3-year tritium (hydrogen-3), 53-day beryllium-7, and 2,700,000-year beryllium-10. Meteorites are found to contain additional small amounts of radioactivity, the result of cosmic ray bombardments during their history outside the Earth's atmospheric shield. Activities as short-lived as 35-day argon-37 have been measured in fresh falls of meteorites. Nuclear explosions since 1945 have injected additional radioactivities into the environment, consisting of both nuclear fission products and secondary products formed by the action of neutrons from nuclear weapons on surrounding matter.

The fission products encompass most of the known beta emitters in the mass region 75–160. They are formed in varying yields, rising to maxima of about 7 percent per fission in the mass region 92–102 (light peak of the fission yield versus atomic mass curve) and 134–144 (heavy peak). Two kinds of delayed hazards caused by radioactivity are recognized. First, the general radiation level is raised by fallout settling to Earth. Protection can be provided by concrete or earth shielding until the activity has decayed to a sufficiently low level. Second, ingestion or inhalation of even low levels of certain radioactive species can pose a special hazard, depending on the half-life, nature of radiations, and chemical behaviour within the body. For a detailed discussion on the biologic effects of radiation, see RADIATION.

Nuclear reactors also produce fission products but under conditions in which the activities may be contained. Containment and waste-disposal practices should keep the activities confined and eliminate the possibility of leaching into groundwaters for times that are long compared to the half-lives. A great advantage of thermonuclear fusion power over fission power, if it can be practically realized, is not only that its fuel reserves, heavy hydrogen and lithium, are vastly greater than uranium, but also that the generation of radioactive fission product wastes can be largely avoided. In this connection, it may be noted that a major source of heat in the interior of both the Earth and the Moon is provided by radioactive decay. Theories about the formation and evolution of the Earth, Moon, and other planets must take into account these large heat production sources.

Heat from radioactive decay

Desired radioactivities other than natural activities and fission products may be produced either by irradiation of certain selected target materials by reactor neutrons or by charged particle beams or gamma ray beams of accelerators.

ENERGETICS AND KINETICS OF RADIOACTIVITY

Energy release in radioactive transitions. Consideration of the energy release of various radioactive transitions leads to the fundamental question of nuclear binding energies and stabilities. A much-used method of displaying nuclear-stability relationships is an isotope chart, those positions on the same horizontal row corresponding to a given proton number (Z) and those on the same vertical column to a given neutron number (N). Such a map is shown in Figure 23. The irregular bold line surrounds the region of presently known nuclei. The area encompassed by this is often referred to as the valley of stability because the chart may be considered a map of a binding energy surface, the lowest areas of which are the most stable.

From *Proceedings of the International Conference on Properties of Nuclei Far from the Region of Beta-Stability*, Leysin, 1970, (CERN 70-30)

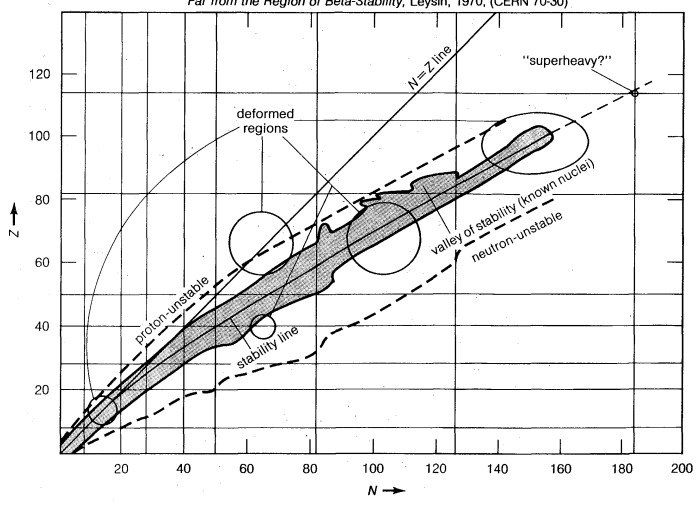


Figure 23: *Map of the nuclei.*
The coordinates are number of neutrons N and number of protons Z. The bold line encloses the region of known nuclei. The dashed lines are estimated limits for barely binding nucleons. The stability line gives the locus of nuclei most stable with respect to beta decay. Nuclei within the circular regions are cigar-shaped; otherwise the nuclei are spherical. The grid line intersections of the graph indicate the closed shell "magic numbers" of extra stability. A possible region of undiscovered longer-lived nuclei is indicated as "superheavy?"

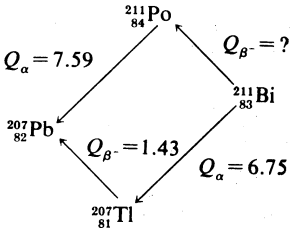
The most tightly bound nuclei of all are the abundant iron and nickel isotopes. Near the region of the valley containing the heaviest nuclei (largest mass number A; *i.e.*, largest number of nucleons, $N+Z$), the processes of alpha decay and spontaneous fission are most prevalent; both these processes relieve the energetically unfavourable concentration of positive charge in the heavy nuclei.

Along the region that borders on the valley of stability on the upper left-hand side are the positron-emitting and electron-capturing radioactive nuclei, with the energy release and decay rates increasing the farther away the nucleus is from the stability line. Along the lower right-hand border region, beta-minus decay is the predominant process, with energy release and decay rates increasing the farther the nucleus is from the stability line.

The grid lines of the graph are at the nucleon numbers corresponding to extra stability, the "magic numbers" (see below *Nuclear models*). The circles labeled "deformed regions" enclose regions in which nuclei should exhibit cigar shapes; elsewhere the nuclei are spherical. Outside the dashed lines nuclei would be unbound with respect to neutron or proton loss and would be exceedingly short-lived (less than 10^{-19} second).

Closed energy cycles

Calculation and measurement of energy. By the method of closed energy cycles, it is possible to use measured radioactive-energy-release (Q) values for alpha and beta decay to calculate the energy release for unmeasured transitions. An illustration is provided by the cycle of four nuclei below:



In this cycle, energies from two of the alpha decays and one beta decay are measurable. The unmeasured beta-decay energy for bismuth-211, $Q_{\beta^-}(\mathrm{Bi})$, is readily calculated because conservation of energy requires the sum of Q values around the cycle to be zero. Thus, $Q_{\beta^-}(\mathrm{Bi}) + 7.59 - 1.43 - 6.75 = 0$. Solving this equation gives $Q_{\beta^-}(\mathrm{Bi}) = 0.59$ MeV. This calculation by closed energy cycles can be extended from stable lead-207 back up the chain of alpha and beta decays to its natural precursor uranium-235 and beyond. In this manner the nuclear binding energies of a series of nuclei can be linked together. Because alpha decay decreases the mass number A by 4, and beta decay does not change A, closed α–β-cycle calculations based on lead-207 can link up only those nuclei with mass numbers of the general type $A = 4n + 3$, in which n is an integer. Another, the $4n$ series, has as its natural precursor thorium-232 and its stable end product lead-208. Another, the $4n + 2$ series, has uranium-238 as its natural precursor and lead-206 as its end product.

In early research on natural radioactivity, the classification of isotopes into the series cited above was of great significance because they were identified and studied as families. Newly discovered radioactivities were given symbols relating them to the family and order of occurrence therein. Thus, thorium-234 was known as UX_1, the isomers of protactinium-234 as UX_2 and UZ, uranium-234 as U_{II}, and so forth. These original symbols and names are occasionally encountered in more recent literature but are mainly of historical interest. The remaining $4n + 1$ series is not naturally occurring but comprises well-known artificial activities decaying down to stable thallium-205.

To extend the knowledge of nuclear binding energies, it is clearly necessary to make measurements to supplement the radioactive-decay energy cycles. In part, this extension can be made by measurement of Q values of artificial nuclear reactions. For example, the neutron-binding energies of the lead isotopes needed to link the energies of the four radioactive families together can be measured by determining the threshold gamma-ray energy to remove a neutron (photonuclear reaction); or the energies of incoming deuteron and outgoing proton in the reaction can be measured to provide this information.

Further extensions of nuclear-binding-energy measurements rely on precision mass spectroscopy (see ANALYSIS AND MEASUREMENT, PHYSICAL AND CHEMICAL). By ionizing, accelerating, and magnetically deflecting various nuclides, their masses can be measured with great precision. A precise measurement of the masses of atoms involved in radioactive decay is equivalent to direct measurement of the energy release in the decay process. The atomic mass of naturally occurring but radioactive potassium-40 is measured to be 39.964008 amu. Potassium-40 decays predominantly by β-emission to calcium-40, having a measured mass 39.962589. Through Einstein's equation, energy is equal to mass (m) times velocity of light (c) squared, or $E = mc^2$, the energy release (Q) and the mass difference, Δm, are related, the conversion factor being one amu, equal to 931.478 MeV. Thus, the excess mass of potassium-40 over calcium-40 appears as the total energy release Q_β in the radioactive decay $Q_{\beta^-} = (39.964008 - 39.962589) \times 931.478$ MeV $= 1.31$ MeV. The other neighbouring isobar (same mass number, different atomic number) to argon-40 is also of lower mass, 39.962384, than potassium-40. This mass difference converted to energy units gives an energy release of 1.5 MeV, this being the energy release for EC decay to argon-40. The maximum energy release for positron emission is always less than that for electron capture by twice the rest mass energy of an electron ($2m_0c^2 = 1.022$ MeV); thus, the maximum positron energy for this reaction is $1.5 - 1.02$, or 0.48 MeV.

Energy–mass conversions

To connect alpha-decay energies and nuclear mass differences requires a precise knowledge of the alpha-particle (helium-4) atomic mass. The mass of the parent minus the sum of the masses of the decay products gives the energy release. Thus, for alpha decay of plutonium-239 to uranium-235 and helium-4 the calculation goes as follows:

$M(^{239}\mathrm{Pu})$	239.05216
$-M(^{235}\mathrm{U})$	−235.04393
$-M(^{4}\mathrm{He})$	− 4.00260
	0.00563×931.478

$$Q_\alpha = 5.24 \text{ MeV}$$

By combining radioactive-decay-energy information with nuclear-reaction Q values and precision mass spectroscopy, extensive tables of nuclear masses have been prepared. From them the Q values of unmeasured reactions or decay may be calculated.

Alternative to the full mass, the atomic masses may be expressed as mass defect, symbolized by the Greek letter delta, Δ (the difference between the exact mass M and the integer A, the mass number), either in energy units or atomic mass units.

Absolute nuclear binding energy. The absolute nuclear binding energy is the hypothetical energy release if a given nuclide were synthesized from Z separate hydrogen atoms and N (equal to $A - Z$) separate neutrons. An example is the calculation giving the absolute binding energy of the stablest of all nuclei, iron-56:

$26 \times M(^1\mathrm{H})$	$26 \times 1.007825 =$	26.20345
$30 \times M(n)$	$30 \times 1.008665 =$	30.25995
$M(^{56}\mathrm{Fe})$	−	55.93493
	binding energy =	$0.52847 \times 931.478 =$ 492.58 MeV

average binding energy per nucleon of ^{56}Fe $= 492.58/56 = 8.796$ MeV

A general survey of the average binding energy per nucleon (for nuclei of all elements grouped according to ascending mass) shows a maximum at iron-56 falling off gradually on both sides to about 7 MeV at helium-4 and to about 7.4 MeV for the most massive nuclei known. Most of

the naturally occurring nuclei are thus not stable in an absolute nuclear sense. Nuclei heavier than iron would gain energy by degrading into nuclear products closer to iron, but it is only for the elements of greatest mass that the rates of degradation processes such as alpha decay and spontaneous fission attain observable rates. In a similar manner, nuclear energy is to be gained by fusion of most elements lighter than iron. The coulombic repulsion between nuclei, however, keeps the rates of fusion reactions unobservably low unless the nuclei are subjected to temperatures of greater than 10^7 K. Only in the hot cores of the Sun and other stars or in thermonuclear bombs or controlled fusion plasmas are these temperatures attained and nuclear-fusion energy released.

NUCLEAR MODELS

The liquid-drop model. The average behaviour of the nuclear binding energy can be understood with the model of a charged liquid drop. In this model, the aggregate of nucleons has the same properties of a liquid drop, such as surface tension, cohesion, and deformation. There is a dominant attractive-binding-energy term proportional to the number of nucleons A. From this must be subtracted a surface-energy term proportional to surface area and a coulombic repulsion energy proportional to the square of the number of protons and inversely proportional to the nuclear radius. Furthermore, there is a symmetry-energy term of quantum-mechanical origin favouring equal numbers of protons and neutrons. Finally, there is a pairing term that gives slight extra binding to nuclei with even numbers of neutrons or protons.

The pairing-energy term accounts for the great rarity of odd–odd nuclei (the terms odd–odd, even–even, even–odd, and odd–even refer to the evenness or oddness of proton number, Z, and neutron number, N, respectively) that are stable against beta decay. The sole examples are deuterium, lithium-6, boron-10, and nitrogen-14. A few other odd–odd nuclei, such as potassium-40, occur in nature, but they are unstable with respect to beta decay. Furthermore, the pairing-energy term makes for the larger number of stable isotopes of even-Z elements, compared to odd-Z, and for the lack of stable isotopes altogether in element 43, technetium, and element 61, promethium.

Mirror nuclei

The beta-decay energies of so-called mirror nuclei afford one means of estimating nuclear sizes. For example, the neon and fluorine nuclei, $^{19}_{10}Ne_9$ and $^{19}_{9}F_{10}$, are mirror nuclei because the proton and neutron numbers of one of them equal the respective neutron and proton numbers of the other. Thus, all binding-energy terms are the same in each except for the coulombic term, which is inversely proportional to the nuclear radius. Such calculations along with more direct determinations by high-energy electron scattering and energy measurements of X rays from mumesic atoms (hydrogen atoms in which the electrons are replaced by negative mu mesons) establish the nuclear charge as roughly uniformly distributed in a sphere of radius $1.2\ A^{1/3} \times 10^{-13}$ centimetre. That the radius is proportional to the cube root of the mass number has the great significance that the average density of all nuclei is nearly constant.

Careful examination of nuclear-binding energies reveals periodic deviations from the smooth average behaviour of the charged-liquid-drop model. An extra binding energy arises in the neighbourhood of certain numbers of neutrons or protons, the so-called magic numbers (2, 8, 20, 28, 50, 82, and 126). Nuclei such as $^{4}_{2}He_2$, $^{16}_{8}O_8$, $^{40}_{20}Ca_{20}$, $^{48}_{20}Ca_{28}$, and $^{208}_{82}Pb_{126}$ are especially stable species, doubly magic, in view of their having both proton and neutron numbers magic. These doubly magic nuclei are situated at the intersections of grid lines on Figure 23 above.

The shell model. In the preceding section, the overall trends of nuclear binding energies were described in terms of a charged-liquid-drop model. Yet there were noted periodic binding-energy irregularities at the magic numbers. The periodic occurrence of magic numbers of extra stability is strongly analogous to the extra electronic stabilities occurring at the atomic numbers of the noble-gas atoms (see above). The explanations of these stabilities are quite analogous in atomic and nuclear cases as arising from filling of particles into quantized orbitals of motion. The completion of filling of a shell of orbitals is accompanied by an extra stability. The nuclear model accounting for the magic numbers is, as previously noted, the shell model. In its simplest form, this model can account for the occurrence of spin zero for all even–even nuclear ground states; the nucleons fill pairwise into orbitals with angular momenta canceling. The shell model also readily accounts for the observed nuclear spins of the odd-mass nuclei adjacent to doubly magic nuclei, such as $^{208}_{82}Pb$. Here, the spins of 1/2 for neighbouring $^{207}_{81}Tl$ and $^{207}_{82}Pb$ are accounted for by having all nucleons fill pairwise into the lowest energy orbits and putting the odd nucleon into the last available orbital before reaching the doubly magic configuration (the Pauli exclusion principle dictates that no more than two nucleons may occupy a given orbital, and their spins must be oppositely directed); calculations show the last available orbitals below lead-208 to have angular momentum 1/2. Likewise, the spins of 9/2 for $^{209}_{82}Pb$ and $^{209}_{83}Bi$ are understandable because spin-9/2 orbitals are the next available orbitals beyond doubly magic lead-208. Even the associated magnetization, as expressed by the magnetic dipole moment, is rather well explained by the simple spherical-shell model.

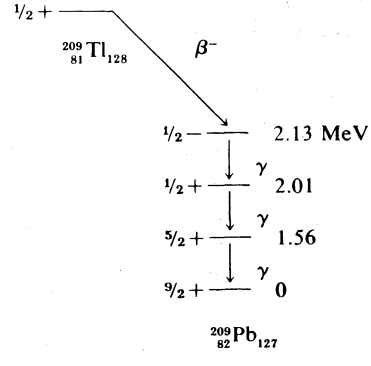
Doubly magic nuclei

The orbitals of the spherical-shell model are labeled in a notation close to that for electronic orbitals in atoms. The orbital configuration of calcium-40 has protons and neutrons filling the following orbitals: $1s_{1/2}$, $1p_{3/2}$, $1p_{1/2}$, $1d_{5/2}$, and $1d_{3/2}$. The letter denotes the orbital angular momentum in usual spectroscopic notation, in which the letters s, p, d, f, g, h, i, etc., represent integer values of l running from zero for s (not to be confused with spins) through six for i. The fractional subscript gives the total angular momentum j with values of $l + 1/2$ and $l - 1/2$ allowed, as the intrinsic spin of a nucleon is 1/2. The first integer is a radial quantum number taking successive values 1, 2, 3, etc., for successively higher energy values of an orbital of given l and j. Each orbital can accommodate a maximum of $2j + 1$ nucleons. Table 4 lists the orbitals comprising each shell, the exact order of various orbitals within a shell differing somewhat for neutrons and protons. The parity associated with an orbital is even (+) if l is even (s, d, g, i) and odd (−) if l is odd (p, f, h).

Table 4: Spherical-Shell-Model Orbitals

shell closure number	
2	$1s_{1/2}$
8	$1p_{3/2}$, $1p_{1/2}$
20	$1d_{5/2}$, $2s_{1/2}$, $1d_{3/2}$
28	$1f_{7/2}$
50	$2p_{3/2}$, $1f_{5/2}$, $2p_{1/2}$, $1g_{9/2}$
82	$1g_{7/2}$, $2d_{5/2}$, $1h_{11/2}$, $2d_{3/2}$, $3s_{1/2}$
126	$2f_{7/2}$, $1h_{9/2}$, $1i_{13/2}$, $3p_{3/2}$, $2f_{5/2}$, $3p_{1/2}$
184 (?)	$2g_{9/2}$, $1i_{11/2}$, $1j_{15/2}$, $3d_{5/2}$, $2g_{7/2}$, $4s_{1/2}$, $3d_{3/2}$

An example of a spherical-shell-model interpretation is provided by the beta-decay scheme of 2.2-minute thallium-209 shown below, in which spin and parity are given for each state. The ground and lowest excited states of lead-209 are to be associated with occupation by the 127th neutron of the lowest available orbitals above the closed shell of 126. From the last line of Table 4, it is to be noted



that there are available $g_{9/2}$, $d_{5/2}$, and $s_{1/2}$ orbitals with which to explain the ground and first two excited states. Low-lying states associated with the $i_{11/2}$ and $j_{15/2}$ orbitals are known from nuclear-reaction studies, but they are not populated in the beta decay.

The 2.13-MeV state that receives the primary beta decay is not so simply interpreted as the other states. It is to be associated with the promotion of a neutron from the $3p_{1/2}$ orbital below the 126 shell closure. The density (number per MeV) of states increases rapidly above this excitation, and the interpretations become more complex and less certain.

By suitable refinements, the spherical-shell model can be extended further from the doubly magic region. Primarily, it is necessary to drop the approximation that nucleons move independently in orbitals and to invoke a residual force, mainly short-range and attractive, between the nucleons. The spherical-shell model augmented by residual interactions can explain and correlate around the magic regions a large amount of data on binding energies, spins, magnetic moments, and the spectra of excited states.

The unified model. For nuclei more removed from the doubly magic regions, the spherical-shell model encounters difficulty in explaining the large observed electric quadrupole moments indicating cigar-shaped nuclei. For these nuclei a hybrid of liquid-drop and shell models, the unified model, has been proposed. (See the circular regions of Figure 23 for occurrence of cigar-shaped nuclei.)

Deformed regions

Nucleons can interact with one another in a collective fashion to deform the nuclear shape to a cigar shape. Such large spheroidal distortions are usual for nuclei far from magic, notably with $150 \lesssim A \lesssim 190$, and $224 \lesssim A$ (the symbol $<$ denotes less than, and $\sim$ means that the number is approximate). In these deformed regions the unified model prescribes that orbitals be computed in a cigar-shaped potential and that the relatively low-energy rotational excitations of the tumbling motion of the cigar shape be taken into account. The unified model has been highly successful in correlating and predicting nuclear properties in deformed regions. An example of a nuclear rotational band (a series of adjacent states) is provided by the decay of the isomer hafnium-180*m*, in Figure 24, through a cascade of gamma rays down the ground rotational band (see below *Gamma transition* for explanation of *M*2, *E*1, *E*2, and *E*3).

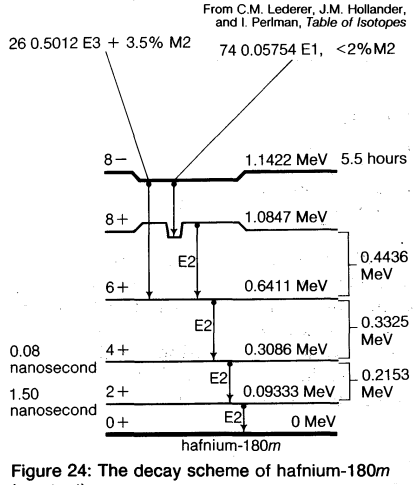


Figure 24: The decay scheme of hafnium-180*m* (see text).

RATES OF RADIOACTIVE TRANSITIONS

Equations governing radioactive decay

There is a vast range of the rates of radioactive decay, from undetectably slow to unmeasurably short. Before considering the factors governing particular decay rates in detail, it seems appropriate to review the mathematical equations governing radioactive decay and the general methods of rate measurement in different ranges of half-life.

Exponential-decay law. Radioactive decay occurs as a statistical exponential rate process. That is to say, the number of atoms likely to decay in a given infinitesimal time interval (dN/dt) is proportional to the number (N) of atoms present. The proportionality constant, symbolized by the Greek letter lambda, λ, is called the decay constant. Mathematically, this statement is expressed by the first-order differential equation,

Decay constant

$$-\frac{dN}{dt} = \lambda N. \tag{1}$$

This equation is readily integrated to give

$$N(t) = N_0 e^{-\lambda t}, \tag{2}$$

in which N_0 is the number of atoms present when time equals zero. From the above two equations it may be seen that a disintegration rate, as well as the number of parent nuclei, falls exponentially with time. An equivalent expression in terms of half-life $t_{1/2}$ is

$$N(t) = N_0(1/2)^r; \; r = t/t_{1/2}.$$

It can readily be shown that the decay constant λ and half-life ($t_{1/2}$) are related as follows: $\lambda = \log_e 2/t_{1/2} = 0.693/t_{1/2}$. The reciprocal of the decay constant λ is the mean life, symbolized by the Greek letter tau, τ.

For a radioactive nucleus such as potassium-40 that decays by more than one process (89 percent $\beta-$, 11 percent electron capture), the total decay constant is the sum of partial decay constants for each decay mode. (The partial half-life for a particular mode is the reciprocal of the partial decay constant times 0.693.) It is helpful to consider a radioactive chain in which the parent (generation 1) of decay constant λ_1 decays into a radioactive daughter (generation 2) of decay constant λ_2. The case in which none of the daughter isotope (2) is originally present yields an initial growth of daughter nuclei followed by its decay. The equation giving the number (N_2) of daughter nuclei existing at time t in terms of the number $N_1(0)$ of parent nuclei present when time equals zero is

$$N_2(t) = \lambda_1 N_1(0) \frac{e^{-\lambda_1 t} - e^{-\lambda_2 t}}{\lambda_2 - \lambda_1}, \tag{3}$$

in which e represents the logarithmic constant 2.71828.

Radioactive decay chain

The general equation for a chain of n generations with only the parent initially present (when time equals zero) is as follows:

$$N_n(t) = N_1(0)(C_1 e^{-\lambda_1 t} + C_2 e^{-\lambda_2 t} + \ldots . C_n e^{-\lambda_n t})\lambda_1\lambda_2 \ldots \lambda_{n-1}, \tag{4}$$

in which e represents the logarithmic constant 2.71828.

$$C_1 = 1/(\lambda_2 - \lambda_1)(\lambda_3 - \lambda_1) \ldots (\lambda_n - \lambda_1),$$

$$C_2 = 1/(\lambda_1 - \lambda_2)(\lambda_3 - \lambda_2) \ldots (\lambda_n - \lambda_2),$$

$$C_n = 1/(\lambda_1 - \lambda_n)(\lambda_2 - \lambda_n) \ldots (\lambda_{n-1} - \lambda_n).$$

These equations can readily be modified to the case of production of isotopes in the steady neutron flux of a reactor or in a star. In such cases, the chain of transformations might be mixed with some steps occurring by neutron capture and some by radioactive decay. The neutron-capture probability for a nucleus is expressed in terms of an effective cross-sectional area. If one imagines the nuclei replaced by spheres of the same cross-sectional area, the rate of reaction in a neutron flux would be given by the rate at which neutrons strike the spheres. The cross section is usually symbolized by the Greek letter sigma, σ, with the units of barns (10^{-24} cm^2) or millibarns (10^{-3} b) or microbarns (10^{-6} b). Neutron flux is often symbolized by the letters nv (neutron density, n, or number per cubic centimetre, times average speed, v) and given in neutrons per square centimetre per second.

The modification of the transformation equations merely involves substituting the product $nv\sigma_i$ in place of λ_i for any step involving neutron capture rather than radioactive decay. Reactor fluxes nv even higher than 10^{15} neutrons per square centimetre per second are available in several research reactors, but usual fluxes are somewhat lower by a factor of 1,000 or so. Tables of neutron-capture cross sections of the naturally occurring nuclei and some radioactive nuclei can be used for calculation of isotope production rates in reactors.

Measurement of half-life. The measurement of half-lives of radioactivity in the range of seconds to a few years commonly involves measuring the intensity of radiation at successive times over a time range comparable to the half-life. The logarithm of the decay rate is plotted against time, and a straight line is fitted to the points. The time interval for this straight-line decay curve to fall by a factor of 2 is read from the graph as the half-life, by virtue of equations (1) and (2). If there is more than one activity present in the sample, the decay curve will not be a straight line over its entire length, but it should be resolvable graphically (or by more sophisticated statistical analysis) into sums and differences of straight-line exponential terms. The general equations (4) for chain decays show a time dependence given by sums and differences of exponential terms, though special modified equations are required in the unlikely case that two or more decay constants are identically equal.

Decay curve for two or more activities

For half-lives longer than several years it is often not feasible to measure accurately the decrease in counting rate over a reasonable length of time. In such cases, a measurement of specific activity may be resorted to; *i.e.*, a carefully weighed amount of the radioactive isotope is taken for counting measurements to determine the disintegration rate, D. Then by equation (1) the decay constant λ_i may be calculated. Alternately, it may be possible to produce the activity of interest in such a way that the number of nuclei, N, is known, and again with a measurement of D equation (1) may be used. The number of nuclei, N, might be known from counting the decay of a parent activity or from knowledge of the production rate by a nuclear reaction in a reactor or accelerator beam.

Half-lives from 100 microseconds to one nanosecond are measured electronically in coincidence experiments. The radiation yielding the species of interest is detected to provide a start pulse for an electronic clock, and the radiation by which the species decays is detected in another device to provide a stop pulse. The distribution of these time intervals is plotted semi-logarithmically, as discussed for the decay-rate treatment, and the half-life is determined from the slope of the straight line.

Half-lives in the range of 100 microseconds to one second must often be determined by special techniques. For example, the activities produced may be deposited on rapidly rotating drums or moving tapes, with detectors positioned along the travel path. The activity may be produced so as to travel through a vacuum at a known velocity and the disintegration rate measured as a function of distance; however, this method usually applies to shorter half-lives in or beyond the range of the electronic circuit.

Species with half-lives shorter than the electronic measurement limit are not considered as separate radioactivities, and the various techniques of determining their half-lives will hence not be cited here.

Decay-rate considerations for various types of radioactivity are given here in the same order as listed above in *Types of radioactivity*.

Alpha decay. Alpha decay, the emission of helium ions, exhibits sharp line spectra when spectroscopic measurements of the alpha-particle energies are made. For even–even alpha emitters the most intense alpha group or line is always that leading to the ground state of the daughter. Weaker lines of lower energy go to excited states, and there are frequently numerous lines observable.

The main decay group of even–even alpha emitters exhibits a highly regular dependence on the atomic number, Z, and the energy release, Q_a. (Total alpha energy release, Q_α, is equal to alpha-particle energy, E_α, plus daughter recoil energy needed for conservation of momentum; $E_{\text{recoil}} = (m_a/[m_a + M_d])E_\alpha$, with m_a equal to the mass of the alpha particle and M_d the mass of the daughter product.) As early as 1911 the German physicist Johannes Wilhelm Geiger, together with the British physicist John Mitchell Nuttall, noted the regularities of rates for even–even nuclei and proposed a remarkably successful equation for the decay constant, $\log \lambda = a + b \log r$, in which r is the range in air, b is a constant, and a is given different values for the different radioactive series. The decay constants of odd alpha emitters (odd A or odd Z or both) are not quite so regular and may be much smaller. The values of the constant b that were used by Geiger and Nuttall implied a roughly 90th-power dependence of λ on Q_a. There is a tremendous range of known half-lives from the 2×10^{15} years of $^{144}_{60}$Nd (neodymium) with its 1.83-MeV alpha-particle energy (E_a) to the 0.3 microsecond of $^{212}_{84}$Po (polonium) with $E_a = 8.78$ MeV.

Geiger–Nuttall law

The theoretical basis for the Geiger–Nuttall empirical rate law remained unknown until the formulation of wave mechanics. A dramatic early success of wave mechanics was the quantitative theory of alpha-decay rates. One curious feature of wave mechanics is that particles may have a nonvanishing probability of being in regions of negative kinetic energy. In classical mechanics a ball that is tossed to roll up a hill will slow down until its gravitational potential energy equals its total energy, and then it will roll back toward its starting point. In quantum mechanics the ball has a certain probability of tunneling through the hill and popping out on the other side. For objects large enough to be visible to the eye, the probability of tunneling through energetically forbidden regions is unobservably small. For submicroscopic objects such as alpha particles, nucleons, or electrons, however, quantum mechanical tunneling can be an important process—as in alpha decay.

The logarithm of tunneling probability on a single collision with an energy barrier of height B and thickness D is a negative number proportional to thickness D, to the square root of the product of B and particle mass m. The size of the proportionality constant will depend on the shape of the barrier and will depend inversely on Planck's constant h.

In the case of alpha decay, the electrostatic repulsive potential between alpha particle and nucleus generates an energetically forbidden region, or potential barrier, from the nuclear radius out to several times this distance. The maximum height (B) of this alpha barrier is given approximately by the expression $B = 2Ze^2/R$, in which Z is the charge of the daughter nucleus, e is the elementary charge in electrostatic units, and R is the nuclear radius. Numerically, B is roughly equal to $2Z/A^{1/3}$, with A the mass number and B in energy units of MeV. Thus, although the height of the potential barrier for $^{212}_{84}$Po decay is nearly 28 MeV, the total energy released is $Q_a = 8.95$ MeV. The thickness of the barrier (*i.e.*, distance of the alpha particle from the centre of the nucleus at the moment of recoil) is about twice the nuclear radius of 8.8×10^{-13} centimetre. The tunneling calculation for the transition probability (P) through the barrier gives approximately

Tunneling of alpha particles

$$P = \exp\left[\left(-\frac{\sqrt{2MB}\,R}{\hbar}\right)\left(\frac{\pi B^{1/2}}{Q^{1/2}} - 4\right)\right], \qquad (5)$$

in which M is the mass of the alpha particle and $\hbar$ is Planck's constant h divided by 2π. By making simple assumptions about the frequency of the alpha particle striking the barrier, the penetration formula (5) can be used to calculate an effective nuclear radius for alpha decay. This method was one of the early ways of estimating nuclear sizes. In more sophisticated modern techniques the radius value is taken from other experiments, and alpha-decay data and penetrabilities are used to calculate the frequency factor.

The form of equation (5) suggests the correlation of decay rates by an empirical expression relating the half-life ($t_{1/2}$) of decay in seconds to the release energy (Q_a) in MeV:

$$\log t_{1/2} = \frac{a}{\sqrt{Q_a}} + b. \qquad (6)$$

Values of the constants a and b that give best fits to experimental rates of even–even nuclei with neutron number greater than 126 are given in Table 5. The nuclei with 126 or fewer neutrons decay more slowly than the heavier nuclei, and constants a and b must be readjusted to fit their decay rates.

The alpha-decay rates to excited states of even–even nuclei and to ground and excited states of nuclei with odd numbers of neutrons, protons, or both may exhibit retardations from equation (6) rates ranging to factors of thousands or more. The factor by which the rate is slower

than the rate formula (6) is the hindrance factor. The existence of uranium-235 in nature rests on the fact that alpha decay to the ground and low excited states exhibits hindrance factors of over 1,000. Thus the uranium-235 half-life is lengthened to 7×10^8 years, a time barely long enough compared to the age of the elements in the solar system for uranium-235 to exist in nature today.

Alpha hindrance factors

The alpha hindrance factors are fairly well understood in terms of the orbital motion of the individual protons and neutrons that make up the emitted alpha particle. The alpha-emitting nuclei heavier than radium are considered to be cigar-shaped, and alpha hindrance factor data have been used to infer the most probable zones of emission on the nuclear surface—whether polar, equatorial, or intermediate latitudes.

Table 5: Semi-empirical Constants*

	a	*b*
98 californium (Cf)	152.86	−52.9506
96 curium (Cm)	152.44	−53.6825
94 plutonium (Pu)	146.23	−52.0899
92 uranium (U)	147.49	−53.6565
90 thorium (Th)	144.19	−53.2644
88 radium (Ra)	139.17	−52.1476
86 radon (Rn)	137.46	−52.4597
84 polonium (Po)	129.35	−49.9229

*From correlation of ground-state decay rates of even–even nuclei with $N > 126$. See equation (6) in text.

Beta decay. The processes separately introduced at the beginning of this section as beta-minus decay, beta-plus decay, and orbital electron capture can be appropriately treated together. They all are processes whereby neutrons and protons may transform to one another by weak interaction. In striking contrast to alpha decay, the electrons (minus or plus charged) emitted in beta-minus and beta-plus decay do not exhibit sharp, discrete energy spectra but have distributions of electron energies ranging from zero up to the maximum energy release, Q_β. Furthermore, measurements of heat released by beta emitters (most radiation stopped in surrounding material is converted into heat energy) show a substantial fraction of the energy, Q_β, is missing. These observations, along with other considerations involving the spins or angular momenta of nuclei and electrons, led Wolfgang Pauli to postulate the simultaneous emission of the neutrino (1931). The neutrino, as a light and uncharged particle with nearly no interaction with matter, was supposed to carry off the missing heat energy. Today, neutrino theory is well accepted with the elaboration that there are six kinds of neutrinos, the electron neutrino, mu neutrino, and tau neutrino and corresponding antineutrinos of each. The electron neutrinos are involved in nuclear beta-decay transformations, the mu neutrinos are encountered in decay of muons (mu mesons) to electrons, and the tau neutrinos are produced when a massive lepton called a tau breaks down.

Although in general the more energetic the beta decay the shorter is its half-life, the rate relationships do not show the clear regularities of the alpha-decay dependence on energy and atomic number.

Fermi theory

The first quantitative rate theory of beta decay was given by Enrico Fermi in 1934, and the essentials of this theory form the basis of modern theory. As an example, in the simplest beta-decay process, a free neutron decays into a proton, a negative electron, and an antineutrino: $n \rightarrow p + e^- + \bar{\nu}$. The weak interaction responsible for this process, in which there is a change of species (n to p) by a nucleon with creation of electron and antineutrino, is characterized in Fermi theory by a universal constant, g. The sharing of energy between electron and antineutrino is governed by statistical probability laws giving a probability factor for each particle proportional to the square of its linear momentum (defined by mass times velocity for speeds much less than the speed of light and by a more complicated, relativistic relation for faster speeds). The overall probability law from Fermi theory gives the probability per unit time per unit electron energy interval, $P(W)$, as follows:

$$P(W) = \frac{64\pi^4 m_0^5 c^4 g^2}{h^7} W(W^2 - 1)^{1/2} (W_0 - W)^2 , \quad (7)$$

in which W is the electron energy in relativistic units ($W = 1 + E/m_0c^2$) and W_0 is the maximum ($W_0 = 1 + Q_\beta/m_0c^2$), m_0 the rest mass of the electron, c the speed of light, and h Planck's constant. This rate law expresses the neutron beta-decay spectrum in good agreement with experiment, the spectrum falling to zero at lowest energies by the factor W and falling to zero at the maximum energy by virtue of the factor $(W_0 - W)^2$.

In Fermi's original formulation, the spins of an emitted beta and neutrino are opposing and so cancel to zero. Later work showed that neutron beta decay partly proceeds with the $\frac{1}{2}\hbar$ spins of beta and neutrino adding to one unit of $\hbar$. The former process is known as Fermi decay (F) and the latter Gamow–Teller (GT) decay, after George Gamow and Edward Teller, the physicists who first proposed it. The interaction constants are determined to be in the ratio $g_{GT}^2/g_F^2 = 1.4$. Thus, g^2 in equation (7) should be replaced by $(g_F^2 + g_{GT}^2)$.

Parity violation in beta decay

The scientific world was shaken in 1957 by the measurement in beta decay of maximum violation of the law of conservation of parity. The meaning of this nonconservation in the case of neutron beta decay considered above is that the preferred direction of electron emission is opposite to the direction of the neutron spin of $\frac{1}{2}\hbar$. By means of a magnetic field and low temperature it is possible to cause neutrons in cobalt-60 and other nuclei, or free neutrons, to have their spins set preferentially in the up direction perpendicular to the plane of the coil generating the magnetic field. The fact that beta decay prefers the down direction for spin means that the reflection of the experiment as seen in a mirror parallel to the coil represents an unphysical situation: conservation of parity, obeyed by most physical processes, demands that experiments with positions reversed by mirror reflection should also occur. Further consequences of parity violation in beta decay are that spins of emitted neutrinos and electrons are directed along the direction of flight, totally so for neutrinos and partially so by the ratio of electron speed to the speed of light for electrons.

The overall half-life for beta decay of the free neutron, measured as 12 minutes, may be related to the interaction constants g^2 (equal to $g_F^2 + g_{GT}^2$) by integrating (summing) probability expression (7) over all possible electron energies from zero to the maximum. The result for the decay constant is

$$\lambda = \frac{64\pi^4 m_0^5 c^4 g^2}{h^7} \left\{ (W_0^2 - 1)^{1/2} \left(\frac{W_0^4}{30} - \frac{3W_0^2}{20} - \frac{2}{15} \right) + \right.$$
$$\left. + \frac{W_0}{4} \ln[W_0 + (W_0^2 - 1)^{1/2}] \right\} , \quad (8)$$

in which W_0 is the maximum beta-particle energy in relativistic units ($W_0 = 1 + Q_\beta/m_0c^2$), with m_0 the rest mass of the electron, c the speed of light, and h Planck's constant. The best g value from decay rates is approximately 10^{-49} erg per cubic centimetre. As may be noted from equation (8), there is a limiting fifth-power energy dependence for highest decay energies.

In the case of a decaying neutron not free but bound within a nucleus, the above formulas must be modified. First, as the nuclear charge Z increases, the relative probability of low-energy electron emission increases by virtue of the coulombic attraction. For positron emission, which is energetically impossible for free protons but can occur for bound protons in proton-rich nuclei, the nuclear coulomb charge suppresses lower energy positrons from the shape given by equation (7). This equation can be corrected by a factor $F(Z, W)$ depending on the daughter atomic number Z and electron energy W. The factor can be calculated quantum mechanically. The coulomb charge also affects the overall rate expression (8) such that it can no longer be expressed as an algebraic function, but tables are available for analysis of beta decay rates. The rates are analyzed in terms of a function $f(Z, Q_\beta)$ calculated by integration of equation (7) with correction factor $F(Z, W)$.

Approximate expressions for the f functions usable for decay energies Q between 0.1 MeV and 10 MeV, in which Q is measured in MeV, and Z is the atomic number of the daughter nucleus, are as follows (the symbol $\approx$ means approximately equal to):

$$f_{\beta^-} \approx 6.0Q^{4-0.005(Z-1)} \cdot 10^{Z/50},$$

$$f_{\beta^+} \approx 6.2Q^4/10^{0.007Z} \cdot 10^{0.009Z(\log 1/3Q)^2}.$$

For electron capture, a much weaker dependence on energy is found:

$$f_{EC} \approx (Z+1)^{3.5}Q/4 \times 10^5.$$

The basic beta decay rate expression obeyed by the class of so-called superallowed transitions, including decay of the neutron and several light nuclei is

$$\lambda_\beta = \frac{64\pi^4 m_0^5 c^4 g^2}{h^7} f_\beta. \qquad (9)$$

Like the ground-to-ground alpha transitions of even–even nuclei, the superallowed beta transitions obey the basic rate law, but most beta transitions go much more slowly. The extra retardation is explained in terms of mismatched orbitals of neutrons and protons involved in the transition. For the superallowed transitions the orbitals in initial and final states are almost the same. Most of them occur between mirror nuclei, with one more or less neutron than protons; *i.e.,* beta-minus decay of hydrogen-3, electron capture of beryllium-7 and positron emission of carbon-11, oxygen-15, neon-19, . . . titanium-43.

The nuclear retardation of beta decay rates below those of the superallowed class may be expressed in a fundamental way by multiplying the right side of equation (9) by the square of a nuclear matrix element (a quantity of quantum mechanics), which may range from unity down to zero depending on the degree of mismatch of initial and final nuclear states of internal motion. A more usual way of expressing the nuclear factor of the beta rate is the log ft value, in which f refers to the function $f(Z,Q_\beta)$. Because the half-life is inversely proportional to the decay constant λ, the product $f_\beta t_{1/2}$ will be a measure of (inversely proportional to) the square of the nuclear matrix element. For the log ft value, the beta half-life is taken in seconds, and the ordinary logarithm to the base 10 is used. The superallowed transitions have log ft values in the range of 3 to 3.5. Beta log ft values are known up to as large as ~23 in the case of indium-115. There is some correlation of log ft values with spin changes between parent and daughter nucleons, the indium-115 decay involving a spin change of four, whereas the superallowed transitions all have spin changes of zero or one.

Beta log ft values

Gamma transition. The nuclear gamma transitions belong to the large class of electromagnetic transitions encompassing radio-frequency emission by antennas or rotating molecules, infrared emission by vibrating molecules or hot filaments, visible light, ultraviolet light, and X-ray emission by electronic jumps in atoms or molecules. The usual relations apply for connecting frequency ν, wavelength λ, and photon quantum energy E with speed of light c and Planck's constant h; namely, $\lambda = c/\nu$ and $E = h\nu$. It is sometimes necessary to consider the momentum (p) of the photon given by $p = E/c$.

Classically, radiation accompanies any acceleration of electric charge. Quantum mechanically there is a probability of photon emission from higher to lower energy nuclear states, in which the internal state of motion involves acceleration of charge in the transition. Therefore, purely neutron orbital acceleration would carry no radiative contribution.

A great simplification in nuclear gamma transition rate theory is brought about by the circumstance that the nuclear diameters are always much smaller than the shortest wavelengths of gamma radiation in radioactivity—*i.e.,* the nucleus is too small to be a good antenna for the radiation. The simplification is that nuclear gamma transitions can be classified according to multipolarity, or amount of spin angular momentum carried off by the radiation. One unit of angular momentum in the radiation is associated with dipole transitions (a dipole consists of two separated equal charges, plus and minus). If there is a change of nuclear parity, the transition is designated electric dipole (E1) and is analogous to the radiation of a linear half-wave dipole radio antenna. If there is no parity change, the transition is magnetic dipole (M1) and is analogous to the radiation of a full-wave loop antenna. With two units of angular momentum change, the transition is electric quadrupole (E2), analogous to a full-wave linear antenna of two dipoles out-of-phase, and magnetic quadrupole (M2), analogous to coaxial loop antennas driven out-of-phase. Higher multipolarity radiation also frequently occurs with radioactivity.

Transition rates are usually compared to the single-proton theoretical rate, or Weisskopf formula, named after the American physicist Victor Frederick Weisskopf, who developed it. Table 6 gives the theoretical reference rate formulas in their dependence on nuclear mass number A and gamma-ray energy $E\gamma$ (in MeV).

Weisskopf formula

Table 6: Gamma Transition Rates*

transition type	partial half-life t_γ (seconds)	illustrative t_γ values for $A = 125$, $E = 0.1$ MeV (seconds)
E1	$5.7 \times 10^{-15} E^{-3} A^{-2/3}$	2×10^{-13}
E2	$6.7 \times 10^{-9} E^{-5} A^{-4/3}$	1×10^{-6}
E3	$1.2 \times 10^{-2} E^{-7} A^{-2}$	8
E4	$3.4 \times 10^{4} E^{-9} A^{-8/3}$	9×10^{7}
E5	$1.3 \times 10^{11} E^{-11} A^{-10/3}$	1×10^{15}
M1	$2.2 \times 10^{-14} E^{-3}$	2×10^{-11}
M2	$2.6 \times 10^{-8} E^{-5} A^{-2/3}$	1×10^{-4}
M3	$4.9 \times 10^{-2} E^{-7} A^{-4/3}$	8×10^{2}
M4	$1.3 \times 10^{5} E^{-9} A^{-2}$	8×10^{9}
M5	$5.0 \times 10^{11} E^{-11} A^{-8/3}$	1×10^{17}

*The energies E are expressed in MeV. The nuclear radius has been taken as 1.3 fermis. It is to be noted that t_γ is the partial half-life for γ emission only; the occurrence of internal conversion will always shorten the measured half-life.

It is seen for the illustrative case of gamma energy 0.1 MeV and mass number 125 that there occurs an additional factor of 10^7 retardation with each higher multipole order. For a given multipole, magnetic radiation should be a factor of 100 or so slower than electric. These rate factors ensure that nuclear gamma transitions are nearly purely one multipole, the lowest permitted by the nuclear spin change. There are many exceptions, however; mixed M1–E2 transitions are common, because E2 transitions are often much faster than the Weisskopf formula gives and M1 transitions are generally slower. All E1 transitions encountered in radioactivity are much slower than the Weisskopf formula. The other higher multipolarities show some scatter in rates, ranging from agreement to considerable retardation. In most cases the retardations are well understood in terms of nuclear model calculations.

Though not literally a gamma transition, electric monopole (E0) transitions may appropriately be mentioned here. These may occur when there is no angular momentum change between initial and final nuclear states and no parity change. For spin-zero to spin-zero transitions, single gamma emission is strictly forbidden. The electric monopole transition occurs largely by the ejection of electrons from the orbital cloud in heavier elements and by positron–electron pair creation in the lighter elements.

Electric monopole transitions

APPLICATIONS OF RADIOACTIVITY

In medicine. Radioisotopes have found extensive use in diagnosis and therapy, and this has given rise to a rapidly growing field called nuclear medicine. These radioactive isotopes have proven particularly effective as tracers in certain diagnostic procedures. As radioisotopes are identical chemically with stable isotopes of the same element, they can take the place of the latter in physiological processes. Moreover, because of their radioactivity, they can be readily traced even in minute quantities with such detection devices as gamma-ray spectrometers and proportional counters. Though many radioisotopes are used as tracers, iodine-131, phosphorus-32, and technetium-99m are among the most important. Physicians employ iodine-131 to determine cardiac output, plasma volume, and fat metabolism and particularly to measure the activity of the thyroid gland where this isotope accumulates. Phosphorus-

Use as tracers in diagnostic procedures

32 is useful in the identification of malignant tumours because cancerous cells tend to accumulate phosphates more than normal cells do. Technetium-99*m*, used with radiographic scanning devices, is valuable for studying the anatomic structure of organs.

Such radioisotopes as cobalt-60 and cesium-137 are widely used to treat cancer. They can be administered selectively to malignant tumours and so minimize damage to adjacent healthy tissue.

In industry. Foremost among industrial applications is power generation based on the release of the fission energy of uranium (see below *Nuclear fission* and also the article ENERGY CONVERSION: *Nuclear reactors*). Other applications include the use of radioisotopes to measure (and control) the thickness or density of metal and plastic sheets, to stimulate the cross-linking of polymers, to induce mutations in plants in order to develop hardier species, and to preserve certain kinds of foods by killing microorganisms that cause spoilage. In tracer applications radioactive isotopes are employed, for example, to measure the effectiveness of motor oils on the wearability of alloys for piston rings and cylinder walls in automobile engines. For additional information about industrial uses, see RADIATION: *Applications in science and industry.*

Radiometric dating

In science. Research in the Earth sciences has benefited greatly from the use of radiometric-dating techniques, which are based on the principle that a particular radioisotope (radioactive parent) in geologic material decays at a constant known rate to daughter isotopes. Using such techniques, investigators have been able to determine the ages of various rocks and rock formations and thereby quantify the geologic time scale (see GEOCHRONOLOGY: *Radiometric dating*). A special application of this type of radioactivity age method, carbon-14 dating, has proved especially useful to physical anthropologists and archaeologists. It has helped them to better determine the chronological sequence of past events by enabling them to date more accurately fossils and artifacts from 500 to 50,000 years old.

Radioisotopic tracers are employed in environmental studies, as, for instance, those of water pollution in rivers and lakes and of air pollution by smokestack effluents. They also have been used to measure deep-water currents in oceans and snow-water content in watersheds. Researchers in the biologic sciences, too, have made use of radioactive tracers to study complex processes. For example, thousands of plant metabolic studies have been conducted on amino acids and compounds of sulfur, phosphorus, and nitrogen. (J.O.R./E.P.S.)

Energy from atoms

NUCLEAR FISSION

Spontaneous and induced fission

Nuclear fission is the breakup of the nucleus of an atom into two lighter nuclei. Occurring primarily in heavy nuclei, this process may take place spontaneously in some cases or may be induced by the excitation of the nucleus with a variety of particles (*e.g.*, neutrons, protons, deuterons, or alpha particles) or with electromagnetic radiation in the form of gamma rays. In the fission process, a large quantity of energy is released, radioactive products are formed, and several neutrons are emitted. These neutrons can induce fission in a nearby nucleus of fissionable material and release more neutrons that can repeat the sequence, causing a chain reaction in which a large number of nuclei undergo fission and an enormous amount of energy is released. If controlled in a nuclear reactor, such a chain reaction can provide power for society's benefit. If uncontrolled, as in the case of the so-called atomic bomb, it can lead to an explosion of awesome destructive force.

The discovery of nuclear fission has opened a new era—the "Atomic Age." The potential of nuclear fission for good or evil and the risk/benefit ratio of its applications have not only provided the basis of many sociological, political, economic, and scientific advances but grave concerns as well. Even from a purely scientific perspective, the process of nuclear fission has given rise to many puzzles and complexities, and a complete theoretical explanation is still not at hand.

History of fission research and technology. The term fission was first used by the German physicists Lise Meitner and Otto Frisch in 1939 to describe the disintegration of a heavy nucleus into two lighter nuclei of approximately equal size. The conclusion that such an unusual nuclear reaction can in fact occur was the culmination of a truly dramatic episode in the history of science, and it set in motion an extremely intense and productive period of investigation.

The story of the discovery of nuclear fission actually began with the discovery of the neutron in 1932 by James Chadwick in England (see above). Shortly thereafter, Enrico Fermi and his associates in Italy undertook an extensive investigation of the nuclear reactions produced by the bombardment of various elements with this uncharged particle. In particular, these workers observed (1934) that at least four different radioactive species resulted from the bombardment of uranium with slow neutrons. These newly discovered species emitted beta particles and were thought to be isotopes of unstable "transuranium elements" of atomic numbers 93, 94, and perhaps higher. There was, of course, intense interest in examining the properties of these elements, and many radiochemists participated in the studies. The results of these investigations, however, were extremely perplexing, and confusion persisted until 1939 when Otto Hahn and Fritz Strassmann in Germany, following a clue provided by Irène Joliot-Curie and Pavle Savić in France (1938), proved definitely that the so-called transuranic elements were in fact radioisotopes of barium, lanthanum, and other elements in the middle of the periodic table.

Discovery of fission by Hahn and Strassmann

That lighter elements could be formed by bombarding heavy nuclei with neutrons had been suggested earlier (notably by the German chemist Walter Noddack in 1934), but the idea was not given serious consideration because it entailed such a broad departure from the accepted views of nuclear physics and was unsupported by clear chemical evidence. Armed with the unequivocal results of Hahn and Strassmann, however, Meitner and Frisch invoked the recently formulated liquid-drop model of the nucleus (see above) to give a qualitative theoretical interpretation of the fission process and called attention to the large energy release that should accompany it. There was almost immediate confirmation of this reaction in dozens of laboratories throughout the world, and within a year more than 100 papers describing most of the important features of the process were published. These experiments confirmed the formation of extremely energetic heavy particles and extended the chemical identification of the products.

The chemical evidence that was so vital in leading Hahn and Strassmann to the discovery of nuclear fission was obtained by the application of carrier and tracer techniques. Since invisible amounts of the radioactive species were formed, their chemical identity had to be deduced from the manner in which they followed known carrier elements, present in macroscopic quantity, through various chemical operations. Known radioactive species were also added as tracers and their behaviour was compared with that of the unknown species to aid in the identification of the latter. Over the years, these radiochemical techniques have been used to isolate and identify some 34 elements from zinc (atomic number 30) to gadolinium (atomic number 64) that are formed as fission products. The wide range of radioactivities produced in fission makes this reaction a rich source of tracers for chemical, biologic, and industrial use (see above).

Although the early experiments involved the fission of ordinary uranium with slow neutrons, it was rapidly established that the rare isotope uranium-235 was responsible for this phenomenon. The more abundant isotope uranium-238 could be made to undergo fission only by fast neutrons with energy exceeding 1 MeV. The nuclei of other heavy elements, such as thorium and protactinium, also were shown to be fissionable with fast neutrons; and other particles, such as fast protons, deuterons, and alphas, along with gamma rays, proved to be effective in inducing the reaction.

In 1939, Frédéric Joliot-Curie, Hans von Halban, and Lew Kowarski found that several neutrons were emitted

in the fission of uranium-235, and this discovery led to the possibility of a self-sustaining chain reaction. Fermi and his coworkers recognized the enormous potential of such a reaction if it could be controlled. On Dec. 2, 1942, they succeeded in doing so, operating the world's first nuclear reactor. Known as a "pile," this device consisted of an array of uranium and graphite blocks and was built on the campus of the University of Chicago.

First controlled chain reaction

The secret Manhattan Project, established not long after the United States entered World War II, developed the atomic bomb. Once the war had ended, efforts were made to develop new reactor types for large-scale power generation, giving birth to the nuclear power industry.

Fundamentals of the fission process. *Structure and stability of nuclear matter.* The fission process may be best understood through a consideration of the structure and stability of nuclear matter. Nuclei consist of nucleons (neutrons and protons), the total number of which is equal to the mass number of the nucleus. The actual mass of a nucleus is always less than the sum of the masses of the free neutrons and protons that constitute it, the difference being the mass equivalent of the energy of formation of the nucleus from its constituents. The conversion of mass to energy follows Einstein's equation, $E = mc^2$, where E is the energy equivalent of a mass, m, and c is the velocity of light. This difference is known as the mass defect and is a measure of the total binding energy (and, hence, the stability) of the nucleus. This binding energy is released during the formation of a nucleus from its constituent nucleons and would have to be supplied to the nucleus to decompose it into its individual nucleon components.

Binding energy and stability of the nucleus

A curve illustrating the average binding energy per nucleon as a function of the nuclear mass number is shown in Figure 25. The largest binding energy (highest stability) occurs near mass number 56—the mass region of the element iron. Figure 25 indicates that any nucleus heavier than mass number 56 would become a more stable system by breaking into lighter nuclei of higher binding energy with the difference in binding energy being released in the process. (It should be noted that nuclei lighter than mass number 56 can gain in stability by fusing to produce a heavier nucleus of greater mass defect—again, with the release of the energy equivalent of the mass difference. It is the fusion of the lightest nuclei that provides the energy released by the Sun and constitutes the basis of the hydrogen, or fusion, bomb. Efforts to harness fusion reaction for power production are being actively pursued. [See below *Nuclear fusion.*])

On the basis of energy considerations alone, Figure 25 would indicate that all matter should seek its most stable configuration, becoming nuclei of mass number near 56. However, this does not happen, because barriers to such a spontaneous conversion are provided by other factors. A good qualitative understanding of the nucleus is achieved by treating it as analogous to a uniformly charged liquid drop. The strong attractive nuclear force between pairs of nucleons is of short range and acts only between the closest neighbours. Since nucleons near the surface of the drop have fewer close neighbours than those in the interior, a surface tension is developed, and the nuclear drop assumes a spherical shape in order to minimize this surface energy. (The smallest surface area enclosing a given volume is provided by a sphere.) The protons in the nucleus exert a long-range, repulsive (Coulomb) force on each other due to their positive charge. As the number of nucleons in a nucleus increases beyond about 40, the number of protons must be diluted with an excess of neutrons to maintain relative stability.

If the nucleus is excited by some stimulus and begins to oscillate (*i.e.*, deform from its spherical shape), the surface forces will increase and tend to restore it to a sphere, where the surface tension is at a minimum. On the other hand, the Coulomb repulsion decreases as the drop deforms and the protons are positioned farther apart. These opposing tendencies set up a barrier in the potential energy of the system, as indicated in Figure 26.

Fission barrier

The curve in Figure 26 rises initially with elongation, since the strong, short-range nuclear force that gives rise to the surface tension increases. The Coulomb repulsion between protons decreases faster with elongation than the surface tension increases, and the two are in balance at point *B*, which represents the height of the barrier to fission. (This point is called the "saddle point" because, in a three-dimensional view of the potential energy surface, the shape of the pass over the barrier resembles a saddle.) Beyond point *B*, the Coulomb repulsion between the protons drives the nucleus into further elongation until at some point, *S* (the scission point), the nucleus breaks in two. Qualitatively, at least, the fission process is thus seen to be a consequence of the Coulomb repulsion between protons. Further discussion of the potential energy in fission is provided below.

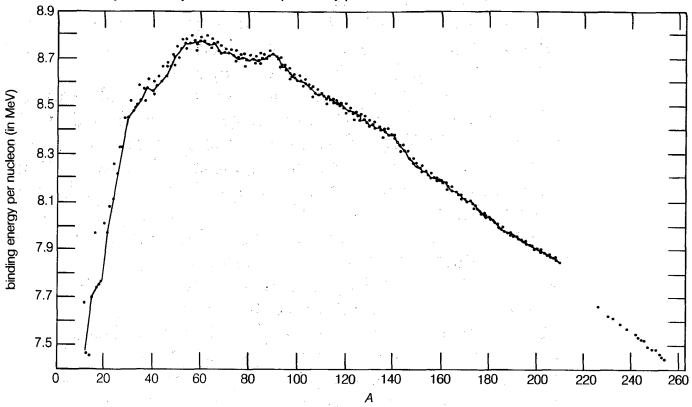


Figure 25: The average binding energy per nucleon as a function of the mass number, *A* (see text). The line connects the odd-*A* points.

Induced fission. The height and shape of the fission barrier are dependent on the particular nucleus being considered. Fission can be induced by exciting the nucleus to an energy equal to or greater than that of the barrier. This can be done by gamma-ray excitation (photofission) or through excitation of the nucleus by the capture of a neutron, proton, or other particle (particle-induced fission). The binding energy of a particular nucleon to a nucleus will depend on—in addition to the factors considered above—the odd–even character of the nucleus. Thus, if a neutron is added to a nucleus having an odd number of neutrons, an even number of neutrons will result, and the binding energy will be greater than for the addition of a neutron that makes the total number of neutrons odd. This "pairing energy" accounts in part for the difference in behaviour of nuclides in which fission can be induced with slow (low-energy) neutrons and those that require fast (higher-energy) neutrons. Although the heavy elements are unstable with respect to fission, the reaction takes place to an appreciable extent only if sufficient energy of acti-

Photofission and particle-induced fission

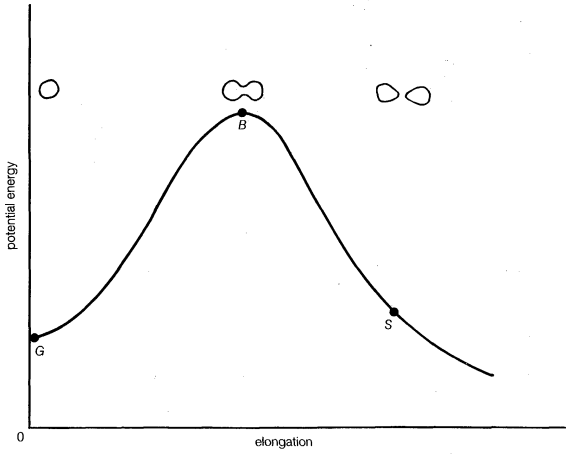


Figure 26: *The potential energy as a function of elongation of a fissioning nucleus.*
G is the ground state of the nucleus; *B* is the top of the barrier to fission (called the saddle point); and *S* is the scission point. The nuclear shape at these points is shown at the top.

vation is available to surmount the fission barrier. Most nuclei that are fissionable with slow neutrons contain an odd number of neutrons (*e.g.*, uranium-233, uranium-235, or plutonium-239), whereas most of those requiring fast neutrons (*e.g.*, thorium-232 or uranium-238) have an even number. The addition of a neutron in the former case liberates sufficient binding energy to induce fission. In the latter case, the binding energy is less and may be insufficient to surmount the barrier and induce fission. Additional energy must then be supplied in the form of the kinetic energy of the incident neutron. (In the case of thorium-232 or uranium-238, a neutron having about 1 MeV of kinetic energy is required.)

Spontaneous fission. The laws of quantum mechanics deal with the probability of a system such as a nucleus or atom being in any of its possible states or configurations at any given time. A fissionable system (uranium-238, for example) in its ground state (*i.e.*, at its lowest excitation energy and with an elongation small enough that it is confined inside the fission barrier) has a small but finite probability of being in the energetically favoured configuration of two fission fragments. In effect, when this occurs, the system has penetrated the barrier by the process of quantum mechanical tunneling. This process is called spontaneous fission because it does not involve any outside influences. In the case of uranium-238, the process has a very low probability, requiring more than 10^{15} years for half of the material to be transformed (its so-called half-life) by this reaction. On the other hand, the probability for spontaneous fission increases dramatically for the heaviest nuclides known and becomes the dominant mode of decay for some—those having half-lives of only fractions of a second. In fact, spontaneous fission becomes the limiting factor that may prevent the formation of still heavier (super-heavy) nuclei.

The stages of fission. A pictorial representation of the sequence of events in the fission of a heavy nucleus is given in Figure 27. In this figure, neutrons are shown in black and protons in white. The approximate time elapse between stages of the process is indicated at the bottom of the Figure.

The phenomenology of fission. When a heavy nucleus undergoes fission, a variety of fragment pairs may be formed, depending on the distribution of neutrons and protons between the fragments. This leads to probability distribution of both mass and nuclear charge for the fragments. The probability of formation of a particular fragment is called its fission yield and is expressed as the percentage of fissions leading to it.

'ission ragments

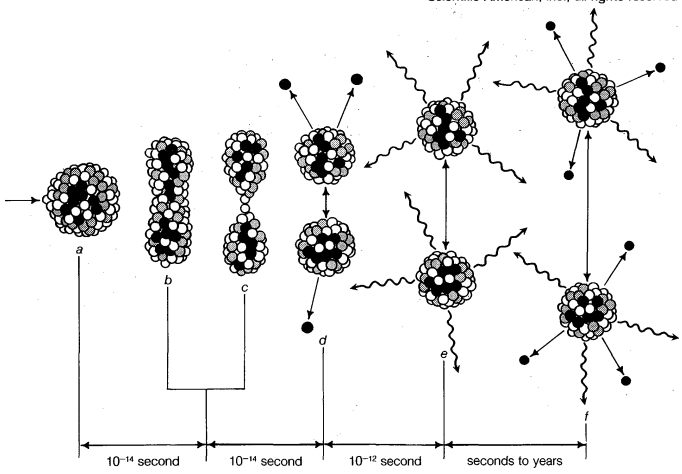


Figure 27: *Sequence of events in the fission of a uranium nucleus by a neutron.*
In *a*, the neutron strikes the nucleus and is absorbed, causing the nucleus to undergo deformation, *b*. In about 10^{-14} second, one of the deformations, *c*, is so drastic that the nucleus cannot recover and fissions, *d*, releasing two or three neutrons. In about 10^{-12} second, the fission fragments lose their kinetic energy and come to rest, emitting a number of gamma rays. At this stage, *e*, they are called fission products. In the final stage, *f*, the fission products lose their excess energy by radioactive decay, emitting beta particles and gamma rays over a time period ranging from seconds to years.

The separated fragments experience a large Coulomb repulsion due to their nuclear charges, and they recoil from each other with kinetic energies determined by the fragment charges and the distance between the charge centres at the time of scission. Variations in these parameters lead to a distribution of kinetic energies, even for the same mass split.

The initial velocities of the recoiling fragments are too fast for the outer (atomic) electrons of the fissioning atom to keep pace, and many of them are stripped away. Thus, the nuclear charge of the fragment is not fully neutralized by the atomic electrons, and the fission fragments fly apart as highly charged atoms. As the nucleus of the fragment adjusts from its deformed shape to a more stable configuration, the deformation energy (*i.e.*, the energy required to deform it) is recovered and converted into internal excitation energy, and neutrons and prompt gamma rays (an energetic form of electromagnetic radiation given off nearly coincident with the fission event) may be evaporated from the moving fragment. The fast-moving, highly charged atom collides with the atoms of the medium through which it is moving, and its kinetic energy is transferred to ionization and heating of the medium as it slows down and comes to rest. The range of fission fragments in air is only a few centimetres.

Fission products

During the slowing-down process, the charged atom picks up electrons from the medium and becomes neutral by the time it stops. At this stage in the sequence of events, the atom produced is called a fission product to distinguish it from the initial fission fragment formed at scission. Since a few neutrons may have been lost in the transition from fission fragment to fission product, the two may not have the same mass number. The fission product is still not a stable species but is radioactive, and it finally reaches stability by undergoing a series of beta decays, which may vary over a time scale of fractions of a second to many years. The beta emission consists of electrons and antineutrinos, often accompanied by gamma rays and X rays.

The distributions in mass, charge, and kinetic energy of the fragments have been found to be dependent on the fissioning species as well as on the excitation energy at which the fission act occurs. Many other aspects of fission have been observed, adding to the extensive phenomenology of the process and providing an intriguing set of problems for interpretation. These include the systematics of fission cross sections (a measure of the probability for fission to occur); the variation of the number of prompt neutrons (see below) emitted as a function of the fissioning species and the particular fragment mass split; the angular distribution of the fragments with respect to the direction of the beam of particles inducing fission; the systematics of spontaneous fission half-lives; the occurrence of spontaneous fission isomers (excited states of the nucleus); the emission of light particles (hydrogen-3, helium-3, helium-4, etc.) in small but significant numbers in some fission events; the presence of delayed neutron emitters among the fission products; the time scale on which the various stages of the process take place; and the distribution of the energy release in fission among the particles and radiations produced.

A detailed discussion of all of these facets of fission and how the data were obtained is not possible here, but a few of them are treated to provide some insight into this field of study and a taste of its fascination.

Distribution of fragment masses

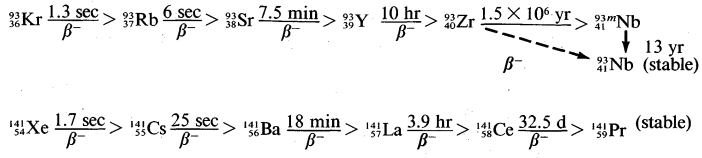
Fission fragment mass distributions. The distribution of the fragment masses formed in fission is one of the most striking features of the process. It is dependent on the mass of the fissioning nucleus and the excitation energy at which the fission occurs. At low excitation energy, the fission of such nuclides as uranium-235 or plutonium-239 is asymmetric—*i.e.*, the fragments are formed in a two-humped probability (or yield) distribution favouring an unequal division in mass. This is illustrated in Figure 28. As will be noted, the light group of fragment masses shifts to higher mass numbers as the mass of the fissioning nucleus increases, whereas the position of the heavy

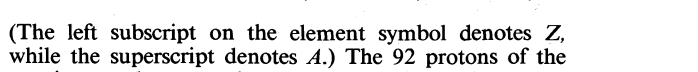
group remains nearly stationary. As the excitation energy of the fission increases, the probability for a symmetric mass split increases, while that for asymmetric division decreases. Thus, the valley between the two peaks increases in probability (yield of formation), and at high excitations the mass distribution becomes single-humped, with the maximum yield at symmetry (see Figure 29). Radium isotopes show interesting triple-humped mass distributions, and nuclides lighter than radium show a single-humped, symmetric mass distribution. (These nuclides, however, require a relatively high activation energy to undergo fission.) For very heavy nuclei in the region of fermium-260, the mass-yield curve becomes symmetric (single-humped) even for spontaneous fission, and the kinetic energies of the fragments are unusually high. An understanding of these mass distributions has been one of the major puzzles of fission, and a complete, theoretical interpretation is still lacking, albeit much progress has been made (see below).

Fission decay chains and charge distribution. In order to maintain stability, the neutron-to-proton (n/p) ratio in nuclei must increase with increasing proton number. The ratio remains at unity up to the element calcium, with 20 protons. It then gradually increases until it reaches a value of about 1.5 for the heaviest elements. When a heavy nucleus fissions, a few neutrons are emitted; however, this still leaves too high an n/p ratio in the fission fragments to be consistent with stability for them. They undergo radioactive decay and reach stability by successive conversions of neutrons to protons with the emission of a negative electron (called a beta particle, β^-) and an antineutrino. The mass number of the nucleus remains the same, but the nuclear charge (atomic number) increases by one, and a new element is formed for each such conversion. The successive beta decays constitute an isobaric, fission-product decay chain for each mass number. The half-lives for the decay of the radioactive species generally increase as they approach the stable isobar of the chain. (Species of the same element characterized by the same nuclear charge, Z [number of protons], but differing in their number of neutrons [and therefore in mass number A] are called isotopes. Species that have the same mass number, A, but differ in Z are known as isobars.)

Charge distribution in fission

For a typical mass split in the neutron-induced fission of uranium-235, the complementary fission-product masses of 93 and 141 may be formed following the emission of two neutrons from the initial fragments. The division of charge (*i.e.*, protons) between the fragments represents an important parameter in the fission process. Thus, for the mass numbers 93 and 141, the following isobaric fission-product decay chains would be formed (the half-lives for the beta-decay processes are indicated above the arrows):





(The left subscript on the element symbol denotes Z, while the superscript denotes A.) The 92 protons of the uranium nucleus must be conserved, and complementary fission-product pairs, such as krypton-36 with beryllium-56, rubidium-37 with cesium-55, or strontium-38 with xenon-54, would be possible.

The percentage of fissions in which a particular nuclide is formed as a primary fission product (*i.e.*, as the direct descendant of an initial fragment following its de-excitation) is called the independent yield of that product. The total yield for any nuclide in the isobaric decay chain is the sum of its independent yield and the independent yields of all of its precursors in the chain. The total yield for the entire chain is called the cumulative yield for that mass number.

Extensive radiochemical investigations have suggested that the most probable charge division is one that is displaced from stability about the same distance in both chains. This empirical observation is called the equal charge displacement (ECD) hypothesis, and it has been confirmed by several physical measurements. In the above example the ECD would predict the most probable charges at about rubidium-37 and cesium-55. A strong shell effect modifies the ECD expectations for fragments having 50 protons. The dispersion of the charge formation probability about the most probable charge (Z_p) is rather narrow and approximately Gaussian in shape and is nearly independent of the mass split as well as of the fissioning species. The most probable charge for an isobaric chain is a useful concept in the description of the charge dispersion, and it need not have an integral value. As the

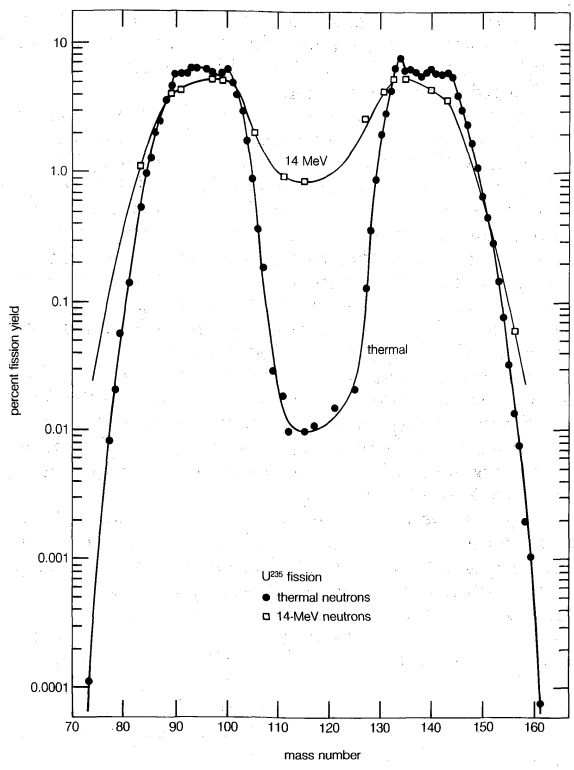


Figure 29: *Mass distribution dependence on the energy excitation in the fission of uranium-235.* At still higher energies, the curve becomes single-humped, with a maximum yield for symmetric mass splits (see text).

From A.C. Wahl, *Symposium on Physics and Chemistry of Fission* (1965); International Atomic Energy Agency, Vienna

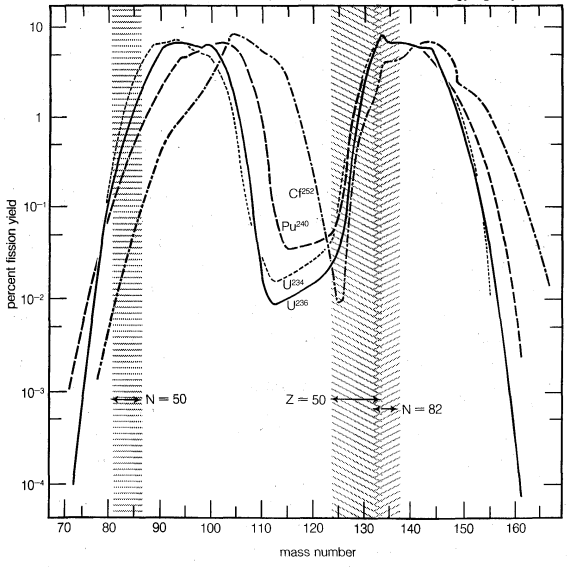


Figure 28: *Mass distributions (or fission-yield curves) for the thermal-neutron fission of uranium-233, uranium-235, and plutonium-239 and the spontaneous fission of californium-252.* The light mass group shifts to higher masses as the mass of the fissioning nucleus increases, while the heavy group remains nearly stationary. The shaded areas show the location of the closed shells of 50 protons, 50 neutrons, and 82 neutrons (see text).

energy of fission increases, the charge division tends toward maintaining the n/p ratio in the fragments the same as that in the fissioning nucleus. This is referred to as an unchanged charge distribution.

Prompt neutrons in fission. The average number of neutrons emitted per fission (represented by the symbol $\bar{\nu}$) varies with the fissioning nucleus. It is about 2.0 for the spontaneous fission of uranium-238 and 4.0 for that of fermium-257. In the thermal-neutron induced fission of uranium-235, $\bar{\nu} = 2.4$. The actual number of neutrons emitted, however, varies with each fission event, depending on the mass split. Although there is still controversy regarding the number of neutrons emitted at the instant of scission, it is generally agreed that most of the neutrons are given off by the recoiling fission fragments soon after scission occurs. The number of neutrons emitted from each fragment depends on the amount of energy the fragment possesses. The energy can be in the form of internal excitation (heat) energy or stored as energy of deformation of the fragment to be released when the fragment returns to its stable equilibrium shape.

Figure 30 shows the number of neutrons emitted per fragment as a function of the fragment mass number in the thermal neutron fission of uranium-235. The mass-yield distribution for the same isotope also is shown. This "sawtooth" neutron emission curve is typical of many fissioning systems at low excitation energy and provides another interesting phenomenon of fission. It is directly correlated with the fragment deformations at scission.

Delayed neutrons in fission. A few of the fission products have beta-decay energies that exceed the binding energy of a neutron in the daughter nucleus. This is likely to happen when the daughter nucleus contains one or two neutrons more than a closed shell of 50 or 82 neutrons, since these "extra" neutrons are more loosely bound. The beta decay of the precursor may take place to an excited state of the daughter from which a neutron is emitted. The neutron emission is "delayed" by the beta-decay half-life of the precursor. Six such delayed neutron emitters have been identified, with half-lives varying from about 0.5 to 56 seconds. The yield of the delayed neutrons is only about 1 percent of that of the prompt neutrons, but they are very important for the control of the chain reaction in a nuclear reactor.

Energy release in fission. The total energy release in a fission event may be calculated from the difference in the rest masses of the reactants (*e.g.*, $^{235}U + n$) and the final stable products (*e.g.*, $^{93}Nb + {}^{141}Pr + 2n$). The energy equivalent of this mass difference is given by the Einstein relation, $E = mc^2$. The total energy release depends on the mass split, but a typical fission event would have the total energy release distributed approximately as follows for the major components in the thermal neutron-induced fission of uranium-235:

Energy component	number per fission	total energy
Kinetic energy of fission fragments	2	170 MeV
Kinetic energy of prompt neutrons	2.5	5
Binding energy from capture of prompt neutrons	2.5	~12
Prompt gamma rays	8	8
		Total = 195 MeV

(The energy release from the capture of the prompt neutrons depends on how they are finally stopped, and some will escape the core of a nuclear reactor.)

This energy is released on a time scale of about 10^{-12} second and is called the prompt energy release. It is largely converted to heat within an operating reactor and is used for power generation. Also, there is a delayed release of energy from the radioactive decay of the fission products varying in half-life from fractions of a second to many years. The shorter-lived species decay in the reactor, and their energy adds to the heat generated; however, the longer-lived species remain radioactive and pose a problem in the handling and disposition of the reactor fuel elements when they need to be replaced. The antineutrinos that accompany the beta decay of the fission products are unreactive, and their kinetic energy (about 10 MeV per fission) is not recovered. Overall, about 200 MeV of energy per fission may be recovered for power applications.

The prompt energy release

Fission theory. Nuclear fission is a complex process that involves the rearrangement of hundreds of nucleons in a single nucleus to produce two separate nuclei. A complete theoretical understanding of this reaction would require a detailed knowledge of the forces involved in the motion of each of the nucleons through the process. Since such knowledge is still not available, it is necessary to construct simplified models of the actual system to simulate its behaviour and gain as accurate a description as possible of the steps in the process. The successes and failures of the models in accounting for the various observations of the fission process can provide new insights into the fundamental physics governing the behaviour of real nuclei, particularly at the large nuclear deformations encountered in a nucleus undergoing fission.

The framework for understanding nuclear reactions is analogous to that for chemical reactions and involves the concept of a potential-energy surface on which the reaction occurs. The driving force for physical or chemical reactions is the tendency to lower the potential energy and increase the stability of the system. Thus, for example, a stone at the top of a hill will roll down the hill, converting its potential energy at the top to kinetic energy of motion, and will come to rest at the bottom in a more stable state of lower potential energy. The potential energy is calculated as a function of various parameters of the system being studied. In the case of fission, the potential energy may be calculated as a function of the shape of the system as it proceeds over the barrier to the scission point, and the path of lowest potential energy may be determined.

As has been pointed out, an exact calculation of the nuclear potential energy is not yet possible, and it is to approximate this calculation that various models have been constructed to simulate the real system. Some of the models were developed to address aspects of nuclear structure and spectroscopy as well as features of nuclear reactions, and they also have been employed in attempts to understand the complexity of nuclear fission. The models are based on different assumptions and approximations of the nature of the nuclear forces and the dynamics of the path to scission. No one model can account for all of the extensive phenomenology of fission, but each addresses different aspects of the process and provides a foundation for further development toward a complete theory.

Nuclear models and nuclear fission. The nucleus exhibits some properties that reflect the collective motion of all its constituent nucleons as a unit, as well as other properties that are dependent on the motion and state of the individual nucleons.

The analogy of the nucleus to a drop of an incompressible liquid was first suggested by George Gamow in 1935

From J. Terrell, *Physical Review* 127, p. 892 (August 1962)

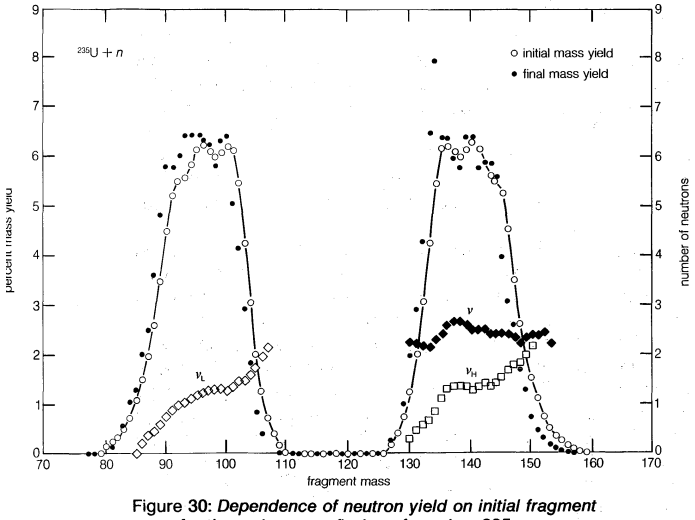


Figure 30: *Dependence of neutron yield on initial fragment mass for thermal-neutron fission of uranium-235.* Average number of neutrons emitted by light and heavy fragments are given the symbols ν_L and ν_H; the total from both fragments is ν. Also shown are the initial (fission fragment) and final (fission product) mass yields.

and later adapted to a description of nuclear reactions (by Niels Bohr [1936]; and Bohr and Fritz Kalckar [1937]) and to fission (Bohr and John A. Wheeler [1939]; and Yakov Frenkel [1939]). Bohr proposed the so-called compound nucleus description of nuclear reactions, in which the excitation energy of the system formed by the absorption of a neutron or photon, for example, is distributed among a large number of degrees of freedom of the system. This excited state persists for a long time relative to the periods of motion of nucleons across the nucleus and then decays by emission of radiation, the evaporation of neutrons or other particles, or by fission. The liquid-drop model of the nucleus accounts quite well for the general collective behaviour of nuclei and provides an understanding of the fission process on the basis of the competition between the cohesive nuclear force and the disruptive Coulomb repulsion between protons (as described above). It predicts, however, a symmetric division of mass in fission, whereas an asymmetric mass division is observed. Moreover, it does not provide an accurate description of fission barrier systematics or of the ground-state masses of nuclei. The liquid-drop model is particularly useful in describing the behaviour of highly excited nuclei, but it does not provide an accurate description for nuclei in their ground or low-lying excited states. Many versions of the liquid-drop model employing improved sets of parameters have been developed. However, investigators have found that mass asymmetry and certain other features in fission cannot be adequately described on the basis of the collective behaviour posited by such models alone.

Limitations of the liquid-drop model in describing certain fission features

A preference for the formation of unequal masses (*i.e.*, an asymmetric division) was observed early in fission research, and it has remained the most puzzling feature of the process to account for. Investigators have invoked various models other than that of the liquid drop in an attempt to address this question. Dealing with the mutual interaction of all the nucleons in a nucleus has been simplified by treating it as if it were equivalent to the interaction of one particle with an average, spherical static potential field that is generated by all the other nucleons. The methods of quantum mechanics provide the solution for the motion of a nucleon in such a potential. A characteristic set of energy levels for neutrons and protons is obtained, and, analogous to the set of levels of the electrons in an atom, the levels group themselves into shells at certain so-called magic numbers of nucleons. (For both neutrons and protons, these numbers are 2, 8, 20, 28, 50, 82, and 126.) Shell closures at these nuclear numbers are marked by especially strong binding, or extra stability. This constitutes the essence of the spherical-shell model (sometimes called the independent-particle, or single-particle, model), as developed by Maria Goeppert Mayer and J. Hans D. Jensen and their colleagues (1949). It accounts well for ground-state masses and spins, and for the existence of isomeric nuclear states (excited states having measurable half-lives) that occur when nuclear levels of widely differing spins lie relatively close to each other. The agreement with observations is excellent for spherical nuclei with nucleon numbers near the magic shell numbers. The spherical-shell model, however, does not agree well with the properties of nuclei that have other nucleon numbers—*e.g.*, the nuclei of the lanthanide and actinide elements, with nucleon numbers between the magic numbers.

Inadequacy of the shell model

In the lanthanide and actinide nuclei, the ground state is not spherical but rather deformed into a prolate spheroidal shape—that of a football or watermelon. For such nuclei, the allowed states of motion of a nucleon must be calculated in a potential having a symmetry corresponding to a spheroid rather than a sphere. This was first done by Aage Bohr, Ben R. Mottelson, and Sven G. Nilsson in 1955, and the level structure was calculated as a function of the deformation of the nucleus. A spheroid has three axes of symmetry, and it can rotate in space as a unit about any one of them. The rotation can occur independent of the internal state of excitation of the individual nucleons. Various modes of vibration of the spheroid also may take place. Since this deformed shell model has components of both the independent-particle motion and the collective motion of the nucleus as a whole (*i.e.*, rotations and vibrations), it is sometimes referred to as the unified model.

In Aage Bohr's application of the unified model to the fission process, the sequence of potential-energy surfaces for the excited states of the system are considered to be functions of a deformation parameter (*i.e.*, elongation) characterizing the motion toward fission and evaluated at the saddle point. As the system passes over the saddle point, most of its excitation energy is used up in deforming the nucleus, and the system remains "cold"—*i.e.*, it manifests little excitation, or heat, energy. Thus, only the low-lying excited states are available to the system. The spin and parity of the particular state (or channel) in which the system exists as it passes over the saddle point are then expected to determine the fission properties. In this channel (or transition-state) analysis of fission, a number of characteristics of the process are qualitatively accounted for. Hence, fission thresholds would depend on the spin and parity of the compound nuclear state, the fission fragment angular distribution would be governed by the collective rotational angular momentum of the state, and asymmetry in the mass distribution would result from passage over the barrier in a state of negative parity (which does not possess reflection symmetry). This model gives a good qualitative interpretation of many fission phenomena, but it must assume that at least some of the properties of the transition state at the saddle point are not altered by dynamical considerations in the descent of the system to the scission point. It is the only model that provides a satisfactory interpretation of the angular distributions of fission fragments, and it has attractive features that must be included in any complete theory of fission.

Application of the unified model to the fission process

The first application of the spherical-shell model to fission was the recognition that the positions of the peaks in the fission mass distribution correlated fairly well with the magic numbers and suggested a qualitative interpretation of the asymmetric mass division. Thus, a preference for the formation of nuclei with neutron numbers close to 82 would favour the formation of nuclides near the peak in the heavy group and would thus determine the mass split for the fissioning system (see Figure 28). Some extra stability for nuclear configurations of 50 protons would also be expected, but this is not particularly evident. In fact, the so-called doubly magic nucleus tin-132, with 50 protons and 82 neutrons, has a rather low yield in low-energy fission.

A more quantitative application of the spherical-shell model to fission was undertaken by Peter Fong in the United States in 1956. He related the probability of formation of a given pair of fragments to the available density of states for that pair of fragments at the scission point in a statistical-model approach. A model of this sort predicts that the system, in its random motions, will experience all possible configurations and so will have a greater probability of being in the region where the greatest number of such configurations (or states) is concentrated. The model assumes that the potential energy at the saddle point is essentially all converted to excitation energy and that a statistical equilibrium among all possible states is established at the scission point. The extra binding energy for closed-shell nuclei leads to a higher density of states at a given excitation energy than is present for other nuclei and, hence, leads to a higher probability of formation. An asymmetric mass distribution in good agreement with that observed for the neutron-induced fission of uranium-235 is obtained. Moreover, the changes in the mass distribution with an increased excitation energy of fission (*e.g.*, an increase in the probability of symmetric fission relative to asymmetric fission) are accounted for by the decrease in importance of the shell effects as the excitation energy increases. Other features of the fission process also are qualitatively explained; however, extensive changes in the parameters of the model are required to obtain agreement with experiments for other fissionable nuclides. Then, too, there are fundamental problems concerning the validity of some of the basic assumptions of the model.

Statistical-model approach

The fundamental question as to the validity of models that evaluate the properties of the system at the scission point (the so-called scission-point models of fission) is whether the system remains long enough at this point

cission-
oint
iodels

on the steep decline of the potential-energy surface for a quasi-equilibrium condition to be established. There is some evidence that such a condition may indeed prevail, but it is not clearly established. Nonetheless, such models have proved quite useful in interpreting observations of mass, charge, and kinetic energy distributions, as well as of neutron emission dependence on fragment mass. It seems very likely that the fragment shell structure plays a significant role in determining the course of the fission process.

Although the single-particle models provide a good description of various aspects of nuclear structure, they are not successful in accounting for the energy of deformation of nuclei (*i.e.*, surface energy), particularly at the large deformations encountered in the fission process. A major breakthrough occurred when a hybrid model incorporating shell effects as a correction to the potential energy of the liquid-drop model was proposed by the Soviet physicist V.M. Strutinskii in 1967. This approach retains the dominant collective surface and Coulomb effects while adding shell and pairing corrections that depend on deformation. Shell corrections of several million electron volts are calculated, and these can have a significant effect on a liquid-drop barrier of about 5 MeV. The nucleon numbers at which the shells appear depend on the deformation and may differ from the spherical model magic numbers. In the vicinity of the fission barrier, the shells introduce structure in the liquid-drop potential-energy curve, as illustrated in Figure 31. The relative heights and widths of the two peaks vary with the mass and charge of the fissioning system.

trutin-
kii's
ybrid
iodel

The double-humped barrier (Figure 31) provides a satisfactory explanation for a number of puzzling observations in fission. The existence of short-lived, spontaneous fission isomers, for example, is understood as the consequence of the population of states in the second well (class II). These isomers have a much smaller barrier to penetrate and so exhibit a much shorter spontaneous fission half-life. The change in shape associated with these states, as compared to class I states, also hinders a rapid return to the ground state by gamma emission. (Class II states are also called shape isomers.) The systematics of neutron-induced fission cross sections and structure in some fission-fragment angular distributions also find an interpretation in the implications of the double-humped barrier.

From R. Vandenbosch and J.F. Huizenga, *Nuclear Fission* (1973); Academic Press

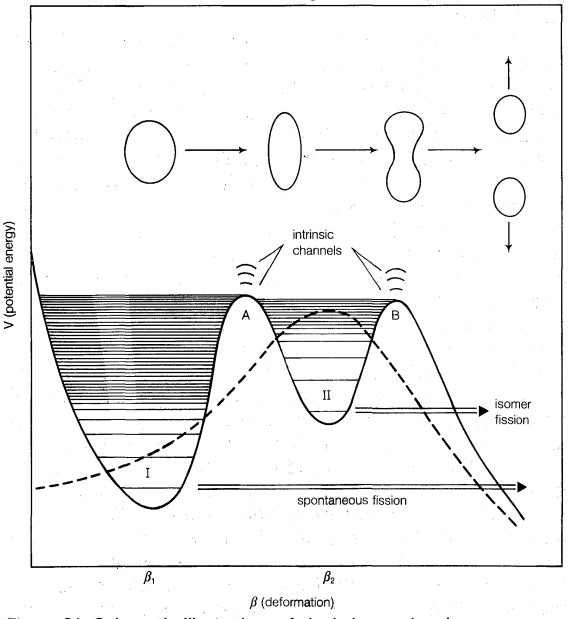


Figure 31: *Schematic illustrations of single-humped and double-humped fission barriers.*
The former are represented by the dashed line and the latter by the continuous line. Intrinsic excitations in the first and second wells at deformations β_1 and β_2 are designated class I and class II states, respectively. Intrinsic channels at the two barriers also are illustrated. The transition in the shape of the nucleus as a function of deformation is schematically represented in the upper part of the figure. Spontaneous fission of the ground state and isomeric state occurs from the lowest energy class I and class II states, respectively.

The Strutinskii procedure provided a strong stimulus for calculations of the potential-energy surfaces appropriate to fissioning systems, since it provided a consistent and useful prescription for treating both the macroscopic (liquid-drop) and microscopic (single-particle) effects in deformed nuclei. Many calculations of the potential-energy surface employing different model potentials and parameters have been carried out as functions of the shapes of the system. The work of the American nuclear physicists W.J. Swiatecki, James R. Nix, and their collaborators has been particularly noteworthy in such studies, which also include some attempts to treat the dynamical evolution of the fission process.

Calculations for the actinide elements indicate that, at deformations corresponding to the second barrier (Figure 31), the potential energy for asymmetric mass splits is lower than that for symmetric ones; hence, the former are favoured at that stage of the process. For larger deformations, however, a single potential does not represent the incipient formation of two fragments very well. In fact, a discontinuity occurs at the scission point, and the results of the calculation depend on whether the scission configuration is treated as one nucleus or as two separate nuclei.

Two-centre potential

A two-centre potential may also be used to represent the nature of the forces at work in a fissioning nucleus. In such a model, the potential energy surfaces are represented by two overlapping spheres or spheroids. It is equivalent to a one-centre potential when there is a complete overlap at small deformations, and it has the correct asymptotic behaviour as the nascent fragments separate. This approach indicates a preformation of the final shell structure of the fragments early in the process.

Although the validity of the assumptions inherent in scission-point models may be in question, the results obtained with them are in excellent agreement with observation. Representative of such a model is the Argonne Scission-Point model, which uses a macroscopic-microscopic calculation with deformed fragment shell and pairing corrections to determine the potential energy of a system of two nearly touching spheroids and which includes their interaction in terms of a neck connecting them. Models of this kind provide a simple approach to a highly quantitative and detailed study of the dependence of the probability of formation of a given fragment pair on the neutron and proton number and on the deformation in each fragment. They account very well for the mass, charge, and kinetic-energy distributions and the neutron-emission dependence on mass number for a broad range of fissioning nuclei. The scission-point models, however, do not address questions of fission probability or the angular distributions of the fragments. As the fission-excitation energy increases, the shell correction diminishes and the macroscopic (liquid-drop) behaviour dominates.

Nuclides in the region of fermium-264 have been observed to undergo symmetric fission with unusually high fragment kinetic energies. This appears to be the consequence of the stability for the magic number configurations of 50 protons and 82 neutrons. The formation of two doubly magic fragments of tin-132 is strongly favoured energetically, whereas the formation of only one such fragment in the low-energy fission of uranium or plutonium isotopes is not. The fragments of tin-132 are spherical rather than deformed, and a more compact configuration at the scission point (with the charge centres closer together) leads to higher fragment kinetic energies.

It is evident that shell effects, both in the fissioning system at the saddle point and in the deformed fragments near the scission point, are important in interpreting many of the features of the fission process. The stage of the process at which the various fragment distributions are determined is, however, not clearly established. All the components of a reasonable understanding of fission seem to be at hand, but they have yet to be synthesized into a complete, dynamic theory.

Considerations of the dynamics of the descent of the system on the potential-energy surface from the saddle point to the scission point involve two extreme points of

Dynamics of fission

view. An "adiabatic" approximation may be valid if the collective motion of the system is considered to be so slow or the coupling between the collective and internal single-particle degrees of freedom (*i.e.*, between macroscopic and microscopic behaviour) so weak that the fast single-particle motions can readily adjust to the changes in shape of the fissioning nucleus as it progresses toward scission. In this case, the changes in the system take place without the gain or loss of heat energy. The decrease in potential energy between the saddle and scission points will then appear primarily in the collective degrees of freedom at scission and be associated with the kinetic energy of the relative motion of the nascent fragments (referred to as pre-scission kinetic energy). On the other hand, if the collective motion toward scission is relatively fast or the coupling-to-particle motion stronger, collective energy can be transformed into internal excitation (heat) energy of the nucleons. (This is analogous to heating in the motion of a viscous fluid.) In such a "non-adiabatic" process the mixing among the single-particle degrees of freedom may be sufficiently complete that a statistical model may be applicable at the scission point. Either extreme represents an approximation of complex behaviour, and some experimental evidence in support of either interpretation may be advanced. As in most such instances in nature, the truth probably lies somewhere between the extremes, with both playing some role in the fission process.

Fission chain reactions and their control. The emission of several neutrons in the fission process leads to the possibility of a chain reaction if at least one of the fission neutrons induces fission in another fissile nucleus, which in turn fissions and emits neutrons to continue the chain. If more than one neutron is effective in inducing fission in other nuclei, the chain multiplies more rapidly. The condition for a chain reaction is usually expressed in terms of a multiplication factor, k, which is defined as the ratio of the number of fissions produced in one step (or neutron generation) in the chain to the number of fissions in the preceding generation. If k is less than unity, a chain reaction cannot be sustained. If $k = 1$, a steady-state chain reaction can be maintained; and if k is greater than 1, the number of fissions increases at each step, resulting in a divergent chain reaction. The term critical assembly is applied to a configuration of fissionable material for which $k = 1$; if $k > 1$, the assembly is said to be supercritical. A critical assembly might consist of the fissile material in the form of a metal or oxide, a moderator to slow the fission neutrons, and a reflector to scatter neutrons that would otherwise be lost back into the assembly core.

Critical assembly

In a fission bomb it is desirable to have k as large as possible and the time between steps in the chain as short as possible so that many fissions occur and a large amount of energy is generated within a brief period ($\sim 10^{-7}$ second) to produce a devastating explosion. If one kilogram of uranium-235 were to fission, the energy released would be equivalent to the explosion of 20,000 tons of the chemical explosive trinitrotoluene (TNT). In a controlled nuclear reactor, k is kept equal to unity for steady-state operation. A practical reactor, however, must be designed with k somewhat greater than unity. This permits power levels to be increased if desired, as well as allowing for the following: the gradual loss of fuel by the fission process; the buildup of "poisons" among the fission products being formed that absorb neutrons and lower the k value; and the use of some of the neutrons produced for research studies or the preparation of radioactive species for various applications (see below). The value of k is controlled during the operation of a reactor by the positioning of movable rods made of a material that readily absorbs neutrons (*i.e.*, one with a high neutron-capture cross section), such as boron, cadmium, or hafnium. The delayed-neutron emitters among the fission products increase the time between successive neutron generations in the chain reaction and make the control of the reaction easier to accomplish by the mechanical movement of the control rods.

Fission reactors

Fission reactors can be classified by the energy of the neutrons that propagate the chain reaction. The most common type, called a thermal reactor, operates with thermal neutrons (those having the same energy distribution as gas molecules at ordinary room temperatures). In such a reactor the fission neutrons produced (with an average kinetic energy of more than 1 MeV) must be slowed down to thermal energy by scattering from a moderator, usually consisting of ordinary water, heavy water (D_2O), or graphite. In another type termed an intermediate reactor the chain reaction is maintained by neutrons of intermediate energy, and a beryllium moderator may be used. In a fast reactor fast fission neutrons maintain the chain reaction, and no moderator is needed. All of the reactor types require a coolant to remove the heat generated; water, a gas, or a liquid metal may be used for this purpose, depending on the design needs. For details about reactor types, see ENERGY CONVERSION: *Nuclear reactors.*

Uses of fission reactors and fission products. A nuclear reactor is essentially a furnace used to produce steam or hot gases that can provide heat directly or drive turbines to generate electricity. Nuclear reactors are employed for commercial electric-power generation throughout much of the world and as a power source for propelling submarines and certain kinds of surface vessels. Another important use for reactors is the utilization of their high neutron fluxes for studying the structure and properties of materials and for producing a broad range of radionuclides, which, along with a number of fission products, have found many different applications. Heat generated by radioactive decay can be converted into electricity through the thermoelectric effect in semiconductor materials and thereby produce what is termed an atomic battery. When powered by either a long-lived, beta-emitting fission product (*e.g.*, strontium-90, calcium-144, or promethium-147) or one that emits alpha particles (plutonium-238 or curium-244), these batteries are a particularly useful source of energy for cardiac pacemakers and for instruments employed in remote, unmanned facilities, such as those in outer space, the polar regions of the Earth, or the open seas. There are many practical uses for other radionuclides, as discussed above in *Applications of radioactivity.* (E.P.S.)

Atomic batteries

NUCLEAR FUSION

Nuclear fusion is the process by which nuclear reactions between light elements form heavier elements (up to iron) and, in most cases, result in the release of energy. Such reactions constitute the fundamental energy source of stars, including the Sun. Stellar evolution can be viewed as the passing of a star through various stages as thermonuclear reactions and nucleosynthesis cause compositional changes over long time periods. Hydrogen "burning" initiates the fusion energy source of stars and leads to the formation of helium. Generation of fusion energy for practical use also relies on fusion reactions between the lightest elements that burn to form helium. In fact, the heavy isotopes of hydrogen, deuterium ($^2_1$H, or D) and tritium ($^3_1$H, or T), react more efficiently with each other and yield more energy per reaction than do two hydrogen nuclei (protons) when they undergo fusion. The deuterium nucleus has one proton and one neutron, while that of tritium consists of a proton bound together with two neutrons.

Energy source of stars

Fusion reactions between light elements, like fission reactions that split heavy elements, release energy due to a key feature of nuclear matter. A parameter called the binding energy of the nucleus is a measure of the efficiency with which its constituent nucleons are bound together. Take, for example, an element with Z protons and N neutrons in its nucleus. The element's atomic mass number or atomic weight, A, is $Z + N$ and its atomic number is Z. The binding energy is the energy associated with the mass difference between the Z protons and N neutrons considered separately and the nucleons bound together ($Z + N$) in a nucleus of mass, M. The formula is

$$B = [Zm_p + Nm_n - M(Z,N)]c^2, \qquad (10)$$

where m_p and m_n are the proton and neutron masses, and c is the speed of light. It has been determined experimentally that the binding energy per nucleon is a maximum of about 8.8 MeV at an atomic mass number of approximately 60. Accordingly, the fusion of lighter elements or the splitting of heavier ones generally leads to a net release of energy.

History of fusion research and technology. The fusion process has been studied as part of nuclear physics for much of the 20th century. In the late 1930s the German-born physicist Hans A. Bethe first recognized that the fusion of hydrogen nuclei to form deuterium is exoergic (*i.e.,* there is a net release of energy) and, together with subsequent reactions, accounts for the energy source in stars. Work proceeded over the next two decades, motivated by the need to understand nuclear matter and forces, to learn more about the nuclear physics of stellar objects, and to develop thermonuclear weapons (the so-called hydrogen bomb) and predict their performance. During the late 1940s and early 1950s, research programs in the United States, United Kingdom, and Soviet Union began to yield a better understanding of nuclear fusion, and investigators embarked on ways of exploiting the process for practical energy production. This work focused on the use of magnetic fields and electromagnetic forces to contain extremely hot gases called plasmas. A plasma consists of unbound electrons and positive ions whose motion is dominated by electromagnetic interactions. It is the only state of matter in which thermonuclear reactions can occur in a self-sustaining manner. Astrophysics and magnetic fusion research, among other fields, require extensive knowledge of how gases behave in the plasma state.

Research on plasma containment

The inadequacy of the then-existent knowledge became clearly apparent in the 1950s as the behaviour of plasma in many of the early magnetic confinement systems proved too complex to understand. Moreover, researchers found that confining fusion plasma in a "magnetic trap" was far more challenging than they had anticipated. Plasma must be heated to tens of millions of degrees kelvin or higher to induce and sustain the thermonuclear reaction required to produce usable amounts of energy. At temperatures this high, the nuclei in the plasma move rapidly enough to overcome their mutual repulsion and fuse. It is exceedingly difficult to contain plasmas at such a temperature level because the hot gases tend to expand and escape from the enclosing structure (see below).

The work of the major American, British, and Soviet fusion programs was strictly classified until 1958. That year, research objectives were made public, and many of the topics being studied were found to be similar, as were the problems encountered. Since that time, investigators have continued to study and measure fusion reactions between the lighter elements and have arrived at more accurate determinations of reaction rates. Also, the formulas developed by nuclear physicists for predicting the rate of fusion-energy generation have been adopted by astrophysicists to derive new information about the structure of the stellar interior and about the evolution of stars.

The late 1960s witnessed a major advance in efforts to harness fusion reactions for practical energy production: the Soviets announced the achievement of high plasma temperature (about 3,000,000 K), along with other physical parameters, in a tokamak, a toroidal magnetic confinement system in which the plasma is kept generally stable both by an externally generated, doughnut-shaped magnetic field and by electric currents flowing within the plasma itself. (The basic concept of the tokamak had been first proposed by Andrey D. Sakharov and Igor Y. Tamm about 1950.) Since its development, the tokamak has been the focus of most research, though other approaches have been pursued as well. Employing the tokamak concept, physicists have attained conditions in plasmas that approach those required for practical fusion-power generation (see below).

The tokamak

Work on another major approach to fusion energy, called inertial confinement fusion (ICF), has been carried on since the early 1960s. Initial efforts were undertaken in 1961 with a then-classified proposal that large pulses of laser energy could be used to implode and shock-heat matter to temperatures at which nuclear fusion would be vigorous. Aspects of inertial confinement fusion were declassified in the 1970s, but a key element of the work—specifically the design of targets containing pellets of fusion fuels—still is largely secret. Very painstaking work to design and develop suitable targets continues today. At the same time, significant progress has been made in developing high-energy, short-pulse drivers with which to implode millimetre-radius targets. The drivers include both high-power lasers and particle accelerators capable of producing beams of high-energy electrons or ions. Lasers that produce more than 100,000 joules in pulses on the order of one nanosecond (10^{-9} second) have been developed, and the power available in short bursts exceeds 10^{14} watts. Best estimates are that practical inertial confinement for fusion energy will require either laser or particle-beam drivers with an energy of 5,000,000 to 10,000,000 joules capable of delivering more than 10^{14} watts of power to a small target of deuterium and tritium (see below).

Research on inertial confinement fusion

Types of fusion reactions. Fusion reactions are of two basic types: (1) those that preserve the number of protons and neutrons, and (2) those that involve a conversion between protons and neutrons. Reactions of the first type are most important for practical fusion-energy production, whereas those of the second type are crucial to the initiation of star burning. An arbitrary element is indicated by the notation ${}^{A}_{Z}\mathrm{X}$, where Z is the charge of the nucleus and A is the atomic weight. An important fusion reaction for practical energy generation is that between deuterium and tritium. It produces helium (${}^{4}_{2}\mathrm{He}$) and a neutron (${}^{1}_{0}n$); and is written

$$\mathrm{D} + \mathrm{T} \rightarrow \mathrm{He} + n. \tag{11}$$

To the left of the arrow (before the reaction), there are two protons and three neutrons. The same is true on the right. Another reaction, that which initiates star burning, is the fusion of two hydrogen nuclei to form deuterium:

$$\mathrm{H} + \mathrm{H} \rightarrow \mathrm{D} + \beta^{+} + \nu, \tag{12}$$

where β^{+} represents a positron and ν stands for a neutrino. Before the reaction, there are two hydrogen nuclei. Afterward, there is one proton and one neutron, bound together as the nucleus of deuterium, plus a positron and a neutrino produced as a consequence of the conversion of one proton to a neutron. Both fusion reactions (11) and (12) are exoergic and so yield energy. Bethe proposed that fusion reaction (12) could occur with a net release of energy and provide, along with subsequent reactions, the fundamental energy source sustaining the stars. Practical energy generation requires reaction (11) rather than (12) for two reasons: first, the rate of reactions between deuterium and tritium is much faster than that between protons; and, second, the net energy release from reaction (11) is 40 times greater than that from reaction (12).

Energy released in fusion reactions. Energy is released in a nuclear reaction if the total mass of the resultant particles is less than the mass of the initial reactants. To illustrate, suppose two nuclei, labeled X and a, react to form two other nuclei, Y and b, denoted

$$\mathrm{X} + \mathrm{a} \rightarrow \mathrm{b} + \mathrm{Y}, \tag{13}$$

or X(a,b)Y. The particles a and b are often nucleons, either protons or neutrons, but in general can be any nuclei. Assuming that none of the particles are internally excited (*i.e.,* are in their ground state), the energy quantity called the Q-value for this reaction is defined as

$$Q = (m_{\mathrm{x}} + m_{\mathrm{a}} - m_{\mathrm{b}} - m_{\mathrm{y}})c^{2}, \tag{14}$$

where the m-letters refer to the mass of each particle and c is the speed of light. If the energy value Q is positive, the reaction is exoergic. If Q is negative, the reaction is endoergic (*i.e.,* absorbs energy). When the total proton number and total neutron number are each preserved before and after the reaction (such as in reaction 11), then the Q-value can be expressed in terms of the binding energy, B, of each particle as

$$Q = B_{\mathrm{y}} + B_{\mathrm{b}} - B_{\mathrm{x}} - B_{\mathrm{a}}. \tag{15}$$

Fusion reaction (11) between deuterium and tritium has a positive Q-value of 17.58 MeV, which is equivalent to 2.8×10^{-12} joule. Fusion reaction (12) also is exoergic, with a Q-value of 0.420 MeV, or 6.7×10^{-14} joule. To develop a sense for these figures, one might consider that one metric ton (2,205 pounds) of deuterium atoms is equal to 3×10^{32} atoms. If one ton of deuterium is consumed through burning via fusion reaction (11) with

tritium, the energy released would be 8.4×10^{20} joules. This can be compared to the energy content of one ton of coal—namely, 2.9×10^{10} joules. In other words, one ton of deuterium has the energy equivalent of approximately 29,000,000,000 tons of coal.

Rate and yield of fusion reactions. The energy yield of a reaction between nuclei and the rate of such reactions are equally important. These quantities have a profound influence in both nuclear astrophysics and practical fusion-energy considerations. When a particle of one type passes through a collection of particles of the same or different type, there is a measurable chance that the particles will interact. The particles may interact in many ways, such as simply scattering, which means that they change direction and exchange energy. Or they may undergo a nuclear fusion reaction such as (11) or (12). The measure of the likelihood that particles will interact is called the cross section, σ. The cross section depends on the type of interaction and the state and energy of the particles. When σ is multiplied by the atomic density of target particles, N, the resultant product, $N\sigma$, is denoted Σ, the macroscopic cross section. The inverse of Σ is the mean distance an incident particle will travel before interacting with a target particle and is called the mean free path. The rate of interactions of a given type is ($vN\sigma$), where v is the speed of the incident particle. Cross sections are measured by producing a beam of one particle at a given energy, allowing the beam to interact with a (usually thin) target made of either the same or a different material, and measuring deflections or reaction products. In this way it is possible to determine the relative likelihood of one type of fusion reaction versus another, as well as the optimal conditions for a particular reaction.

The cross sections of fusion reactions can be measured experimentally or calculated theoretically and have been determined for many reactions over a wide range of particle energies. They are well known for practical fusion-energy applications and are reasonably well known, though with significant gaps, for stellar evolution. Fusion reactions between nuclei, each with a positive charge of one or more, are the most important for both practical applications and the nucleosynthesis of the light elements in the burning stages of stars. Yet, it is well known that two positively charged nuclei electrostatically repel each other—*i.e.*, they experience a repulsive force inversely proportional to the square of the distance separating them. This repulsion is called the Coulomb barrier. It is highly unlikely that two positive nuclei will approach each other closely enough to undergo a fusion reaction unless they have sufficient energy to overcome the Coulomb barrier. As a result, the cross section for fusion reactions between charged particles is very small unless the energy of the particles is high, at least 10^4 eV and often more than 10^5 or 10^6 eV. This explains why the centre of a star is hot and why fuel material for practical fusion energy must be heated to 50,000,000 K or more. Only then will a reasonable fusion reaction rate and power output be achieved.

Significance of the Coulomb barrier

The phenomenon of the Coulomb barrier also explains a fundamental difference between energy generation by nuclear fusion and nuclear fission. While fission of heavy elements can be induced by either protons or neutrons, generation of fission energy for practical applications is dependent on neutrons to induce fission reactions in uranium or plutonium. Having no electric charge, the neutron is free to enter the nucleus even if its energy corresponds to room temperature. Fusion energy, relying as it does on the fusion reaction between light nuclei, occurs only when the particles are sufficiently energetic to overcome the Coulomb repulsive force. This requires the production and heating of the gaseous reactants to the high temperature state known as the plasma state.

Plasma state. Gas containing unbound electrons and an equal amount of positive charge constitutes plasma, which is commonly considered the fourth state of matter. Very high temperature plasmas are fully ionized gases, which means that the ratio of neutral gas atoms to charged particles is extremely small. For example, the ionization energy of hydrogen is 13.6 eV, while the average energy of a hydrogen ion in a plasma at 50,000,000 K is 6,462 eV. One can reasonably expect that essentially all of the hydrogen in this plasma is ionized.

Fully ionized gases

A reaction-rate parameter more appropriate to the plasma state is obtained by accounting for the fact that the particles in a plasma, as in any gas, have a distribution of energies. That is to say, not all particles have the same energy. In simple plasmas, this energy distribution is called the Maxwell–Boltzmann distribution, and the temperature of the gas or plasma is defined to be two-thirds of the average particle energy—*i.e.*, the relationship between the average energy, denoted $\bar{E}$, and temperature is $\bar{E} = (3/2)kT$. The quantity k is the Boltzmann constant, 8.62×10^{-5} eV per kelvin. The intensity of nuclear fusion reactions in a plasma is derived by averaging the quantity, $v\sigma$, over a distribution of speeds corresponding to a Maxwell–Boltzmann distribution. The cross section for the reaction depends on the energy or speed of the particles. The averaging process yields a function for a given reaction that depends only on the temperature and can be denoted $f(T)$. The rate of energy released (*i.e.*, the power released) in a reaction between two species, a and b, is

$$P_{ab} = n_a n_b f_{ab}(T) U_{ab}, \tag{16}$$

where n_a and n_b are the density of species a and b in the plasma, respectively, and U_{ab} is the energy released each time a and b undergo a fusion reaction. The parameter P_{ab} properly takes into account both the rate of a given reaction and the energy yield per reaction (see Figure 32).

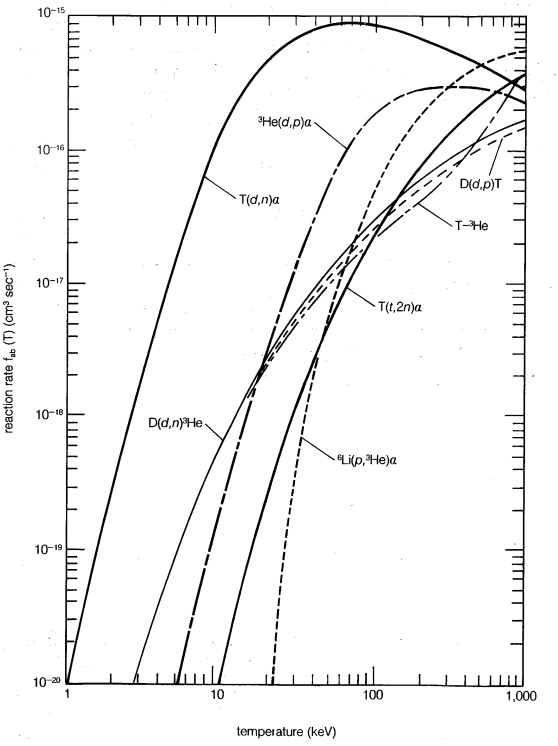


Figure 32: *The reaction rate as a function of plasma temperature, expressed in kiloelectron volts (keV).*
One keV is equivalent to a temperature of 11,000,000 K. The rate of reaction between deuterium and tritium is seen to be higher than all others and is very substantial, even at temperatures in the 5- to 10-keV range (see text).

Fusion reactions in stars. Fusion reactions are the primary energy source of stars and the mechanism for the nucleosynthesis of the light elements. The synthesis of helium from hydrogen is the main source of energy emitted by normal stars, such as the Sun, where the burning-core plasma has a temperature of less than 15,000,000 K. However, because the gas from which the star is formed often contains some heavier elements, notably carbon and nitrogen, it is important to include nuclear reactions between protons and these nuclei. The reaction chain between protons that ultimately leads to helium is the proton-proton (*pp*) chain. When protons also induce the burning of carbon and nitrogen, the CN cycle must be considered; and

Mechanism for nucleosynthesis

when oxygen is included, still another alternative scheme, the CNO bi-cycle must be accounted for.

The proton-proton nuclear fusion cycle in a star containing only hydrogen begins with the reaction

$$H + H \rightarrow D + \beta^+ + \nu; \qquad Q = 1.44 \text{ MeV}, \qquad (17)$$

where the Q-value assumes annihilation of the positron by an electron. The deuterium could react with other deuterium nuclei, but because there is so much hydrogen, the D/H ratio is held to very low values, typically 10^{-18}. Thus the next step is

$$H + D \rightarrow {}^3He + \gamma; \qquad Q = 5.49 \text{ MeV}, \qquad (18)$$

where γ indicates that gamma rays carry off some of the energy yield. The burning of the helium-3 isotope then gives rise to ordinary helium and hydrogen via the last step in the chain:

$${}^3He + {}^3He \rightarrow {}^4He + 2(H); \qquad Q = 12.86 \text{ MeV}. \qquad (19)$$

At equilibrium, helium-3 burns predominantly by reactions with itself because its reaction rate with hydrogen is small, while burning with deuterium is negligible due to the very low deuterium concentration. Once helium-4 builds up, reactions with helium-3 can lead to the production of still heavier elements, including beryllium-7, beryllium-8, lithium-7, and boron-8, if the temperature is greater than about 10,000,000 K.

The stages of stellar evolution are the result of compositional changes over very long periods. The size of a star, on the other hand, is determined by a balance between the pressure exerted by the hot plasma and the gravity force of the star's mass. The energy of the burning core is transported toward the surface of the star, where it is radiated at an effective temperature. The effective temperature of the Sun's surface is about 6,000 K, and significant amounts of radiation in the visible and infrared wavelength range are emitted.

Fusion reactions for controlled power generation. Reaction (11) between deuterium and tritium is one of the most important fusion reactions for controlled power generation, because the cross section for its occurrence is high, the practical plasma temperature required for the net energy release is moderate, and the energy yield of the reaction is high—17.58 MeV. Important reactions of this type involving deuterium include

$$D + T \rightarrow {}^4He + n; \qquad Q = 17.58 \text{ MeV}, \qquad (20)$$

$$D + D \xrightarrow{50\%} {}^3He + n; \qquad Q = 3.32 \text{ MeV} \qquad (21a)$$

$$D + D \xrightarrow{50\%} H + T; \qquad Q = 4.04 \text{ MeV}, \qquad (21b)$$

$$D + {}^3He \rightarrow H + {}^4He; \qquad Q = 18.34 \text{ MeV}. \qquad (22)$$

It should be noted that any plasma containing deuterium automatically produces some tritium and helium-3 from reactions of deuterium with other deuterium reactions. Significant too is the fact that reactions (11) and (21a) both produce a neutron, whereas reactions (21b) and (22) yield only charged nuclei as reactant products. The practical application of reactions (21a and 21b) or (22) would require plasmas that are three to 10 times hotter than a plasma burning deuterium and tritium, and the D-T plasma would burn 10 to 100 times as fast. Other fusion reactions involving elements with an atomic number (nuclear charge) above 2 can be used but with much greater difficulty. This is because the Coulomb barrier increases with increasing charge of the nuclei, leading to the requirement that the plasma temperature exceed 1,000,000,000 K. Some of the more interesting reactions are

$$H + {}^{11}B \rightarrow 3\,({}^4He); \qquad Q = 8.68 \text{ MeV}, \qquad (23)$$

$$H + {}^6Li \rightarrow {}^3He + {}^4He; \qquad Q = 4.023 \text{ MeV}, \qquad (24)$$

$${}^3He + {}^6Li \rightarrow H + 2({}^4He); \qquad Q = 16.88 \text{ MeV} \qquad (25a)$$

$${}^3He + {}^6Li \rightarrow D + {}^7Be; \qquad Q = 0.113 \text{ MeV}. \qquad (25b)$$

Reaction (24) converts lithium-6 to helium-3 and ordinary helium. Interestingly, if reaction (24) is followed by reaction (25a), then a proton would again be produced and be available to induce reaction (24), thereby propagating the process. Unfortunately, it appears that reaction (25b) is 10 times more likely to occur than reaction (25a).

Methods of achieving fusion energy. Practical efforts to achieve fusion energy involve either of two basic approaches to contain and sometimes sustain a hot plasma of elements that undergo nuclear fusion reactions: magnetic confinement and inertial confinement.

Magnetic confinement. In this approach, the particles and energy of a hot plasma are confined using magnetic fields. A charged particle in a magnetic field experiences a Lorentz force that is proportional to the product of the particle's velocity and the magnetic field. This force causes electrons and ions to spiral about the direction of the magnetic line of force, thereby confining the particles. When the topology of the magnetic field yields an effective magnetic well and the pressure balance between the plasma and the field is stable, plasma can be confined away from material boundaries. Heat and particles are transported both along and across the field, but losses can be prevented in two ways. The first is to increase the strength of the magnetic field at two locations along the field line. Charged particles contained between these points can be made to reflect back and forth, an effect called magnetic mirroring. In a basically straight system with a region of intensified magnetic field at each end, particles can still escape through the ends due to scattering between particles as they approach the mirroring points. Such end losses can be avoided altogether by creating a magnetic field in the topology of a torus (*i.e.*, configuration of a doughnut or inner tube).

External magnets can be arranged to create a magnetic field topology for stable plasma confinement, or they can be used in conjunction with magnetic fields generated by currents induced to flow in the plasma itself. Many schemes have been devised over the years, but the most successful approaches have proved to be toroidal magnetic confinement designs. Examples include the tokamak, pioneered in the Soviet Union and characterized by both a strong externally produced magnetic field and a plasma current in the toroidal direction; the stellarator, pioneered in the United States and involving the use of external magnets only; and the reversed-field pinch (RFP), pursued mainly in the United States, the United Kingdom, and Italy and employing a weak toroidal magnetic field and a strong toroidally flowing plasma current. All three approaches yield magnetic field lines that follow a helical or screwlike path as the lines proceed around the torus. In the tokamak and stellarator, the pitch of the helix is weak, passing several times around the torus before the starting point returns past the initial position. By contrast, the RFP field lines trace out a path much more like a screw, wrapping many times around the cross section before covering one full length of the torus along the toroidal axis. Magnetically confined plasma must be heated to temperatures at which nuclear fusion is vigorous, typically greater than 4,400 eV, or equivalently about 75,000,000 K. This can be achieved by coupling radio-frequency waves to the plasma particles, by injecting energetic beams of neutral atoms that become ionized and heat the plasma, by magnetically compressing the plasma, or by the Joule-heating that occurs when plasma resistance dissipates the energy of electric currents induced to flow in the plasma.

Toroidal magnetic confinement designs

Inertial confinement fusion (ICF). In this approach, a fuel mass is compressed rapidly to densities 1,000 to 10,000 times greater than normal by generating a pressure as high as 10^{17} pascals (10^{12} atmospheres) for periods as short as nanoseconds. Near the end of this time period the implosion speed exceeds about 3×10^5 metres per second. At maximum compression of the fuel, which is now in a cool plasma state, the energy in converging shock waves is sufficient to heat the very centre of the fuel to temperatures high enough to induce fusion reactions. If the product of mass and size of this highly compressed fuel material is large enough, energy will be generated through fusion reactions before the plasma disassembles. Under proper conditions, much more energy

Compressing and shock-heating fuel material

can be released than is required to compress and shock-heat the fuel to thermonuclear burning conditions. The physical processes in ICF bear a relationship to those in thermonuclear weapons and in star formation—namely, gravitational collapse, compression heating, and the onset of nuclear fusion. The situation in star formation differs in one respect: after gravitational collapse ceases and a star begins to expand again due to heat from exoergic nuclear fusion reactions, the expansion is arrested by the gravity force associated with the enormous mass of the star. In a star a state of equilibrium in both size and temperature is achieved. In ICF, by contrast, complete disassembly of fuel occurs.

Conditions for practical fusion yield. Two conditions must be met to achieve a practical yield of fusion energy. First, the plasma temperature must be high enough so that fusion reactions occur at a sufficient rate. Second, the energy content of the plasma at the required temperature must be confined long enough so that the energy released by fusion reactions, when deposited in the plasma, is sufficient to maintain the temperature against heat lost by such phenomena as conduction, convection, or radiation. When this condition is first achieved, the plasma is said to be ignited. At such a temperature or higher, the energy redeposited in the plasma from fusion reactions is capable of balancing or exceeding the rate of plasma heat loss. In the case of stars or some approaches to fusion by magnetic confinement, a steady state can be achieved and no energy beyond that from fusion reactions is needed to sustain the system. In other cases, such as the ICF approach, there is a large temperature excursion once fuel ignition is achieved. The energy yield can far exceed the energy required to attain plasma ignition conditions. The energy yielded, however, is released in a burst, and the process would have to be repeated roughly once every second.

The conditions for plasma ignition are readily derived. When fusion reactions are occurring in a plasma, the power released is given by expression (16), which is proportional to the square of plasma ion density, n^2. The plasma loses energy when electrons scatter from positively charged ions, accelerating and radiating in the process. Such radiation is called bremsstrahlung and is proportional to $(n^2T^{1/2})$, where T is the plasma temperature. Assume that all other mechanisms by which heat can escape the plasma lead to a characteristic energy-loss time denoted τ. The energy content of the plasma at temperature T is $3nkT$, where k is the Boltzmann constant (see above). The rate of energy loss by mechanisms other than bremsstrahlung is simply $(3nkT)/(\tau)$. The energy balance of the plasma is the balance between the fusion energy heating the plasma and the energy loss rate, which is the sum of $3nkT/(\tau)$ and the bremsstrahlung. The condition satisfying this balance is called the ignition condition. An equation relates the product of density and energy confinement time, denoted $n\tau$, to a function that depends only on the plasma temperature and the type of fusion reaction. For example, when the plasma is composed of deuterium and tritium, the smallest value of $n\tau$ required to achieve ignition is about 2×10^{20} (particles per cubic metre times second, or m^{-3}—s) and the required temperature is equal to about 25,000 eV. If the only energy losses are due to bremsstrahlung escaping from the plasma (meaning τ is infinite), the ignition temperature decreases to 4,400 eV. Hence, the keys to generating usable amounts of fusion energy are to attain a sufficient plasma temperature and a sufficient confinement quality, as measured by the product, $n\tau$. At a temperature equivalent to 10,000 eV, the $n\tau$ product must be about 3×10^{20} (m^{-3}—s).

Energy loss by bremsstrahlung

Magnetic fusion energy generally creates plasmas with a density of about 3×10^{20} particles per cubic metre, which is about 10^{-8} of normal density. Hence, the characteristic time for heat to escape must be greater than about one second. This is a measure of the required degree of "magnetic" insulation for the heat content. Under these conditions the plasma remains in energy balance and can operate continuously if the ash of the nuclear fusion, namely helium, is removed (otherwise it will quench the plasma) and if fuel is replenished.

ICF creates plasmas of much higher density, generally about 10^{31} to 10^{32} particles per cubic metre, or 1,000 to 10,000 times normal density. As such, the confinement time need be only about 20×10^{-12} second, or 20 picoseconds. The objective in ICF is to achieve a temperature of 4,400 eV at the centre of the highly compressed fuel mass while still having sufficient mass left around the centre so that the disassembly time will exceed the minimum burn time of 20 or more picoseconds. The requirement is usually expressed in terms of the product of the mass density, ρ, given in grams per cubic centimetre, and the radius, r, of the compressed fuel mass. An $n\tau$ value of 3×10^{20} (m^{-3}—s) is roughly equal to a value of ρr of 0.3 gram per square centimetre.

Fusion energy and electric-power production. Practical fusion reactors are not yet available because the physics of containing and heating plasmas to thermonuclear conditions in a controlled manner has proved extraordinarily difficult. Large-scale fusion experiments, however, have been conducted in various countries, and the necessary conditions of plasma temperature and heat insulation have been largely achieved, suggesting that fusion energy for electric-power production is now a serious possibility.

Commercial fusion-power stations would provide an inexhaustible source of electricity. A facility of this type would of course derive its primary energy from nuclear fusion in a hot plasma. The reaction easiest to initiate is that between deuterium and tritium, though ultimately commercial fusion-power stations may employ other reactions, such as between deuterium nuclei or between deuterium and helium-3. Perhaps some of the other reactions discussed earlier will prove feasible as well. The use of deuterium as the primary fuel seems to be the most promising option, however, since it can be extracted at relatively low cost from ordinary water.

From a practical viewpoint, the initiation of nuclear fusion in a hot plasma is but the first step in a whole sequence of steps required to convert fusion energy to electricity. The intervening steps in this process would transform the fusion energy into heat at conditions appropriate, for example, to generate steam that can be employed in a Carnot cycle to drive a turbine and produce electricity. An alternative but more difficult approach may be possible in some fusion systems. The products of fusion reactions are often energetic charged nuclei, as, for example, the doubly charged helium-4 produced in reaction (11) or the proton and helium-4 released in reaction (22). The energy of these charged particles can be converted directly to electricity. The particles must be guided magnetically to a system that decelerates their initially high energy, extracting the energy as a voltage. By decelerating the particles to a lower voltage and collecting them as a current, electric power equal to the voltage difference times the collected current can be generated. The efficiency of such direct conversion can be considerably higher than the roughly 40-percent efficiency obtained in converting steam heat to electricity. This approach of direct conversion, however, is likely to be feasible only in a few specialized approaches to fusion energy. In the end, successful fusion reactors must be capable of safely producing electricity in a cost-effective manner while yielding a minimum of radioactive material and environmental impact. (Ro.W.C.)

BIBLIOGRAPHY

General history: STEVEN WEINBERG, *The Discovery of Subatomic Particles* (1983), concise exposition emphasizing 19th- and early 20th-century discoveries; EMILIO SEGRÈ, *From Falling Bodies to Radio Waves: Classical Physicists and Their Discoveries* (1984), and *From X-Rays to Quarks: Modern Physicists and Their Discoveries* (1980; originally published in Italian, 1976), readable works giving a comprehensive history of thought on the atom from the mid-1850s; GINESTRA AMALDI, *The Nature of Matter: Physical Theory from Thales to Fermi* (1966, reprinted 1982; originally published in 1961), nonmathematical history from the Greeks to 1960; HENRY A. BOORSE and LLOYD MOTZ (eds.), *The World of the Atom,* 2 vol. (1966), reprints of many original papers influential in the development of thought on the atom, highly recommended for its lively and thorough commentary; and ROBERT P. CREASE and CHARLES C. MANN, *The Second Creation: Makers of the Revolution in Twentieth-Century Physics* (1986), readable account of physicists and their discoveries in subatomic physics from Bohr to the 1970s.

Components and properties of atoms: KENNETH S. KRANE, *Introductory Nuclear Physics* (1987), includes many applications to other fields of science and technology; CHARLES KITTEL, *Introduction to Solid State Physics,* 6th ed. (1986), undergraduate-level treatment of the properties of solids emphasizing electron bands; and ROBERT EISBERG and ROBERT RESNICK, *Quantum Physics of Atoms, Molecules, Solids, Nuclei, and Particles,* 2nd ed. (1985), for readers with calculus background but no previous quantum mechanics. For the properties of chemical bonds in terms of electron orbitals, see LINUS PAULING, *The Nature of the Chemical Bond and the Structure of Molecules and Crystals: An Introduction to Modern Structural Chemistry,* 3rd ed. (1960); and C.A. COULSON, *The Shape and Structure of Molecules,* 2nd ed. rev. by ROY MCWEENY (1982). PETER RING and PETER SCHUCK, *The Nuclear Many-Body Problem* (1980), treats advanced theory of nuclear structure. See also *Atomic Data and Nuclear Data Tables,* a bimonthly journal devoted to the compilation of the properties of atoms and nuclei.

(S.McG./G.F.B.)

Isotopes: F.W. ASTON, *Mass Spectra and Isotopes,* 2nd ed. (1942), the history of the discovery of radioactive and stable isotopes; GERHART FRIEDLANDER *et al., Nuclear and Radiochemistry,* 3rd ed. (1981); KEITH J. LAIDLER, *Chemical Kinetics,* 3rd ed. (1987); STELIO VILLANI, *Isotope Separation,* trans. from Italian (1967); and JAMES W. TRURAN, "Nucleosynthesis," *Annual Review of Nuclear and Particle Science,* 34:53–97 (1984).

(G.F.H.)

Radioactivity: BERNARD G. HARVEY, *Introduction to Nuclear Physics and Chemistry,* 2nd ed. (1969), an excellent introductory text on nuclear phenomena; AAGE BOHR and BEN R. MOTTELSON, *Nuclear Structure,* 2 vol. (1969); C. MICHAEL LEDERER and VIRGINIA S. SHIRLEY, *Table of Isotopes,* 7th ed. (1978), a comprehensive table that lists all the known radioactive and stable isotopes and their properties; and ALFRED ROMER, *The Restless Atom: The Awakening of Nuclear Physics* (1960, reprinted 1982), a popular account of the discovery of radioactivity and research in that field. Collections of articles and reports are FREDERICK SODDY, *Radioactivity and Atomic Theory* (1975); and ALFRED ROMER (ed.), *The Discovery of Radioactivity and Transmutation* (1964). Applications of radiation are discussed in INTERNATIONAL ATOMIC ENERGY AGENCY, *Industrial Application of Radioisotopes and Radiation Technology* (1982); and HOWARD J. GLENN (ed.), *Biologic Applications of Radiotracers* (1982), on the use of small animals in radiotracer research.

Nuclear fission: LOUIS A. TURNER, "Nuclear Fission," *Reviews of Modern Physics,* 12(1):1–29 (Jan. 1940), an excellent review of the early studies on nuclear fission; HENRY DEWOLF SMYTH, *Atomic Energy for Military Purposes: The Official Report on the Development of the Atomic Bomb Under the Auspices of the United States Government, 1940–1945,* new and enlarged ed. (1948, reprinted 1978); and SAMUEL GLASSTONE, *Sourcebook on Atomic Energy,* 3rd ed. (1967, reprinted 1979), a comprehensive text on the atom and nuclear energy. For a detailed, authoritative treatment of all aspects of nuclear fission, see EARL K. HYDE, ISADORE PERLMAN, and GLENN T. SEABORG, *The Nuclear Properties of the Heavy Elements,* vol. 3, *Fission Phenomena* (1964, reissued 1971); and ROBERT VANDENBOSCH and JOHN R. HUIZENGA, *Nuclear Fission* (1973). Also useful are WOLF-UDO SCHRÖDER (ed.), *Nuclear Fission and Heavy-Ion-Induced Reactions* (1987), papers from a conference; and a multivolume proceedings series published by the International Atomic Energy Agency, "Physics and Chemistry of Fission." For more popular accounts of nuclear energy and its uses, see GRACE MARMOR SPRUCH and LARRY SPRUCH (eds.), *The Ubiquitous Atom* (1974); and MARTIN MANN, *Peacetime Uses of Atomic Energy,* 3rd rev. ed. (1975), a brief description of nuclear reactors and the uses of radioisotopes in industry, medicine, and scientific research. The story of the atomic bomb is told in WILLIAM L. LAURENCE, *Men and Atoms: The Discovery, the Uses, and the Future of Atomic Energy* (1959, reissued 1962); JAMES W. KUNETKA, *City of Fire: Los Alamos and the Atomic Age, 1943–1945,* rev. ed. (1978); and RICHARD RHODES, *The Making of the Atomic Bomb* (1986). (E.B.S.)

Nuclear fusion: DONALD D. CLAYTON, *Principles of Stellar Evolution and Nucleosynthesis* (1968, reprinted 1983), a description of nuclear astrophysics, covering energy generation and transport in stars, thermonuclear fusion reactions, and star burning; FRANCIS F. CHEN, *Introduction to Plasma Physics and Controlled Fusion,* vol. 1, *Plasma Physics,* 2nd ed. (1984), a basic introduction; V.E. GOLANT, A.P. ZHILINSKY, and I.E. SAKHAROV, *Fundamentals of Plasma Physics* (1980; originally published in Russian, 1977), an advanced text; J. RAEDER *et al., Controlled Nuclear Fusion: Fundamentals of Its Utilization for Energy Supply* (1986; originally published in German, 1981), an introduction to fusion energy, its technology, and the engineering aspects of conceptual fusion power reactors; ROBERT W. CONN, "The Engineering of Magnetic Fusion Reactors," *Scientific American,* 249(4):60–71 (Oct. 1983), a descriptive article on the technology of fusion machines and future fusion-energy reactors; R. STEPHEN CRAXTON, ROBERT L. MCRORY, and JOHN M. SOURES, "Progress in Laser Fusion," *Scientific American,* 255(2):68–79 (Aug. 1986), a description of inertial confinement fusion and specifically laser fusion; and EDWARD TELLER (ed.), *Fusion,* vol. 1, *Magnetic Confinement,* 2 vol., pt. A and B (1981), a series of technical articles on confinement approaches to fusion, such as the tokamak, stellarator, and magnetic mirror, and on the technology of fusion energy. (Ro.W.C.)

Attention

Attention, in psychology, is awareness of the here and now in a focal and perceptive way. For early psychologists, such as Edward Bradford Titchener, attention determined the content of consciousness and influenced the quality of conscious experience. In subsequent years less emphasis was to be placed on the subjective element of consciousness and more on the behaviour patterns by which attention could be recognized in others. Although human experience is determined by the way people deploy attention, it is evident that they do not have complete control over such deployment. There are, for example, times when an individual has difficulty in maintaining as much attention on a task, a conversation, or other set of events as he would desire. There are other times when an individual's attention is "captured" by an unexpected event rather than voluntarily directed toward it.

Attention has to do with the immediate experience of the individual; it is a state of current awareness. There are, of course, myriad events taking place in the world all of the time, impinging upon people's senses in great profusion. There are events taking place within the body affecting attention, and there are representations of past events stored away in memory but accessible to awareness under appropriate circumstances.

At first sight it might be expected that current awareness is the totality of all those events at any given moment, but clearly this is not the case. Within this vast field of potential experiences an individual focuses upon—or attends to—some limited subset, which constitutes the subjective field of awareness. It is possible to determine the reason for this limitation. Control and coordination of the many inputs and stored experiences and the organization of appropriate patterns of response are the province of the brain. The brain has impressive processing capabilities, but it has a limited capacity. A person simply cannot consciously experience all of the events and information available to him at any one time; nor is it possible to initiate at the same time an unlimited number of different actions. The question becomes one of how an appropriate subset of inputs, intermediate processes, and outputs are selected to command attention and engage available resources.

Attention, then, may be conceived as a condition of selective awareness, governing the extent and quality of man's transactions with his environment, although it is not necessarily held under voluntary control. Some of the history of attention and the methods by which psychologists and others have come to characterize and understand it are presented in the discussion that follows.

This article is divided into the following sections:

EARLY VIEWS ON ATTENTION

Following the break between psychology and the introspective traditions of philosophy, psychologists began in the latter part of the 19th century and early years of the 20th century to emphasize attention. Those philosophers who had previously considered attention at all had usually done so within the context of apperception. Thus Gottfried Wilhelm Leibniz suggested that one's loss of awareness of the constant sound of a waterfall illustrates how events cease to be apperceived (represented in consciousness) without specific attention. He suggested that attention determines what will and will not be apperceived. The term apperception was still employed in the 19th century by Wilhelm Wundt, one of the founders of modern psychology. Wundt, however, was among the first to point out the distinction between the focal and more general features of human awareness of the world. He wrote of the wide field of awareness (which he called the *Blickfeld*) within which lay the more limited focus of attention (the *Blickpunkt*). He suggested that the range of the *Blickpunkt* was about six items or groups and speculated that attention is a function of the frontal lobes of the brain.

The concept of apperception

One of the most influential psychologists at the turn of the century was William James. In his major work, *The Principles of Psychology,* published in 1890, he says:

> Every one knows what attention is. It is the taking possession by the mind, in clear and vivid form, of one out of what seem several simultaneously possible objects or trains of thought. Focalization, concentration, of consciousness are of its essence. It implies withdrawal from some things in order to deal effectively with others, . . .

James held that attention made humans perceive, conceive, distinguish, and remember more effectively and sped their reactions.

In 1906, another prominent psychologist, W.B. Pillsbury, suggested methods for measuring attention. He distinguished three separate approaches: the first relied upon tests that measured attention through performance of a task judged to require a high degree of attention; the second measured diminished attention through performance decrements; and the third gauged the strength of attention by the stimulus level required to distract the individual.

As the 20th century progressed, psychology and the study of behaviour were subject to new influences that had far-reaching consequences for notions of attention. One such area of influence originated in the work of Ivan Petrovich Pavlov who reported what is now usually referred to as the orienting response. In dogs and other animals this includes such signs of attention as pricked-up ears, turning the head toward the stimulus, increased muscular tension, and physiological changes detectable with instruments. A further influence on psychology was the work of Pavlov and that of his fellow Russian Vladimir M. Bekhterev on reflexology. Many psychologists came to regard the conditioned reflex (an involuntary response conditioned by reward), first demonstrated by these physiologists, as the basic building block of all human learning.

The influence of Pavlov

THE INFLUENCE OF BEHAVIOURISM

During this period the psychological school of behaviourism was born. Its principal advocate, John B. Watson, in 1913 defined psychology uncompromisingly as the science of human behaviour. His concern was primarily with stimulus–response relations. Attention seemed an unnecessary concept in a system of this kind, which rejected mentalistic notions, such as volition, free will, introspection, and consciousness. If used at all, the term attention was operationally defined in terms of discriminative responses to external stimuli. Ultimately it became apparent that the scientific explanations offered did not deal adequately with situations in which multiple stimuli compete with one another for attention. This led to emphasis' being placed on notions of set, attitude, and expectancy, and to a renewed interest in attention.

RELATION TO INFORMATION THEORY

Interest in attention had been subdued between World Wars I and II, and it was not until the 1940s that a true resurgence occurred. It was during this time that engineers and psychologists became involved in problems of man–machine interaction in various military contexts and found a need to formulate a new communications theory. Faced with this new range of problems, such as maintaining vigilance in soldiers watching radar systems, applied psychologists found no help in existing academic theories. As the occupational psychologist D.E. Broadbent expressed it, "attention had to be brought back into respectability." Gradually the individual came to be viewed as a processor of information, and much of the research during the 1940s and 1950s was carried out in pursuit of information theory.

Renewed interest in the 1940s and 1950s

Dependence of attention both on the unexpectedness of events and on their familiar association may seem paradoxical. This may be resolved, or at least redefined, by considering events in relation to the information they convey. Information theory suggests that the significance of any event can only be estimated in terms of what else might have happened; hence, its tendency to attract attention is considered a function of its statistical improbability. The degree of novelty, which is estimated according to the number of times an event has been experienced previously, provides a measure of its surprise value. Thus an event that has never been experienced before has a high surprise value and should attract attention, even without specific associations or consequences.

The attempts to apply information theory to a diversity of psychological problems met in the end with limited success. Nevertheless, the view of the human brain as an information processor, a type of computer, was becoming more prevalent, and the notion that one might be able to quantify the gain or flow of information proved attractive. Information itself was defined as that which reduces or removes uncertainty. The process of removing uncertainty was seen as a series of binary (yes or no) choices, rather like the game of Twenty Questions. The unit of information that expressed this primitive choice between two alternatives, or halving the residual uncertainty, was called the bit (short for binary digit). In the terms of this theory, humans are seen as a communication channel, through which information is transmitted at the rate of so many bits per second. Attempts were made to measure the capacity of this channel in many areas of human activity, but ultimately these attempts became fewer as the results were found to be too inconsistent to be useful. Finally, cognitive psychologists largely abandoned information theory, recognizing the incalculable effect of past experience on the information carried by any bit.

The brain as a computer

THE FILTER THEORY OF SELECTIVE ATTENTION

Following the decline of behaviourism in psychology, the main function of the term attention has been to provide a label for some of the internal mechanisms that determine the significance of stimuli. There is little dispute that human beings and other animals selectively attend to some of the information available to them through their senses at the expense of the remainder. One reason advanced for this is the limited capacity of the brain, which cannot process all available information simultaneously.

The perennial question arises as to whether an individual can attend to more than one thing at a time. Everyday experience leads to the conclusion that people are able to do several things at the same time. When driving an automobile they can apparently watch the road, turn the steering wheel, change gears, and apply the brakes simul-

taneously if necessary. This is not to say, however, that people attend to all of these activities simultaneously. It may be that only one of them, such as the road or its traffic, is at the forefront of awareness, while the others are dealt with relatively automatically. Another kind of evidence indicates that, when two stimuli are presented at the same time, quite frequently only one is perceived, while the other is completely ignored. In those instances when both are perceived, the responses made to them tend to be in succession, not together.

The conclusion reached and embodied in theories of the 1950s was that somewhere in the system was a bottleneck. Views differed as to where the bottleneck occurred. One of the most influential of the psychological models of selective attention was that put forward by Broadbent in 1958. He postulated that the many signals entering the central nervous system in parallel with one another are held for a very short time in a temporary "buffer." At this point the signals are analyzed for features such as their location in space, their tonal quality, their size, their colour, or other basic physical properties. They then pass through a selective "filter" that allows only those signals with the appropriate, selected properties, to proceed along a single channel for further analysis. Part of the lower priority information held in the buffer will fail to pass this stage before the time limit on the buffer expires. Items lost in this way have no further effect on behaviour. The original theory held that signals from only one source at a time could proceed. This view was modified, however, to suggest that the filter does not completely block, but simply attenuates, the nonattended signals.

Broadbent's "filter" theory

With the notion of attenuation, rather than exclusion, of nonattended signals came the idea of the establishment of thresholds. Thus threshold sensitivity might be set quite low for certain priority classes of stimuli, which, even when basically unattended and hence attenuated, may nevertheless be capable of activating the perceptual systems. Examples would be the sensitivity displayed to hearing one's own name spoken or the mother's sensitivity to the cry of her child in the night. This latter example demonstrates how processing at some level occurs even in sleep. Before attention can be said to be deployed on the activating event, however, the brain must return to a state of wakefulness. Some theorists have considered that there is no real need to postulate an early filter at all. They suggest that all signals reach central brain structures, which are, according to current circumstances, weighted to take account of particular properties. Some have a high weighting, for example, in response to one's own name; others are weighted according to the immediate task or interest. Among the concurrently active structures, that with the highest weighting gains awareness and is most directly responded to.

The information weighting system

Some critics of the above theories consider that they overemphasize the serial elements in attention. Apart from the everyday instances of tasks performed in parallel, as in the example of driving, they point to experimental evidence for highly demanding combinations of concurrent activities. As early as 1887 the French philosopher Frédéric Paulhan reported the ability to write one poem while reciting another. More recently it has been shown that some music students can sight read and play piano music while at the same time repeating aloud a prose passage. Of course it can still be held that when two such tasks are being performed together one of them is being done automatically and essentially without direct attention. An alternative explanation might be that attention alternates between them in a rapid, and frequently imperceptible, way. An analogy would be "time-sharing" on a large modern computer, where many users may be in simultaneous contact with the machine, although it is in practice servicing their demands in very rapid alternation. Each user remains relatively unaware that the interactive process is not absolutely continuous.

Another criticism of the theories already considered is that they deal only with the passive aspects of attention and that there is more to attention than mere selection. Such critics point out that there is also the question of the degree or intensity with which attention is applied to a particular task or situation. These "intensive" aspects of attention may be regarded as a subset of the broader dimension of arousal; that is to say, they relate to the continuum of awareness that extends from sleep, or even coma, at one end to alert wakefulness at the other. The topic of arousal is discussed later; for the present it is sufficient to note that the level of arousal can be determined by the demands of the task or activity in which the individual is engaged or by internal states; these are sometimes manifested as instinctive drives and frequently accompanied by high emotions, ranging from keen excitement to unpleasant stress. In the case of some drive states the high arousal may be directed to the satisfaction of a particular need. The consequences for attention can be the allocation of a high priority, or weighting, to all stimuli that relate to satisfaction of the need.

The intensity of attention

By contrast, the level of arousal associated with a particular task varies from moment to moment as the task demands change; in other words, it is very much dependent upon overall stimulus load. One of the consequences of high-demand tasks is that spare capacity decreases. At full load, virtually all attention must be concentrated on the main task, leaving little attention available for perceptual monitoring of the surrounding.

In recent years the direction of attention in response to task demands has often been spoken of in terms of the deployment of mental effort. The implication is that the intensive aspects of attention correspond to effort rather than just wakefulness. Effort, like arousal, is subject to task demands and available capacity. It is regarded as being mobilized in response to such demands, although the degree of voluntary control of effort is limited. Effort is not simply to be equated with the amount of work required by a task. Much mental activity takes place without the investment of a large amount of conscious effort.

Attempts to accommodate the selective and intensive aspects of attention and its links with both awareness and more automatic processes have led to the formulation of a number of "two-process" theories of attention. One of the most influential was that advanced by the American psychologists Richard M. Shiffrin and Walter Schneider in 1977 on the basis of experiments involving visual search. In their theory of detection, search, and attention, they distinguish between two modes of processing information: controlled search and automatic detection. Controlled search is highly demanding of attentional capacity and is usually serial in nature. It is easily established and is largely under the individual's control in that it can be readily altered or even reversed. It is strongly dependent on load. It has been suggested that it uses short-term brain storage. By contrast, automatic detection, or automatic processing, operates in long-term memory and is dependent upon extensive learning. It comes into operation without active control or attention by the individual, it is difficult to alter or suppress, and it is virtually unaffected by load. An example of controlled search would be having to identify, say, the letters *k, t,* and *v* in an array of many different letters. Automatic detection, by contrast, is exemplified by having to identify instances of, say, the number 4 in an array of letters, all of which are *c.*

The automatic and focal modes

Broadly speaking the two types of attention can be characterized as focal and automatic. Someone who is focally attentive is highly aware, consciously in control, and selective in handling sensory phenomena. A person in such a state is also employing his brain for short-term storage. (Indeed, some focal attention is almost certainly necessary for storing information in the memory at all.) Focal attention is flexible but makes great demands on brain capacity. Automatic attention makes fewer demands but is relatively inflexible. Although it deploys well-learned skills swiftly and smoothly—even when they are quite complex—it cannot cope with the unexpected. Automatic attention is largely uncontrolled, although its workings may possibly be modified by changes in the threshold at which it is triggered. The focal and automatic modes may be illustrated by a driving example: a new driver has to attend to gear-shifting in a focal way—actively thinking about it; an experienced driver, on the other hand, changes gears automatically—not having to think about it.

Attention and memory

An important aspect of the control process in many circumstances is rehearsal, meaning here the mental repetition of incoming information. One consequence of rehearsal is that input items spend an extended period of time in the short-term memory store. It is also generally the case that what is attended to and rehearsed is what is eventually stored in long-term memory, suggesting a close relationship between the conditions for awareness and those for storage in memory. Evidence for learning during sleep has sometimes been cited as contradicting this assertion that people remember only those things of which they were consciously aware at the time they occurred. It is now generally accepted, however, that the original evidence for sleep learning was suspect. In subsequent studies, when more stringent electrophysiological measures were made to ensure that individuals were in fact asleep, no clear evidence for learning during sleep could be found. There has been an indication that some type of conditioning may be possible during sleep, but, generally, awareness appears to be necessary for learning to take place.

As already noted, one of the conditions for becoming aware, or selectively engaged, is when current expectations are violated. Just as people learn skills to the point where they become subject to automatic programs of input and ouput processing, they also encode current experience into patterns of expectation that, as long as they continue to be fulfilled, need not engage focal processing resources. On entering a room a person may be aware of the regular one-second tick of a grandfather clock, but soon it fades from awareness as other things command attention. One is likely to remain unaware of it unless it stops and established expectations are violated, or unless other demands upon attention drop to the point where the person has sufficient spare focal capacity to monitor the environment more directly and again becomes at least partially aware of the sound.

Measurement of repetitive stimuli

The basic mechanism whereby response to novelty wanes with repeated and regular presentation of the same signal is usually referred to as habituation. Thus habituation may be described as a progressive loss of behavioral responsivity to a stimulus as its lack of adaptive significance is discovered. Repetition of the signal merely serves to facilitate this discovery and confirm the inappropriateness of deploying further attention upon it. In practice, the shorter the interval of time between signals the more rapidly responsiveness drops. Naturally, if the signals have special significance for the individual, they will continue to be attended to and responded to even though they may be repetitive. For example, if, to check its accuracy, a person counts the ticks of the clock, habituation to the tick will not take place. In other circumstances where stimuli have special signal properties, habituation may take place, but only very slowly. Among the factors that can affect the magnitude of response to a signal and the rate at which habituation takes place is the intensity of the signal; for example, its loudness or brightness. Although such response enhancement and resistance to habituation are primarily linked with increased stimulus intensity, they can also occur in reaction to quite faint signals, close to the threshold at which they are experienced. These observations of changes in attention with time and signal properties raise the wider question of how attention behaves over long periods of time.

VIGILANCE

Sustained attention, or vigilance, as it is more often called, refers to the state in which attention must be maintained over time. Often this is to be found in some form of "watchkeeping" activity when an observer, or listener, has continuously to monitor a situation in which significant, but usually infrequent and unpredictable, events may occur. An example would be watching a radar screen in order to make the earliest possible detection of a blip that might signify the approach of an aircraft or ship.

Attention decline during vigilance

Vigilance is difficult to sustain, however. Over any given time individuals become increasingly poor at detecting infrequent signals, and measurement of this forms the basis for studying vigilance. No single theory explains vigilance satisfactorily, probably because of its complexity. In the first place, there is a clear distinction between sustaining attention in a detection task, where the overall work load is high, and sustaining it when little is happening except for the occasional looked-for events. Under both conditions performance can show a decline over time. An important factor is the allocation of neural resources to deal with the task. These resources are to some extent fixed by virtue of the limits on processing capacity already mentioned. When the task is complex, detection difficult, time limited, and a series of decisions in using variable data is required, the brain may not succeed in coping. Long, boring, and for the most part uneventful tasks result in lowered performance with regard to both speed and accuracy in detecting looked-for events. If the task is interesting or is taking place in a stimulating environment, it is easier to sustain attention and maintain performance.

Among the factors that influence vigilance performance, the frequency with which task-relevant events occur is among the most important. Generally speaking, the more frequent the events are the better is the performance; long periods of inactivity constitute the worst case for performance. Surprisingly, the ratio of signals to nonsignal stimuli makes little difference to performance. The magnitude of the signal, however, is significant. During the course of a watch, expectancies develop about the frequency with which signals appear. If a signal occurs after an atypical interval it is less likely to be detected. Up to a point, increasing task complexity serves to improve performance, and in some vigilance situations the introduction of a secondary task can actually improve performance on the primary task. Performance is also enhanced when feedback is provided to give the individual knowledge of the results. A noisy environment can be detrimental to a vigilance task, particularly if the noise is high-pitched and loud and the task is difficult. If the task is simple and the noise is low in pitch the effect is likely to be small. Less surprisingly, lack of sleep impairs performance. Conversely, vigilance can be improved—or at least lapses prevented—by short periods of rest or by conversation or other mild forms of diversion. Monetary or other rewards tend to improve performance, as do some stimulant drugs.

THE NEUROPHYSIOLOGY OF ATTENTION

Accompanying the external manifestations of attention are physiological changes taking place within the body, particularly within the brain and nervous system. A useful starting point for examining such changes is through those circumstances where they are elicited by novel stimuli. Following the pioneering research of Pavlov, the orienting response to novel stimuli has been strongly represented in the work of Soviet neuroscientists, who have accorded it a central position in the study of mental processes. Both in the Soviet Union and in the West, the orienting response has come to be characterized less by its behavioral signs than by a broad complex of physiological changes. These embrace changes in heart rate, in the electrical conductivity of the skin, in the size of the pupils of the eyes, in the pattern of respiration, and in the level of tension in the muscles. If the novel signal is an interesting one, the heart transiently slows down; if, on the other hand, it is startling the heart transiently speeds up. Most of the other types of change reflect similar reactions. Thus, the startling signal increases the level of skin conductance and the size of the pupils of the eyes, causes respiration to pause or briefly become irregular, and increases tension in certain muscles. Closer inspection reveals many more such changes: for example, in the size of blood vessels and consequently in blood circulation, in digestive processes, and in other bodily functions. The majority of these changes are regulated by the autonomic nervous system. They prepare the individual to respond to new and potentially threatening situations. Senses become temporarily more responsive to signals from the outside world. Overall the pattern is one of preparing the individual to take in information rapidly and efficiently and of giving priority to those systems that might have to respond promptly to that information. The glands of the endocrine system come into operation, releasing hormonal agents that circulate around the body and further facilitate the preparatory process. Once the

Physical reactions during attention

novel signal has been fully assessed and classed as nonthreatening or of no continuing importance, the defenses are "stood down." As might be predicted from the behavioral evidence, repetitive signals lead to habituation of the physiological responses as novelty dissipates.

One of the crucial factors in this process is the evaluation of the signal and the assessment of its significance. Physiologically this entails shifting the level of arousal and focusing available resources (attention) on the demands the signal makes.

Signal transfers in the brain

Sensory inputs travel to the brain via primary sensory pathways that converge on a central relay structure, the thalamus, from which they are sent to relatively specific and localized receiving areas in the higher (cortical) levels of the brain. On their way from the sensory receptors to the thalamus the signals pass an area of the brain stem and midbrain to which the sensory pathways have lateral connections. This area, called the reticular formation, is important in changing the overall level of arousal. When it is damaged the individual may be unarousable. It has interconnections with the higher brain centres, and it projects pathways to the cerebral cortex. Unlike the primary sensory projections, which are limited to specific sensory modalities, many of the reticular formation cells respond to signals from any of the sensory modalities.

When this ascending reticular activating system is operating the individual is alert, aroused, and attentive. Reduction of its activity results in somnolence or inattentiveness; extreme reduction (for example, by anesthesia or concussion) may lead to confusion or unconsciousness, even though the senses still pass messages to the brain over the direct pathways. The reticular system seems to account physiologically for the sustained, tonic shifts in an individual's level of involvement with the environment, including the control of sleep-wakefulness. One nonspecific route to the cerebral cortex via the thalamus, the diffuse thalamic projection system, appears concerned with moment-to-moment fluctuations in the focus of attention. Collectively the primary sensory pathways, associated areas of the cerebral cortex, and these more diffuse projection systems cooperate in the process of registering the incoming sensory signal, evaluating its contents, and mobilizing brain resources in response to the demands made.

Inevitably this account is an oversimplification. In human beings, other brain structures, particularly the hypothalamus, are involved in regulating states of sleep and wakefulness, and limbic structures, such as the hippocampus, take part in arousal when rewards, punishments, or other emotional factors are involved. Much of our understanding of these systems and their interactions comes from the study of animal brains and from observing what happens in the human brain when things go wrong. There is, however, another important source of information about what is taking place in the healthy human brain when it processes incoming information. This is through the associated electrical changes that take place within the brain; these changes can be recorded from electrodes attached to the scalp. Such recording, known as electroencephalography, involves amplification of the very weak neuroelectric signals, often followed by computer analysis and display. Electroencephalography enables observation of the minute patterns of voltage fluctuation that take place as the brain cells process information and relay messages.

Measuring brain activity

Often the patterns of these intrinsic brain rhythms are modified by attention to external events and by thinking and other internal activity. The clearest effect of this kind is the inhibition (blocking) of so-called alpha rhythms, usually when the eyes are opened or when the person is thinking about a task, especially one involving visual imagery. Alpha rhythms are more or less regular electric oscillations at a frequency of about 10 cycles per second (hertz, or Hz), seen principally in the hindmost (visual) part of the brain. In the majority of people they tend to be most prominent when the mind is relatively blank and the eyes are closed. Absence of rhythmic features in the electroencephalogram (EEG) is generally regarded as evidence of arousal as long as there are signs of less rhythmic (asynchronous) activity; total lack of electric discharge is a serious sign of brain morbidity.

Buried within the fluctuating pattern of voltage changes are more consistent patterns that accompany the registration and evaluation of each discrete piece of sensory information. These changes are referred to as evoked potentials or, more precisely, as event-related potentials (ERP). They extend over the period of half a second or so immediately following the onset of the signal concerned. ERPs are composed of a relatively consistent pattern of positive and negative electrical peaks that vary systematically when the properties of the signal that elicits them change. The whole waveform is divided into components, which are roughly approximate to its peaks and troughs, though not exactly because there is overlap between adjacent components. Each component has its own pattern of distribution over the brain and varies according to the properties of the eliciting signal.

Use of event-related potentials (ERP)

This succession of ERP components constitutes a convenient and meaningful indicator of the various aspects of information processing being carried out on the signal. Moreover, because recording of such electrical potentials offers no insult to the brain or serious interference with the performance of a wide range of tasks, many aspects of processing can be studied in healthy as well as disordered brains. Among the components of the ERP are several related to attention in its various forms. For example, about 100 milliseconds after a novel sound is heard, a prominent negative component is produced, which, if the sound becomes repetitive, diminishes (habituates). The closer together in time the sounds occur, the smaller the component becomes. Components with similar characteristics, but varying slightly in time, are found following novel visual and tactile stimuli. In situations where the individual must pay particular attention to a signal, these electrically negative components become larger. Conversely, if the individual is not paying attention—but is, perhaps, reading a book when the sound occurs—the component is smaller. This physiological sign of selective attention can be shown to be larger to all stimuli in an attended channel than in a nonattended channel. For example, if an individual who is hearing different voices speaking simultaneously in each ear is told to listen for a particular word spoken by one voice only, all words spoken by the "attended" voice elicit a larger 100-millisecond component than those spoken by the other voice. Only the designated word, however, elicits a later prominent, electrically positive component, occurring about 300 milliseconds after it is spoken. These responses appear to offer physiological support for the behavioral view that there is an early filtering for broad characteristics, followed by a later one of the more complex task-relevant properties.

Although such an explanation is plausible, the indications are that selection is not a simple serial process taking place at two discrete stages. When the task is relatively simple, looked-for properties can be distinguished substantially earlier than 100 milliseconds. There is also evidence of a more sustained, electrically negative change that can begin before 100 milliseconds and continue for perhaps several hundred milliseconds. This overlaps several components, supporting the idea that much processing must take place in parallel. Another component with attentional properties occurs just after 200 milliseconds when the incoming signal and current expectations are mismatched.

Apart from these transient electrical links with selective attention, there are rather longer electrical changes in preparatory states when attention is directed toward making a rapid response to an expected signal. When the signal that requires the rapid response (the imperative stimulus) is preceded at a short but fixed interval by a warning or "get-ready" signal, the warning signal triggers in the cerebral cortex a slowly rising negative voltage, which reaches its maximum by the time the imperative signal arrives. (A race starter saying "on your marks, get set," may be a warning signal; the following pistol shot, an imperative stimulus.) When the response has been made the voltage returns to its normal level. This slow potential change, contingent on the association of the two stimuli and the individual's intended response to the imperative second stimulus, has been termed the contingent negative variation (CNV). It appears as a correlate of focal attention,

Contingent negative variation (CNV) factor

and it has been suggested that one of its functions may be to prime the appropriate areas of the cerebral cortex to expected stimuli. The expectation must be focal; *i.e.,* in the forefront of conscious awareness. If the preparation and response are well learned or the individual is distracted, the CNV is reduced. Conversely, if the individual concentrates on the task or is highly motivated to perform it well, the CNV increases in size. The CNV offers links with two-process theories of attention in that it seems to provide a physiological distinction between the more demanding focal, flexible mode (large CNV) and the less demanding automatic mode (small CNV). If the priming view of the CNV is correct, it may well indicate recruitment of the necessary resources to deal with the (imperative) task.

Although these are small beginnings in the quest to find a physiological basis for attention, they constitute an important first step toward the integration of behavioral and neurophysiological evidence and theory. Evidence from both humans and other animals already links the cortical CNV with similar processes taking place in the brain stem reticular formation. Several theoretical models of selective attention based on ERP evidence have been advanced. They have in common an attempt to use systematic patterns of ERP change as an index of the cerebral mechanisms underlying cognitive processes.

Chemical reactions

All of the electrical changes considered are in practice electrochemical. That is, complex chemical changes underlie the electrical correlates of attention. To take just one instance, the passage of electrical signals from nerve cell to nerve cell is dependent upon a range of neurotransmitter substances. Each of the neural systems already discussed is dependent upon the action of one—or sometimes combinations—of these neurochemicals. One transmitter substance, noradrenaline, is particularly prominent in alerting processes, along with its close relative dopamine. The total amount of another transmitter substance, acetylcholine, in the brain is found to be inversely related to the level of central nervous system activity at any given time. For example, if an individual is anesthetized, the electrical activity of the brain is reduced, and the content of acetylcholine is found to be increased. Direct electrical stimulation of the brain, or the convulsant action of certain drugs, tends to decrease brain levels of acetylcholine. This transmitter seems to be involved in a wide range of behaviour and functions. Among those related to attentional and arousal states are stress, awakening from sleep, and exploring behaviour. Certain amino acids, such as gamma-aminobutyric acid (GABA) and glycine, appear to play an inhibitory role in the brain and nervous system. Hence they too may be involved in reciprocal inhibitory processes accompanying some attentional states.

PERCEPTION AND RECALL

The vast subject of memory is beyond the scope of this survey of attention, but a few pointers to the interactions that take place between what is attended to, how it is perceived and recognized, and factors that govern its subsequent recall are relevant. Memorizing is not simply a matter of repetition; attention plays a role in organizing material in ways that can influence its later recall. One example, known as the Von Restorff effect, is that, in any given number of items to be learned, an item that is notably different from the rest in size, colour, or other basic characteristics will be more readily recalled than the others. Unfortunately there is a price to be paid for this improvement; other "standard" items will be less well recalled than they otherwise would have been.

It is also important to realize that what is actually perceived is not a neutral, objective representation of what exists in the external world. It is coloured by past experiences stored in memory and by current expectations, to the extent that substantial distortions can occur to make a perceived item fit those experiences and expectations. Perceptions are frequently formed on the basis of quite limited cues; the art of camouflage utilizes this characteristic to the benefit of both humans and other animals in certain situations. It seems that even the culture within which a person lives determines the way he perceives the world. Following a study of the Hopi and Shawnee languages, the linguist Benjamin Whorf concluded that what these American Indian peoples perceived was itself different from the perceptions of English-speaking Americans, by virtue of the way their languages were structured.

INATTENTION AND DISTRACTIBILITY

Limited processing capacity entails that there is invariably competition for attention. While awake humans are essentially always attending to something. The term inattention usually implies that, at a given moment, the thing being attended to is either not what it was intended to be or not what adaptively it ought to be. People will often report, "I was attending, but found that I was not taking in what was happening." On many such occasions, internal preoccupations (thoughts) have become the object of current attention at the expense of sensory information from the external world. Alternatively, an internal stimulus, such as a pain or hunger, may have captured attention. On other occasions, irrelevant sensory information from the external world may distract individuals from their current focus of attention. When this happens it is either because the intrusive stimulus has high priority, for example, the ringing of a telephone, or perhaps because the task engaged in is simply uninteresting.

Some individuals are more easily distracted than others, but in everyone distractibility varies with circumstances. When motivation and the level of involvement are high, an individual may totally disregard intense and persistent "outside" signals. Such inputs are either heavily filtered or dealt with only at an automatic level. Even when the competing stimulus is pain from an injury sustained, say, by a player in the early stages of a team sport, it is often scarcely noticed until the game ends and attention is no longer absorbed by the game. Nevertheless, because people's ability to focus attention varies, some report "difficulties of concentration" and may find themselves so easily distracted that they can scarcely read a book. There are indications that persons who are chronically anxious may be among those whose attention can readily be distracted by quite modest and irrelevant levels of stimuli. This feature has been noted in a number of psychiatric disorders, and it has been suggested that one cause of these disorders may be a fault in the mechanisms of attention.

Levels of distractibility

LAPSES OF ATTENTION

It has been established that, to conserve limited resources, whole sequences and hierarchies of actions can apparently be elicited without focal attention when they have been well learned or executed many times. There is reason, however, to suppose that at least a minimum level of focal attention may be necessary, if only to ensure that the correct sequence or hierarchy is initiated. Failure of this minimal monitoring can result in the phenomena usually classed as lapses of attention. For example, most people have experienced trivial behavioral slips such as finding themselves taking a regularly used route when they had meant to go in a different direction; attempting to switch off a light when leaving a room in daylight; or, perhaps, pouring tea into the sugar bowl instead of a cup. In each case a well-established action has been inappropriately triggered by partial cues and has slipped past the attentional monitor when it was otherwise engaged. A person realizes in retrospect that the error has occurred because he was "thinking about something else"; he was not paying attention to what he was doing. In many circumstances it is advantageous that automatic sequences of behaviour should be executed with only very limited reference to conscious attention. Musicians, typists, and other skilled persons are well aware that too much attention devoted to the execution of a well-learned skill can disrupt performance. Nevertheless, people cannot dispense entirely with some degree of attentional monitoring if they are to avoid errors. Another kind of lapse entails being unable to remember whether one has performed a particular action as part of a highly automated sequence: "did I or did I not put the sugar in my tea?" Yet another kind results in one automatic action's triggering another unwanted or inappropriate action: "I meant to take off only my shoes but took off my socks as well." Most lapses have in com-

Limited attentional monitoring

mon that they occur when attention has been claimed by an internal preoccupation or external distraction.

CONCLUSION

Attention has sometimes been described not as a single concept but as the name of a complex field of study. This is true only to the extent that around it have grown up a multitude of ancillary and often poorly defined constructs, many of them overlapping. Some, like consciousness and awareness, are related to subjective mental states. Others, like arousal, activation, and orientation, are thought of more in physiological terms. Still others, like alertness and expectancy, have been characterized principally in terms of behaviour and performance. Another dimension considers the process in terms of effort, intention, drive, or motivation, and, yet another, the notion of automaticity. Doubtless each of these facets contributes to the overall picture, but there would seem to be a case for treating the term as referring primarily to that state of the individual which represents the shifting, selective focus of consciousness. This is the state through which learning takes place and one that makes heavy demands upon the brain's processing capacity. Individuals recognize it subjectively in themselves, but it is becoming increasingly recognizable in others through neurophysiological activity as well as by individual behaviour. It is a state of awareness that subserves the more flexible and directable aspects of human transactions with the environment.

BIBLIOGRAPHY. D.E. BROADBENT, *Perception and Communication* (1958), presents an approach that uses communication theory; D.R. DAVIES and G.S. TUNE, *Human Vigilance Performance* (1969), is a review of vigilance research and theory; DANIEL KAHNEMAN, *Attention and Effort* (1973), is a textbook on the psychology of attention, with particular emphasis on selective and intensive dimensions; STEVEN W. KEELE, *Attention and Human Performance* (1973), gives an account of human information processing and attention in relation to memory storage and retrieval; DAVID I. MOSTOFSKI (ed.), *Attention: Contemporary Theory and Analysis* (1970), is a collection of papers on key issues; RAJA PARASURAMAN and D.R. DAVIES (eds.), *Varieties of Attention* (1984), is a comprehensive series of review papers covering the major psychological and physiological aspects of attention; CARL M. STROH, *Vigilance: The Problem of Sustained Attention* (1971), is an account of the factors influencing vigilance, its physiological correlates, and theories of vigilance performance. Current contributions of experimental psychologists can be found in the volumes entitled *Attention and Performance,* a collection of papers presented at various international symposia, beginning in 1966.

(W.C.McC.)

Augustine

Saint Augustine (in Latin, Augustinus), bishop of Hippo in Roman Africa from 396 to 430, and the dominant personality of the Western Church of his time, is generally recognized as having been the greatest thinker of Christian antiquity. His mind was the crucible in which the religion of the New Testament was most completely fused with the Platonic tradition of Greek philosophy; and it was also the means by which the product of this fusion was transmitted to the Christendoms of medieval Roman Catholicism and Renaissance Protestantism.

This unique significance would have belonged to Augustine had he never written the famous *Confessions,* in which at the age of about 45 he told the story of his own restless youth and of the stormy voyage that had ended, as he believed, 12 years before he put it in writing, in the haven of the Catholic Church. It is easy to forget that the real work of Augustine's life did not begin until the last scene of the *Confessions* was already receding for him into a remembered past. Moreover, the *Confessions* themselves are not so much autobiography as they are devotional outpourings of penitence and thanksgiving. Augustine's conscientious memory generally can be trusted for the facts: his reflections upon them are those of the bishop on his knees. This is not to say that, in any attempt to understand or appreciate the mind of the bishop, the *Confessions* can be neglected. The picture must, however, be drawn in proper proportion; it is essential to avoid giving undue prominence to what should be no more than its background.

Youth and conversion. Hippo Regius is the modern Annaba on the Algerian coast, in what was then the Roman province of Numidia. Augustine, named Aurelius Augustinus, was born on November 13, 354, of middle-class parents at Tagaste (modern Souk-Ahras), a small town about 45 miles (72 kilometres) to the south. His father, Patricius, was and remained until late in life a pagan; his mother, Monica, was a Christian of intense but simple piety, from whose early teaching Augustine retained a reverence for the "name of Christ" that never left him. But he was not baptized in infancy. He went through primary and secondary schooling and soon displayed such intellectual promise that the modest family funds were banked upon securing him an academic career that would qualify him for government service. As a 19-year-old student at Carthage he was stirred by the reading of a treatise of Cicero—the now lost *Hortensius*—and was filled with an enthusiasm for "philosophy," which meant not only a devotion to the pursuit of truth but a conviction of the superiority of a life devoted to that pursuit (the *vita contemplative*) over any aims of secular ambition. The faith of the Catholic Church seemed to him too hopelessly unphilosophical for any man of culture to entertain; and he was easily carried away by the discovery in Manichaeism of a religion that professed to appeal to reason rather than authority.

Enthusiasm for philosophy

Alinari—Art Resource/EB Inc.

St. Augustine, fresco by Sandro Botticelli, 1480. In the Church of Ognissanti, Florence.

Influence of Manichaeism. The Manichaean system as propagated in the Western Roman Empire was a materialistic dualism that accounted for the creation of the world as the product of a conflict between light and dark substances and for the soul of man as an element of the light entangled in the dark. Manichaeism claimed to be

the true Christianity, preaching Christ as the Redeemer who enables the imprisoned particles of light to escape and return to their own region. In the Manichaean Church the higher order of "elect" were strictly ascetic and celibate, all physical generation being held to serve the realm of darkness. After an adolescence that probably was no more licentious than was common in his time and country, Augustine had formed a liaison with a woman of low birth by whom he had a son and to whom he remained loyally attached throughout the nine years of his association with the Manichaeans, and he was therefore allowed to join that sect's lower order as one of the "hearers," to whom marriage was permitted as a concession to human weakness.

His first zeal for this "religion of enlightenment" did not last long, however, for the Manichaean experts were intellectually second rate and proved incapable of dealing with the questions he put to them. He became increasingly disillusioned and was already falling into a general agnosticism when, at the age of about 28, he left Carthage, where he had worked as a free-lance teacher of rhetoric, and went to Rome in search of more satisfactory pupils. There he made connections that led to an official professorship at Milan, where the Western emperor then resided.The bishop of Milan was Ambrose, the most eminent Christian churchman of the day. Augustine was introduced to Ambrose but never came to know him well. He went to hear him preach, however, and this, his first contact with the mind of a Christian intellectual, was enough to shake Augustine's prejudice against Catholic teaching. Although he had abandoned the doctrines of Manichaeism, he retained its materialistic presuppositions, which left him still a skeptic with no satisfying alternative to Manichaean notions of ultimate reality. The being of God and the nature and origin of evil remained for him problems as insoluble as they had ever been.

Contact with St. Ambrose

Influence of Neoplatonism. The solution of both problems was given to him by a chance introduction to Neoplatonic writings, for which he may well have been prepared by Ambrose's use of them in some of his sermons. Neoplatonism, in the work of the 3rd-century philosopher and mystic Plotinus, its greatest exponent, is a spiritual monism—a philosophical doctrine holding that there is only one reality—according to which the universe exists as a series of emanations or degenerations from absolute unity. From the transcendent One arises self-conscious mind or spirit; from mind comes soul or life; and soul is the intermediary between the spheres of spirit and of sense. Matter is the lowest and last product of the supreme unity; and since the One is also the real and the good, the potentiality of evil is identified with unformed matter as the point of maximum departure from the One. Evil itself is thus the least real of all things, being simply the privation or absence of good. Neoplatonic mysticism relies on the principle that the inward is superior to the outward: to reach the good, which is the real, one must "return into" oneself; for it is the spirit at the heart of man's inmost self that links him to the ultimate unity.

In the seventh book of the *Confessions,* Augustine tells how in such an act of introspection he found God—the "changeless light," at once immanent and transcendent, which is the source of every intuitive recognition of truth and goodness. This discovery of God was more than the conclusion of a process of reasoning: it was a mystical experience, a vision or touch that came and went. But it left behind it the answer to Augustine's unsatisfied questionings. God is light, and evil is darkness, as the Manichaeans said. But neither is a material substance: the changeless light of God is pure spiritual being, and the evil is nonentity, as darkness is but the absence of light.

Conversion to Christianity. Augustine's mystical experience, his awareness of God, had been momentary and fleeting. He believed that this could be only because he had not made for himself the necessary total identification of supreme value with spirit; he was still himself entangled with the flesh. In fact, Neoplatonism had reinforced the Manichaean principle that the way of return to God must be through escape from the body; and for Augustine this meant primarily and immediately escape from the ties of sexuality. The immortal story of his conversion in the eighth book of the *Confessions* tells of his coming to learn of the heroic achievements of Christian asceticism in East and West, of the self-contempt induced in him by the contrast of his own weakness, and of the final breakdown of resistance in a Milan garden, when, at the sound of a child's voice calling "*tolle, lege: tolle, lege*" ("take up and read"), he opened the New Testament Letters and read in Letter of Paul to the Romans the words, ". . . put on the Lord Jesus Christ, and make no provision for the flesh, to gratify its desires" (Rom. 13:14).

Moral struggle

This was in the late summer of the year 386. Vacation was near, and Augustine resigned his teaching chair and went with some young pupils, his son Adeodatus, and his mother Monica to a reading party at a country house lent by a friend. Out of their literary study and philosophical discussions there came the earliest of Augustine's surviving works—the dialogues, which display so little of the storm and stress of a religious conversion and so little concern with specifically Christian themes that critics have been led to question the accuracy of the *Confessions* story written many years later. It is true that Augustine's struggle against the domination of his sexual nature can be regarded as the final phase in that fluctuating pursuit of the "philosophic life" first presented to him by Cicero's *Hortensius.* But there is no sufficient reason for doubting that he was a Catholic Christian in intention when he received Baptism at the hands of Ambrose in the spring of 387. It is certain that three or four years later, when he wrote his treatise *De vera religione* (*Of True Religion*), he was still thinking of Christianity in Neoplatonic terms. In this treatise, the divine Word (Logos) in Christ is the mind or spirit of Plotinus, illuminating the reason, through whom the human soul has access to the transcendent Godhead. Christ's human life is man's example of the ascetic victory over the pains and pleasures of the flesh; Christian morals serve only to purify the soul for the life of contemplation; and Christian faith is the necessary acceptance of the church's authority in this preliminary stage of training.

Bishop and Christian philosopher. Shortly after his Baptism, Augustine left Milan, with his mother and a small party of friends, to return to Africa. At Rome's port city of Ostia, his mother died; and Augustine recorded his last talk with her, in which son led mother, through a discourse formed on the pattern of the Neoplatonic "ascent" from this world to the other, to share with him a momentary experience of the life eternal. Home again at Tagaste, the friends formed a little community devoted to the religious life of contemplation and study. But its peace was soon broken when, on a visit to Hippo in 391, Augustine was forced to accept ordination as assistant priest to its old bishop, Valerius. Five years later Valerius died, and Augustine entered the episcopate in which he was to labour until his death. The bishop in Roman Africa was not only the pastor of a parish, the busy teacher and preacher, but the presiding judge in a much-frequented court of summary jurisdiction in civil cases. Augustine never enjoyed robust health, and the vast extent of his literary output was made possible only by the constant services of stenographers and by an extraordinary capacity for the extempore formulation of ordered thought, of which at least 400 sermons remain as proof. He was not a systematic theologian. Much of his writing was in response to the appeals that his growing reputation in the Christian world brought to him for the solution of the most diverse problems. Over 200 of his letters have been preserved, many of them having the scale of minor treatises. He was tireless in controversy with heretics—Manichaeans, Donatists, and Pelagians. But his deepest thought, the real Augustinianism, is to be found in his scripture commentaries and homilies, especially his expositions of the Psalms and his writings on the Gospel and First Letter of John. The characteristic pattern he imposed upon Christian theology was not the outcome of controversy.

Pastor, teacher, and judge in Hippo

The decisive turn was given to his thinking by his ordination to the priesthood, which dragged him against his will from the *vita contemplative* into the world and at the same time diverted his studies from philosophy to Scripture. The realities of pastoral experience among the very imperfectly Christianized people of an African seaport,

Biblical and philosophical theologian

together with the rapid impregnation of his mind with the categories of biblical religion, made it impossible for him to overlook the differences between Neoplatonism and Pauline Christianity. The knowledge of God and of the soul always remained from the time of his Baptism the one and only knowledge that he desired; and Plotinus had not been mistaken in bidding him look within himself if he would find God, for the Bible also tells of a likeness to God imprinted on the soul. But although for the Neoplatonist the soul's likeness to God is that of a, so to speak, reduced divinity, for the Christian it is that of a temporal and mutable image of the "eternal and changeless." Augustine was assured that it is the task of a Christian philosophy, guided by the scriptural revelation, to seek to know God through his image in the soul; and this was the path he followed in his great treatise *De Trinitate* (*On the Trinity*). He insisted that a true knowledge of the soul's nature can be based only on the immediate awareness of self-consciousness; and the soul's awareness of itself is of a trinity in unity that reflects "as in a glass darkly" the being of its Maker. He claimed that knowledge of one's own being, of one's own thinking, of one's own willing is not open to doubt; there is an ego that exists, knows, and wills. But in none of these aspects is the ego self-sufficient or independent: it cannot maintain its own being, produce its own knowledge, or satisfy its own desires. Augustine believed that he had learned from the Platonists to find in God "the author of all existences, the illuminator of all truth, the bestower of all beatitude" (*De civitate Dei* viii, 4). But his theories of the universe, of knowledge, and of ethics were his own. The following three paragraphs summarize these theories.

Theory of the universe. Creation in Plotinus is motiveless and purposeless, the automatic by-product of the divine self-contemplation; in Augustine its source is "the will of a good God that good things should be" (*De civitate Dei* xi, 21). The outgoing energy of creative love forms the basic principle of his entire theology. Since nothing can come into being or continue in it but by this divine will to create, all that exists is good "in so far as it has being"; and because there are evidently degrees of goodness, there must also be degrees of being. But even the formless matter that is nearest to "not being" is essentially good because God made it; the origin of evil is not to be sought in material existence. Augustine persistently refused to unload upon the material conditions of human life the responsibility for human wickedness.

Theory of knowledge. Following Plato, Augustine argued that the ability to make true judgments never can be inserted into the mind from outside. The human teacher never can do more than help his pupil to see for himself what he already knew without being aware of it. Augustine's favourite examples of these intuitive judgments are the propositions of mathematics and the appreciation of moral values; they are not the construction of the individual mind, because when properly formulated they are accepted by all minds alike. The individual thinker does not make the truth, he finds it; and he is able to do so because Christ, the revealing Word of God, is the *magister interior,* the "inward teacher," who enables him to see the truth for himself when he listens to him.

Ethics. Augustine accepts the basic assumption of ancient ethical theory that conduct is properly directed to the achievement of *eudaimonia*—the happiness or well-being that is taken to be the one universal desire of humanity. Augustine's cosmos is an ordered structure in which the degrees of being are at the same time degrees of value. This universal order requires the subordination of what is lower in the scale of being to what is higher: body is to be subject to spirit, and spirit to God. Man must know his place in the order of the universe and, knowing it, must voluntarily accept it; that is, he must set upon himself and upon everything else the relative value that is properly due. Augustine's word for the ethical valuation that influences conduct is *amor* ("love"). *Amor* is the moral dynamic that impels man to action. If it is rightly directed man will never set a higher value on what is lower in the scale. All lesser goods are to be "used" as means or aids toward the higher; only the highest is to be "enjoyed" as the ultimate end on which the heart is set. The supreme good in whose fruition alone man reaches his perfection is for Augustine the God whose nature is *agapē*, love in the New Testament sense of the word. If, then, man's love, his *amor,* can rise to the enjoyment of God, it will become a participation in the divine *agapē,* love itself. God will have given himself to men, and by sharing in his love men will love one another as he loves them, drawing from him the power to give themselves to others.

Struggle with the Donatist schism. The energies of Augustine, both pastoral and literary, were for the first 15 years of his episcopate distracted by the wearisome struggle to end the schism in the African Church that had persisted for nearly a century. The Donatists, a Christian sect (named after Donatus, one of its leaders) the members of which outnumbered the Catholics in the country districts and in many towns, claimed to be the only true church on the ground that their ministry was the only one the succession of which had not been stained by apostasy in the great persecution of the years 303–313, which had begun under the emperor Diocletian. Imperial attempts to suppress the schism had stimulated the martyr spirit that had always marked African Christianity and gained Donatism the support of strong elements in the native population whose grievances were social and economic rather than ecclesiastical. The schism maintained itself by fanatical violence, and Augustine's persevering attempts to settle the questions at issue by peaceful discussion were fruitless. In the end, the imperial government became convinced that the Donatists were a danger to the security of Africa. The Donatist bishops were compelled to meet their Catholic rivals at a formal conference held under an official arbitrator at Carthage in 411, the foregone conclusion of which was a Catholic victory.

Answer to the Donatists

Donatists and Catholics agreed that the power of the Holy Spirit is conveyed to the believer through the sacraments, which are administered by the church through the clergy. The Donatists alleged, however, that the sacraments require for their validity a ministry undefiled by serious sin; for the Spirit departs from the sinner, who cannot therefore "confer what he does not possess." Augustine replied that the sacraments convey the Spirit in virtue of Christ's ordinance alone and that this validity is unaffected by the worthiness or unworthiness of the human minister. The church's unity depends on the Spirit's supreme gift of charity, of which schism is the denial. Unfortunately, Augustine, who had for long opposed the use of any means but persuasion to end the schism, eventually was induced to approve the enforcement of legal penalties upon the schismatics, in the interest, as he believed, of the many whose fear of Donatist violence had kept them from returning to the church. His famous saying, "Love, and do what thou wilt," was in fact a defense of compulsion in the service of charity.

Struggle with the Pelagian heresy. As the Donatist controversy was ending, the Pelagians were already beginning to threaten the traditional doctrines of sin and redemption in the Western Church. Pelagius had set himself to resist the slackening of Christian moral standards. Against those who pleaded human frailty in excuse for their failings, he insisted that God has made every man alike free to choose and to perform the good; that it is the essence of sin to be a voluntary act that God's law forbids and that the sinner was free to avoid; and that, were not this freedom real, there could be no justice in God's punishments and rewards. This reduction of Christianity to a bleak moralism could not avoid conflict with the plain implications of the church's sacramental and liturgical practice. Baptism had always been "for the remission of sins," and infants were held to need it because they inherit the guilt of Adam's transgression, which, as St. Paul taught, brought death upon the whole race of men. The doctrine of original sin was firmly established in the Western Church before Augustine's time; and when it was openly rejected by Pelagius' disciple Celestius, there was no escape for Pelagianism from being branded as a heresy. The prevarications of Pelagius were able to persuade Pope Zosimus (417–418) to reverse the condemnation pronounced by his predecessor, Innocent I. But in the spring of 418 the African bishops

obtained from the emperor Honorius an edict banishing the heretics; and Zosimus was obliged to come into line.

Augustine was the soul of the Church's resistance. He had seen Pelagianism at once as not merely a denial of the virtue of Christian Baptism but also as a fatal misconception of the relationship between God and man. For to assert that man can achieve righteousness by his own effort is to contradict the fundamental truth that God is the giver of all good.

Original sin and its propagation

Before the controversy began, Augustine had worked out his own rationalizations of the doctrines of original sin and divine grace—rationalizations that the church was to prove unwilling to accept fully. He accepted the traditional belief in the fact and in the penal consequences of Adam's transgression, defining the fact as man's refusal to accept his place in the created order, and the consequences as a dislocation of the order of man's own nature—the revolt of flesh against spirit. He argued that not only are all men involved in Adam's guilt and punishment but also that this involvement takes effect through the dependence of human procreation on the sexual passion, in which the spirit's inability to control flesh is evident. It was this linking of original sin with human sexuality that exposed Augustine in his old age to the most damaging criticisms of the Pelagian bishop Julian of Eclanum, who boldly asserted the moral neutrality of the instincts that belong to man's created nature and charged Augustine with relapsing into Manichaeism in his argument that an impulse that a man is bound to fight and conquer must therefore be evil.

For Augustine the fall of man means that in all men the true order of love has been violated. Departing from the love of God above him, man has followed the love of self and become subject to what is below him. Man has fallen by the act of his own will. He cannot by a similar exercise of will reverse the consequences of that fall. The subjection of spirit to flesh is a slavery from which the perverted will has no power to deliver itself, just because it cannot will the deliverance. What is needed is a kind of reversal of gravity—the substitution of an uplifting for a down-dragging love. And Augustine believed that this could happen only by that gracious descent of the divine love to dwell within the sinner: the gospel of the incarnation and of Pentecost.

Pelagius, on the other hand, argued that all men have been created free to do what is right when they see it, and that Christians have received the needed moral enlightenment in Christ's teaching and example. Augustine knew the unreality of the Pelagian conception of freedom as an innate and absolute power of choice, unaffected by circumstances. He pointed to the inescapable conditioning of all moral activity by the situation of the agent—outside whose control are in general not only the presentation of an object but also the kind of feeling that the presentation excites. Moreover, the act of will is dependent on feeling as well as on cognition. In Augustine's words:

> Men will not do what is right, either because the right is hidden from them or because they find no delight in it. But that what was hidden may become clear, what delighted not may become sweet—this belongs to the grace of God" (*De peccatorum meritis et remissione*).

Augustine insisted that without this delight in righteousness there can be no true freedom in well-doing, but only a servile obedience to law. The love of God, which is the motive of the Christian life, must be free. Yet love of God, as St. Paul said, enters man's heart by the gift of the Holy Spirit; and Augustine found it increasingly difficult to leave room in his doctrine of grace for a genuinely free response on man's part to the Spirit's gift. The unexamined assumption that everything in human life must be ascribed either to God's or to man's working compelled him to hold that God alone is the cause of every human movement toward good. In the first year of his episcopate, the study of St. Paul's argument in Rom. 9–11 had convinced him that no event in time can alter the eternal setting of God's will toward any human soul: his elect are chosen before the foundations of the world. God knows—not before, but apart from, the time process—how each individual in the course of time will respond to the particular form in which grace is offered to him; and the elect alone receive the grace that will win their acceptance.

The City of God

The rigour of this doctrine did not soften in face of the Pelagian challenge. In *De civitate Dei* (*The City of God*), the masterpiece on which Augustine was working throughout the Pelagian controversy, he drew a picture, as majestic as it is appalling, of the "beginnings, course and destined ends" of the two invisible societies of the elect and the damned. The work seems to have been in his mind before the capture of Rome by the Visigoths in 410 had shaken the empire; but it took the form of a Christian apologetic against the pagan claim that the disaster was consequence and punishment of Rome's apostasy from its ancestral religion. Augustine's two cities are not to be identified with the Christian Church and the pagan or secular state. They are symbolic embodiments of the two spiritual powers that have contended for allegiance in God's creation ever since the fall of the angels—faith and unbelief, "the love of self extending to contempt for God, and the love of God extending to contempt of self." Neither power is embodied in its purity in any earthly institution; in this world the heavenly and earthly cities are inextricably intermingled. If there is a philosophy of history in the *De civitate Dei,* it is the religious philosophy of predestination.

Augustine found it difficult in his old age to reassure some of his own disciples, to whom his doctrine seemed to make moral effort futile and praise and blame alike groundless. But he would retract nothing. His last completed treatises drew out the logic of predestination to its most ruthless conclusions. Though his doctrine in its final form was never accepted by the church, it reappeared virtually unmodified in the writings of both St. Thomas Aquinas and John Calvin, the most acute thinkers, respectively, of Scholasticism and Reform. It may indeed be regarded as product of the too audacious attempt of the time-bound human mind to contemplate existence with the eye of the eternal God.

The influence of Augustine. The end of Roman civilization in Africa was near and the Vandal armies were besieging Hippo when Augustine died there on August 28, 430. Not many years later, Vincent of Lérins defined Catholic orthodoxy in the famous phrase, *Quod ubique quod semper quod ab omnibus creditum est* ("What is everywhere, what is always, what is by all people believed"). He dared not call Augustine a heretic in so many words, but it was against the extravagances that he rightly detected in Augustinian doctrine that his definition was aimed. That these extravagances have been a noxious legacy to theology because of their author's authority cannot be denied. But that should not prevent the grateful acknowledgment of the debt that Christian thinking has owed through the centuries to Augustine's influence, which has spanned and may one day reconcile the divisions of Western Christendom. The secret of that influence is to be found not so much in the brilliance and profundity of his intellect, the magic of his style, or the validity of his constructions as in the unique power of his religious genius. St. Anselm of Canterbury, St. Bernard of Clairvaux, the makers of *The Book of Common Prayer,* St. Francis de Sales, Blaise Pascal, Jacques-Bénigne Bossuet, Joseph Butler, Jacques Maritain, Reinhold Niebuhr, and Paul Tillich—all these have in their different ways drawn inspiration from one in whom they have been compelled to recognize "the heart of the matter." *Verus philosophus est amator Dei* ("The true philosopher is the lover of God"). In those words from the *De civitate Dei,* Augustine has left at once the best portrait of himself and the fullest justification of his life's work.

St. Augustine has been revered as a doctor of the church since the early Middle Ages. His feast is celebrated on August 28.

(Jo.Bu.)

MAJOR WORKS

TEXT AND TRANSLATIONS: Modern critical editions of St. Augustine's works in the original Latin are in process of publication in the *Corpus Scriptorum Ecclesiasticorum Latinorum* and in the *Corpus Christianorum;* but the only available edition complete except for the Sermons is still that of the Benedictines of Saint-Maur (1670–1700), reprinted in Migne's

Patrologia Latina. There are no complete English translations of all St. Augustine's works. The largest separate collection is in the series "Nicene and Post-Nicene Fathers of the Christian Church" (N.P.N.F.). Translations of most of the Major Works listed below can be found either in this collection or in one or other of the following more recent series: "Ancient Christian Writers" (A.C.W.); "The Fathers of the Church" (F.C.); "The Library of Christian Classics" (L.C.C.).

GENERAL: *Confessiones* (*c.* 400; *The Confessions,* L.C.C.); *De doctrina Christiana* (397–428; *Christian Instruction,* F.C.); *De Trinitate* (400–416; *On the Trinity,* N.P.N.F.); *De civitate Dei* (413–426; *The City of God,* F.C.); *Enchiridion ad Laurentium de fide, spe, et caritate* (421; *Enchiridion to Laurentius on Faith, Hope, and Love,* L.C.C.); *Sermones* (from 391; *Selected Sermons,* ed. by *Quincy Howe,* 1966); *Epistolae* (from 386; *Letters,* F.C.).

EXEGETICAL: *De Genesi ad litteram* (401–415), a commentary on the first three chapters of Genesis; *De sermone Domini in monte* (393–394; *Commentary on the Lord's Sermon on the Mount,* F.C.); *Enarrationes in Psalmos* (391–420; *Expositions on the Book of Psalms,* 1847–57; A.C.W. incomplete); *Tractatus in Joannis Evangelium* (407–418; *Homilies on the Gospel of John,* N.P.N.F.); *Tractatus in Epistolam Joannis ad Parthos* (*c.* 415; *Homilies on St. John's Epistle,* L.C.C.).

CONTROVERSIAL: (ANTI MANICHAEAN): *De vera religione* (*c.* 390; *Of True Religion,* L.C.C.); *De libero arbitrio* (389–395; *On Free Will,* L.C.C.). (ANTI-DONATIST): *De Baptismo, contra Donatistas* (400–401; *On Baptism, Against the Donatists,* N.P.N.F.); *Contra litteras Petiliani* (400–403; *Answers to Letters of Petilian,* N.P.N.F.). (ANTI-PELAGIAN): *De spiritu et littera* (412; *The Spirit and the Letter,* L.C.C.); *De natura et gratia* (415; *On Nature and Grace,* N.P.N.F.); *De gratia Christi et de peccato originali* (418; *On the Grace of Christ, and on Original Sin,* N.P.N.F.); *De gratia et libero arbitrio* (426 or 427; *On grace and Free Will,* N.P.N.F.).

BIBLIOGRAPHY. A comprehensive bibliography of works dealing with St. Augustine is CARL ANDRESEN (ed.), *Bibliographia Augustiniana,* 2nd ed. (1973); TARSICIUS J. VAN BAVEL and F. VAN DER ZANDE, *Répertoire bibliographique de Saint Augustin* (1963), covers material that appeared between 1950 and 1960; for the years 1970–80, see TERRY L. MIETHE, *Augustinian Bibliography: 1970–1980* (1982). Annual bibliographies are provided in *L'Année philologique* (1924–); *Revue des études augustiniennes* (quarterly); and *Recherches augustiniennes* (1958–).

Biography: WARREN T. SMITH, *Augustine: His Life and Thought* (1980), is a good introduction. A scholarly and readable biography is PETER R.L. BROWN, *Augustine of Hippo* (1967). GERALD BONNER, *St. Augustine of Hippo: Life and Controversies* (1963), is also valuable. Of literary interest is REBECCA WEST, *St. Augustine* (1933). See also KARL ADAM, *Saint Augustine: The Odyssey of His Soul* (1932; originally published in German, 1931); and HUGH POPE, *Saint Augustine of Hippo* (1937, reissued 1961). The problems concerning the chronology and nature of Augustine's conversion, especially as related in his *Confessions,* are dealt with in PAUL AUBIN, *Le Problème de la "Conversion"* (1963); J.M. LE BLOND, *Les Conversions de Saint Augustin* (1950); A.M. LA BONNARDIÈRE, *Recherches de chronologie augustinienne* (1965); PIERRE P. COURCELLE, *Recherches sur les Confessions de Saint Augustin,* new ed. (1968); JOHN J. O'MEARA, *The Young Augustine* (1954, reissued 1980); and MICHELE PELLEGRINO, *Les Confessions de Saint Augustin* (1961). Augustine's maturity is described in FREDERIK VAN DER MEER, *Augustine the Bishop* (1961; originally published in Dutch, 1947).

Thought: A general outline of Augustine's thought is provided in PROSPER ALFARIC, *L'Évolution intellectuelle de Saint Augustin* (1918); also good introductory texts are HENRI I. MARROU, *St. Augustine and His Influence Through the Ages* (1957; originally published in French, 1956), and *Saint Augustin et la fin de la culture antique,* 4th ed. (1958); and EUGÈNE PORTALIÉ, *A Guide to the Thought of Saint Augustine* (1960, reprinted 1975). His philosophy is considered in JAKOB BARION, *Plotin und Augustinus* (1935); and ÉTIENNE GILSON, *The Christian Philosophy of St. Augustine* (1960; 2nd French ed., 1943). His political theory and view of history, especially as propounded in *De civitate Dei,* is the subject of REGINALD H. BARROW, *Introduction to St. Augustine: The City of God* (1950); JOHN H.S. BURLEIGH, *The City of God* (1949); HERBERT A. DEANE, *The Political and Social Ideas of St. Augustine* (1963); and GORDON L. KEYES, *Christian Faith and the Interpretation of History: A Study of St. Augustine's Philosophy of History* (1966). See also ROBERT A. MARKUS (ed.), *Augustine: A Collection of Critical Essays* (1972); and ROBERT E. MEAGHER, *An Introduction to Augustine* (1978), an anthology of passages extracted from Augustine's writings, with commentary.

Theology: For general accounts of Augustine's theology, see JOHN BURNABY, *Amor Dei: A Study of the Religion of St. Augustine* (1938, reprinted 1960); HENRI DE LUBAC, *Augustinianism and Modern Theology* (1969; originally published in French, 1965); and E.A. TESELLE, *Augustine the Theologian* (1970); and PAUL HENRI, *The Path to Transcendence: From Philosophy to Mysticism in Saint Augustine* (1981; originally published in French, 1938).

Special topics: For Christology, see TARSICIUS J. VAN BAVEL, *Recherches sur la Christologie de Saint Augustin* (1954); for the Eucharist, GASTON LECORDIER, *La Doctrine de l'eucharistie chez S. Augustin* (1930); for biblical exegesis, MAURICE PONTET, *L'Exégèse de S. Augustin, prédicateur* (1946); for predestination and grace, HENRI RONDET, *Essais sur la théologie de la grâce* (1964).

(Jo.Bu./Ed.)

Augustus

Gaius Octavius, subsequently known as Gaius Julius Caesar Octavianus and still later as Augustus or Caesar Augustus, was the first Roman emperor, following the republic, which had been finally destroyed by the dictatorship of Julius Caesar, his great-uncle and adoptive father. His autocratic regime is known as the principate because he was the *princeps,* the first citizen, at the head of that array of outwardly revived republican institutions that alone made his autocracy palatable. With unlimited patience, skill, and efficiency, he overhauled every aspect of Roman life and brought durable peace and prosperity to the Greco-Roman world.

Gaius Octavius was born on September 23, 63 BC, of a prosperous family that had long been settled at Velitrae (Velletri), southeast of Rome. His father, who died in 59 BC, had been the first of the family to become a Roman senator and was elected to the high annual office of the praetorship, which ranked second in the political hierarchy to the consulship. Gaius Octavius' mother, Atia, was the daughter of Julia, the sister of Julius Caesar; and it was Caesar who launched the young Octavius in Roman public life. At the age of 12 he made his debut by delivering the funeral speech for his grandmother Julia. Three or four years later he received the coveted membership of the board of priests (*pontifices*). In 46 he accompanied Caesar, now dictator, in his triumphal procession after his victory in Africa over his opponents in the Civil War; and in the following year, in spite of ill health, he joined the dictator in Spain. He was at Apollonia (now in Albania), completing his academic and military studies when, in 44 BC, he learned that Julius Caesar had been murdered.

Rise to power. Returning to Italy, he was told that in his will Caesar had adopted him as his son and had made him his chief personal heir. He was only 18 when, against the advice of his stepfather and others, he decided to take up this perilous inheritance and proceeded to Rome. Mark Antony (Marcus Antonius), Caesar's chief lieutenant, who had taken possession of his papers and assets and had expected that he himself would be the principal heir, refused to hand over any of Caesar's funds, forcing Octavius to pay the late dictator's bequests to the Roman populace from such resources as he could raise. Caesar's assassins, Brutus and Cassius, ignored him and withdrew to the east. Cicero, the famous orator who was one of Rome's principal elder statesmen, hoped to make use of him but underestimated his abilities.

Celebrating public games, instituted by Caesar, to ingratiate himself with the city populace, Octavius succeeded

Augustus, bronze sculpture from Meroe, Sudan, 1st century AD. In the British Museum.

By courtesy of the trustees of the British Museum

in winning considerable numbers of the dictator's troops to his own allegiance. The Senate, encouraged by Cicero, broke with Antony, called upon Octavius for aid (granting him the rank of senator in spite of his youth), and joined the campaign of Mutina (Modena) against Antony, who was compelled to withdraw to Gaul. When the consuls who commanded the Senate's forces lost their lives, Octavius' soldiers compelled the Senate to confer a vacant consulship on him. Under the name of Gaius Julius Caesar he next secured official recognition as Caesar's adoptive son. Although it would have been normal to add "Octavianus" (with reference to his original family name), he preferred not to do so. Today, however, he is habitually described as Octavian (until the date when he assumed the designation Augustus).

The Second Triumvirate

Octavian soon reached an agreement with Antony, as well as with another of Caesar's principal supporters, Lepidus, who had succeeded him as chief priest. On November 27, 43 BC, the three men were formally given a five-year dictatorial appointment as triumvirs for the reconstitution of the state (the Second Triumvirate—the first having been the informal compact between Pompey, Crassus, and Julius Caesar). The east was occupied by Brutus and Cassius, but the triumvirs divided the west among themselves. They also drew up a list of "proscribed" political enemies, and the consequent executions included 300 senators (one of whom was Antony's enemy Cicero) and 2,000 members of the class immediately below the senators, the equites or knights. Julius Caesar's recognition as a god of the Roman state in January 42 BC enhanced Octavian's prestige as son of a god.

He and Antony crossed the Adriatic and under Antony's leadership (Octavian being ill) won the two battles of Philippi against Brutus and Cassius, both of whom committed suicide. Antony, the senior partner, was allotted the east (and Gaul); and Octavian returned to Italy, where difficulties caused by the settlement of his veterans involved him in the Perusine War (decided in his favour at Perusia, the modern Perugia) against Antony's brother and wife. In order to appease another potential enemy, Sextus Pompeius (Pompey the Great's son), who had seized Sicily and the sea routes, Octavian married Sextus' relative Scribonia (though before long he divorced her for personal incompatibility). These ties of kinship did not deter Sextus, after the Perusine War, from making overtures to Antony; but Antony rejected them and reached a fresh understanding with Octavian at the treaty of Brundisium, under the terms of which Octavian was to have the whole west (except for Africa, which Lepidus was allowed to keep) and Italy, which, though supposedly neutral ground, was in fact controlled by Octavian. The east was again to go to Antony, and it was arranged that Antony, who had spent the previous winter with Queen Cleopatra in Egypt, should marry Octavian's sister Octavia. The peoples of the empire were overjoyed by the treaty, which seemed to promise an end to so many years of civil war. In 38 BC Octavian formed a significant new link with the aristocracy by his marriage to Livia Drusilla.

But a reconciliation with Sextus Pompeius proved abortive, and Octavian was soon plunged into serious warfare against him. When his first operations against Sextus' Sicilian bases proved disastrous, he felt obliged to make a new compact with Antony at Tarentum (Taranto) in 37 BC. Antony was to provide Octavian with ships, in return for troops Antony needed for his forthcoming war against the empire's eastern neighbour Parthia and its Median allies. Antony handed over the ships, but Octavian never sent the troops. The treaty also provided for renewal of the Second Triumvirate for five years, until the end of 33 BC.

Military successes. In the following year the balance of power began to change: whereas Antony's eastern expedition failed, Octavian's fleet, commanded by his former schoolmate, Marcus Agrippa, who, although unpopular with the influential nobles, was an admiral of genius, totally defeated Sextus Pompeius off Cape Naulochus (Venetico) in Sicily. At this point the third triumvir, Lepidus, seeking to contest Octavian's supremacy in the west by force, was disarmed by Octavian, deprived of his triumviral office, and forced into retirement. Ignoring Antony's right to settle his own veterans in Italy and recruit fresh troops, Octavian discharged many legionaries and founded settlements for them. His deliberate rivalry with Antony for the eventual mastership of the Roman world became increasingly apparent. Octavian's marriage two years earlier had begun to win over some of the nobles who had previously been Antony's supporters. Octavian also launched elaborate religious and patriotic publicity, centring on the classical god of order, Apollo, in contrast to Antony's more un-Roman patron, Dionysus (Bacchus). In addition, Octavian had started to prefix his name with the designation "Imperator," to suggest that he was the commander *par excellence;* and now, although he continued to use his triumviral powers, he omitted all reference to them from his coins, gradually concentrating on the plain, emotive name "Caesar Son of a God."

Retirement of Lepidus

But if Octavian was to compete with Antony's military seniority, successes in a foreign war were necessary; and so Octavian between 35 and 33 BC fought three successive campaigns in Illyricum and Dalmatia (parts of the modern Yugoslavia) in order to protect the northeastern approaches of Italy. With the help of Agrippa, he also lavished large sums on the adornment of Rome. When Octavian fomented public clamour against Antony's territorial gifts to Cleopatra, it was clear that a clash between the two men was imminent.

In 32 BC the triumvirate had officially ended, and Octavian, unlike Antony, professed no longer to be employing its powers. Amid a virulent exchange of propaganda, Antony divorced Octavia, whereupon her brother Octavian seized Antony's will and claimed to find in it damaging proofs of Cleopatra's power over him. Each leader induced the populations under his control to swear formal oaths of allegiance to his own cause. Then, in spite of grave discontent aroused by his exactions in Italy, Octavian declared war—not against Antony but gainst Cleopatra.

Defeat of Antony and Cleopatra

Accompanied by her, Antony had brought up his fleet and army to guard strongpoints along the coast of western Greece; but in 31 BC Octavian dispatched Agrippa very early in the year to capture Methone, at the country's southwestern tip. His enemies were taken by surprise; and after Octavian himself arrived—leaving his Etruscan friend and adviser Maecenas in charge of Italy—he and Agrippa soon shut Antony's fleet inside the Gulf of Ambracia (Arta). At the Battle of Actium Antony tried to extricate his ships in the hope of continuing the fight elsewhere. Though Cleopatra and then Antony succeeded in getting away, only a quarter of their fleet was able to follow them. She and Antony fled to Egypt and committed

suicide when Octavian captured the country in the following year. Executing Cleopatra's son Ptolemy XV Caesar (Caesarion)—whose father she had claimed was Caesar—he annexed Egypt and retained it under his direct control.

The seizure of Cleopatra's treasure enabled him to pay off his veterans and made him finally master of the entire Greco-Roman world. From this point on, by a long and gradual series of tentative, patient measures, he established the Roman principate, a system of government that enabled him to maintain, in all essentials, absolute control. Gradually reducing his 60 legions to 28, he retained approximately 150,000 legionaries, mostly Italian, and supplemented them by about the same number of auxiliaries drawn from the provinces. A permanent bodyguard (the Praetorians), based on the bodyguards maintained by earlier generals, was stationed partly in Rome and partly in other Italian towns. A superb network of roads was created to maintain internal order and facilitate trade; and an efficient fleet was organized to police the Mediterranean. In 28 BC Octavian and Agrippa held a census of the civil population, the first of three during the reign. They also reduced the Senate from about 1,000 to 800 (later 600) compliant members; and Octavian was appointed its president.

Master of the Greco-Roman world

Government and administration. Remembering, however, that Caesar had been assassinated because of his resort to naked power, Octavian realized that the governing class would welcome him as the terminator of civil war only if he concealed his autocracy beneath provisions avowedly harking back to republican traditions. From 31 until 23 BC the constitutional basis of his power remained a continuous succession of consulships, but in January 27 BC he ostensibly "transferred the State to the free disposal of the Senate and people," earning the misleading, though outwardly plausible, tribute that he had restored the republic. At the same time he was granted a ten-year tenure of an area of government (*provincia*) comprising Spain, Gaul, and Syria, the three regions containing the bulk of the army. The remaining provinces were to be governed by proconsuls appointed by the Senate in the old republican fashion. Octavian, however, believed that his supreme prestige—crystallized in the meaningful term *auctoritas*—safeguarded him against any defiance by these personages; and he was indeed able, more or less indirectly, to influence their appointments, just as he was able (on the rare occasions when he regarded it as desirable) to influence the appointments to the consulships and other metropolitan offices that continued to exist in "republican" fashion.

Autocracy in republican guise

Four days after these measures, his name Caesar, acquired through adoption in Julius' will, was supplemented by "Augustus," an appellation with an antique religious ring, believed to be linked etymologically with *auctoritas* and with the ancient practice of augury. The word *augustus* was often contrasted with *humanus;* its adoption as the title representing the new order cleverly indicated, in an extraconstitutional fashion, his superiority over the rest of mankind. With the aid of writers such as Virgil, Livy, and Horace, all of whom in their different ways shared the same ideas, he showed his patriotic veneration of the old Italian faith by reviving many of its ceremonials and repairing numerous temples.

Military operations continued in many frontier areas. In 25, recalcitrant Alpine tribes were reduced, and Galatia (central Asia Minor) was annexed. Mauretania, on the other hand, was transferred from Roman provincial status to that of a client-kingdom, for such dependent monarchies, as in the later republic, bore a considerable part of the burden of imperial defence. Augustus himself visited Gaul and directed part of a campaign in Spain until his health gave out; in 23 he fell ill again and seemed on the point of death. Feeling, amid reports of conspiracies, that new constitutional steps were necessary, he proceeded to terminate his series of consulships in favour of a power (*imperium majus*) which was separated altogether from office and its practical inconveniences. This power raised him above the proconsuls; it was never referred to on the official coinage or in Augustus' political testament but was intended to be exercised mainly in emergencies and on personal visits. He was also awarded the power of a tribune (*tribunicia potestas*) for life. Earlier, he had accepted certain privileges of a tribune. The full power he now assumed carried with it practical advantages, notably the right to convene the Senate. But, more particularly, the office of a tribune, because of the ancient character of the annually elected tribunes of the people as defenders of the *plebs,* surrounded him with a "democratic" aura, one which, perhaps, was needed all the more because Augustus himself—while admittedly supporting the interests of poorer people by a great extension of the right of judicial appeal—tended to back the established classes as the keystone of his system.

Agrippa, too, was granted superiority over proconsuls, presumably in order to ensure that the armies would be in safe hands in case one of Augustus' recurrent illnesses proved fatal. The next to die, however, was the emperor's young nephew Marcellus, who had been married to his daughter Julia and might eventually have been envisaged as his successor. In the same year, 23 BC, Agrippa was sent out to the east as deputy *princeps;* two years later he became Julia's second husband. Meanwhile Augustus himself travelled in Sicily, Greece, and Asia (22–19). Important reorganizations were put into effect wherever he went; and immense satisfaction was caused by an agreement in 20 BC with Parthia, under which the Parthians recognized Rome's protectorate over Armenia and returned the legionary standards captured from Crassus 33 years earlier. In 19 Agrippa completed the subjugation of Spain. In this year there was some adjustment of Octavian's powers to allow him to exercise them more freely in Italy, and the two following years witnessed social legislation attempting to encourage marriage, regulate penalties for adultery, and reduce extravagance. In 17 there were resplendent celebrations of ancient ritual, known as the Secular Games, to purify the Roman people of their past sins and provide full religious inauguration of the new age.

Although the principate was not an office which could be automatically handed on, Augustus seemed to be indicating his views regarding his ultimate successor when he adopted the two sons of his daughter Julia, boys aged three and one who were henceforward known as Gaius Caesar and Lucius Caesar. Their father Agrippa, whose powers had been renewed along with his master's, returned to the east. But now Augustus also gave important employment to his stepsons—his wife Livia's sons by her former marriage—Tiberius and Drusus the elder. Proceeding across the Alps, they annexed Noricum and Raetia, comprising large parts of what are now Switzerland, Austria, and Bavaria, and extended the imperial frontier from Italy to the upper Danube (16–15).

Reform of the administration

It was probably during these years that an executive, or drafting, committee (*consilium*) of the Senate was established in order to help Augustus to prepare senatorial business. His administrative burden was also lightened by the expansion of his own staff (knights, who could also now rise to a number of key posts, and freedmen) to form the beginnings of a civil service, which had never existed before but was destined to become an essential feature of the imperial system. Gradually, too, a completely reformed administrative structure of Rome, Italy, and the whole empire was evolved. The financial system that made this possible was evidently far more effective than anything the empire had ever seen until then. The system was based on the central treasury (*aerarium*), but the details of its relationship with the treasuries of the provinces, and particularly the *provincia* of Augustus, are still imperfectly understood, partly because, although the emperor proudly recorded his gifts to the central treasury, he did not report what funds passed in the opposite direction.

The taxation providing these resources apparently included two main direct taxes: a poll tax (*tributum capitis*), paid in some provinces by all adults and in others by adult males only, and a land tax (*tributum soli*). There were also indirect taxes, which (as in the past) were farmed out to contractors because their yield was unpredictable and the embryonic civil service lacked the resources to handle them. The republican customs dues continued; but its rates were low enough not to hamper trade, which, in the peaceful conditions created by Augustus, flourished in

wholly unprecedented fashion. Industries did not exist on a very large scale, but commerce was greatly stimulated by a sweeping reform and expansion of the Roman coinage. Gold and silver pieces, their designs reflecting many facets of imperial publicity, were issued in great quantities at a number of widely distributed mints; and from about 19 (or perhaps 23) BC onward the absence of bronze token coinage, which had been sparse for many decades, was remedied by the creation of abundant mintages in two bright new metals, yellow brass and red copper. In the West, the principal mint for these pieces, besides Rome, was Lugdunum (Lyon), whose coins displayed a view of the Altar of Rome and Augustus that formed a model for other provincial capitals. The Roman citizen colonies of the West, many of them established by Augustus to settle his veterans, supplemented this output by their own local coinages, and in the East, particularly Asia Minor and Syria, numerous Greek cities were also allowed to issue small change.

Expansion of the empire. The death in 12 BC of Lepidus, who had lived on in retirement for 24 years, enabled Augustus finally to succeed him as the official head of the Roman religion, the chief priest (*pontifex maximus*). In the same year, however, another death came as a severe blow to him, for Agrippa, too, died. Augustus compelled his widow Julia to marry Tiberius against the wishes of both of them. During the next three years, however, Tiberius was away in the field, reducing Pannonia (Yugoslavia and Hungary) up to the middle Danube, while his brother Drusus crossed the Rhine frontier and invaded Germany as far as the Elbe, where he died in 9 BC. In the following year Augustus lost another of his intimates, Maecenas, who had been the adviser of his early days and was an outstanding patron of letters.

Tiberius, who replaced Drusus in Germany, was elevated in 6 BC to a share in his stepfather's tribunician power. But shortly afterward he went into retirement on the island of Rhodes. This was attributed to jealousy of his stepnephew Gaius Caesar, who was introduced to public life with a great fanfare in the following year; and the same compliments were paid to his brother Lucius in 2 BC, the year in which Augustus received his climactic title "father of the country" (*pater patriae*). Gaius was sent to the East and Lucius to the West. Both, however, soon died, Lucius in AD 2 and Gaius in 4. Tiberius returned home in 2, and in 4 Augustus realized that he had to make him his heir.

Tiberius named heir

He adopted Tiberius as his son, who in turn was required to adopt Germanicus, the son of his brother Drusus. The powers conferred upon Tiberius made him almost Augustus' own equal in everything except prestige.

Tiberius' next task was to consolidate the invasion and provincial organization of Germany (4–5); and now that the Elbe was the frontier instead of the Danube, Augustus instructed him to establish a shorter frontier line incorporating Bohemia, which had become the nucleus of a German (Marcomannic) empire ruled by King Marobodduus. An invasion of Bohemia, therefore, was planned, and had already been launched, from two directions, when news came in 6 that Pannonia and Illyricum had revolted. It took three years for the rebellion to be put down; and this had only just been completed when Arminius raised the Germans against their Roman governor Varus and destroyed him and his three legions. As Augustus could not readily replace the troops, the annexation of western Germany and Bohemia was postponed indefinitely; Tiberius and Germanicus were sent to consolidate the Rhine frontier.

Although Augustus was now feeling his age, these years in association with Tiberius were marked by administrative innovations: the annexation of Judaea in AD 6 (its client king Herod the Great had died 10 years previously); the establishment at Rome (in the same year) of a fire brigade with police duties—supplemented seven years later by a regular police force (*cohortes urbanae*); the creation of a military treasury (*aerarium militare*) to defray soldiers' retirement bounties from taxes; and the conversion of the hitherto occasional appointment of prefect of the city (*praefectus urbi*) into a permanent office (AD 13). When, in the same year, the powers of Augustus were renewed for 10 years—such renewals had been granted at intervals throughout the reign—Tiberius was made his equal in every constitutional respect. In April, Augustus deposited his will at the House of the Vestals in Rome. It included a summary of the military and financial resources of the empire (*breviarium totius imperii*) and his ingenious political testament known as the "Res Gestae Divi Augusti" ("Achievements of the Divine Augustus"). The best preserved copy of the latter document is still to be seen on the walls of the Temple of Rome and Augustus at Ankara, Turkey (the Monumentum Ancyranum). In 14 Tiberius was due to leave for Illyricum but was recalled by the news that Augustus was gravely ill. He died on August 19, and on September 17 the Senate enrolled him among the gods of the Roman state. By that time Tiberius had succeeded him as the second Roman emperor, though the formalities involved in the succession proved embarrassing both to himself and to the Senate because the "principate" of Augustus had not, constitutionally speaking, been heritable or continuous. Like other emperors, Tiberius assumed the designation "Augustus" as an additional title of his own. Agrippa Postumus, who had been named his co-heir but was later banished, was put to death. The order to kill him may already have been given by Augustus, but this is not certain.

Personality and achievement. Augustus was one of the great administrative geniuses of history. The gigantic work of reorganization that he carried out in every field of Roman life and throughout the entire empire not only transformed the decaying republic into a new, monarchic regime with many centuries of life ahead of it but created a durable Roman peace, based on easy communications and flourishing trade. It was this Pax Romana that ensured the survival and eventual transmission of the classical heritage, Greek and Roman alike, and provided the means for the diffusion of Judaism and Christianity (Jesus Christ was born during Augustus' reign). Although his regime was an autocracy, Augustus, being a tactful and imaginative master of propaganda of many kinds, knew how to cloak that autocracy in traditionalist forms that would satisfy a warworn generation—perhaps, most of all, the upper bourgeoisie immediately below the leading nobility, since it was they who benefitted from the new order more than anyone. He was also able to win the approbation, through the patronage of Maecenas, of some of the greatest writers the world has ever known, including Virgil, Horace, and Livy.

Pax Romana

Their enthusiasm was partly due to Augustus' conviction that the Roman peace must be under occidental, Italian control. This was in contrast to the views of Antony and Cleopatra, who had envisaged some sort of Greco-Roman partnership such as only began to prevail three or four centuries later. Augustus' narrower view, although modified by an informed admiration of Greek civilization, was based on his small-town Italian origins. These were also partly responsible for his patriotic, antiquarian attachment to the ancient religion and for his puritanical social policy.

Augustus was a cultured man, the author of a number of works (all lost): a pamphlet against Brutus, an exhortation to philosophy, an account of his own early life, a biography of Drusus, poems, and epigrams. The conventional view of his character distinguishes between his cruelty in early years and his mildness in later life. But there was not so much need for cruelty later on, and when it was needed (notably in the suppression of alleged plots), he was still ready to apply it. It is probable that nothing short of this degree of political ruthlessness could have achieved such enormous results. His domestic life, however, was simple and homespun. Within his family, the successive deaths of those he had earmarked as his successors or helpers caused him much sadness and disappointment. His devotion to his wife Livia Drusilla remained constant, though, like other Romans, he was unfaithful. His surviving letters show kindliness to his relations. Yet he exiled his daughter Julia for offending against his public moral attitudes, and he exiled her daughter by Agrippa for the same reason; he also exiled the son of Agrippa and Julia, Agrippa Postumus, though the suspicion that he later had him killed is unproved. As for Augustus' male relatives who were his

helpers, he was loyal to them but drove them as hard as he drove himself. He needed them because the burden was so heavy, and he especially needed them in the military sphere because he was not a great commander. In Agrippa and Tiberius and a number of others he had men who supplied this deficiency, and although, on his deathbed, he is said to have advised against the further expansion of the empire, he himself, with their assistance, had expanded its frontiers in many directions.

His physical condition was subject to a host of ills and weaknesses, many of them recurrent. Indeed, in his early life, particularly, it was only his indomitable will that enabled him to survive—a strange preliminary to an unprecedented and unequalled life's work. His appearance is described by the biographer Suetonius.

> He was unusually handsome and exceedingly graceful at all periods of his life, though he cared nothing for personal adornment. His expression, whether in conversation or when he was silent, was calm and mild . . . He had clear, bright eyes, in which he liked to have it thought that there was a kind of divine power, and it greatly pleased him, whenever he looked keenly at anyone, if he let his face fall as if before the radiance of the sun. His teeth were wide apart, small and ill-kept; his hair was slightly curly and inclining to golden; his eyebrows met . . . His complexion was between dark and fair. He was short of stature, but this was concealed by the fine proportion and symmetry of his figure, and was noticeable only by comparison with some taller person standing beside him.

Augustus' countenance proved a godsend to the Greeks and Hellenized easterners who were the best sculptors of the time, for they elevated his features into a moving, never to be forgotten imperial type, which Napoleon's artists, among others, keenly emulated. The contemporary portrait busts of Augustus, echoed on his coins, formed part of a significant renaissance of the arts in which Italic and Hellenic styles were discreetly and brilliantly blended. Still extant at Rome are the severe yet delicate reliefs of the Ara Pacis ("Altar of Peace"), depicting a religious procession in which the national leaders are taking part; there are also scenes from the Roman mythology. The altar was dedicated by the Senate and people of Rome in 13 BC to commemorate the pacification of Gaul and Spain.

The architectural masterpieces of the time were also numerous; and something of their monumental grandeur and classical purity can be seen today in the remains of the Theatre of Marcellus at Rome and of the massive Forum of Augustus, flanked by colonnades and culminating in the Temple of Mars the Avenger—the Avenger of Julius Caesar. Outside Rome, too, there are abundant memorials of the Augustan age; on either side of the Alps, for example, there are monuments to celebrate the submission and loyalty of the local tribes, an elegant arch at Segusio (Susa) and a square stone trophy, topped by a cylindrical drum, at La Turbie. From Livia's mansion on the outskirts of Rome, at Prima Porta, comes a reminder that not all the art of the day was formal and grand. For one of the rooms is adorned with wall paintings representing an enchanted garden; beyond a trellis are orchards and flower beds, in which birds and insects perch among the foliage. Augustus himself had no interest in personal luxury. Yet if ever he or his associates had any spare time, such were the rooms in which they spent it.

(M.Gr.)

BIBLIOGRAPHY. The principal ancient literary sources are GAIUS SUETONIUS TRANQUILLUS, *De Vita Caesarum,* in Latin, which describes the lives of the Roman emperors from Julius Caesar to Domitian; and books 52–56 of CASSIUS DIO COCCEIANUS, *Romaika,* a history of Rome, written in Greek. Both exist in several English translations, including, respectively, SUETONIUS, *The Twelve Caesars,* trans. by ROBERT GRAVES, rev. ed. (1979); and DIO CASSIUS, *Dio's Roman History,* trans. by EARNEST CARY, 9 vol. (1914–27, reprinted 1961), which includes the Greek text. *Inscriptions: Res Gestae Divi Augusti: The Achievements of the Divine Augustus,* ed. by P.A. BRUNT and JOHN M. MOORE (1967, reprinted 1979); VICTOR EHRENBERG and A.H.M. JONES (comps.), *Documents Illustrating the Reigns of Augustus and Tiberius,* 2nd ed. (1955, reprinted with addenda, 1976). *Coins:* CAROL H.V. SUTHERLAND, *Coinage in Roman Imperial Policy, 31 B.C.–A.D. 68* (1951, reprinted 1978); MICHAEL GRANT, *From Imperium to Auctoritas* (1946, reprinted 1969). *Art:* JOCELYN M.C. TOYNBEE, *The Art of the Romans* (1965); AXEL BOËTHIUS and J.B. WARD-PERKINS, *Etruscan and Roman Architecture* (1970). Modern sources include several books by RONALD SYME: *The Roman Revolution* (1939, reprinted 1974), a scholarly analysis of Augustus' creation of the Roman imperial system; *History of Ovid* (1978), which deals with Augustus' motives for Ovid's exile; and *Roman Papers,* 2 vol. (1979). Other works include ARNOLD H.M. JONES, *Augustus* (1970); MASON HAMMOND, *The Augustan Principate in Theory and Practice During the Julio-Claudian Period,* new ed. (1968); *The Cambridge Ancient History,* vol. 10, *The Augustan Empire, 44 B.C.–A.D. 70* (1934, reprinted 1966); DONALD C. EARL, *The Age of Augustus* (1968, reissued 1980); JOHN M. CARTER, *The Battle of Actium: The Rise and Triumph of Augustus Caesar* (1970); HOWARD H. SCULLARD, *From the Gracchi to Nero,* 5th ed. (1982), and *Roman Britain* (1979); JOHN C. STOBART, *The Grandeur That Was Rome,* 4th ed. (1961, reprinted 1971), a survey of Roman culture and civilization; HERMANN BENGTSON, *Kaiser Augustus: sein Leben und seine Zeit* (1981), an illustrated biography; HELMUT SIGNON, *Agrippa* (1978), which explores Marcus Vipsanius Agrippa's friendship and collaboration with Augustus; COLIN M. WELLS, *The German Policy of Augustus: An Examination of the Archaeological Evidence* (1972); MAX CARY and HOWARD H. SCULLARD, *A History of Rome Down to the Reign of Constantine,* 3rd ed. (1975), a standard account; and BARRY BALDWIN, *The Roman Emperors* (1980), a study based on primary sources.

ultural
eritage

Australia

Australia is located in the Southern Hemisphere between the Indian Ocean and the South Pacific, on an island continent that has been called both the Oldest Continent and the Last of Lands. Although possibly poetic, neither description is accurate. It is the oldest continent only in the senses that much of Australia is formed of rocks laid down during the Precambrian (4,600,000,000 to 570,000,000 years ago), and that it has altered relatively little since life first appeared on Earth. It is the last of lands only in the sense that it was the last continent (excluding Antarctica) to be discovered and explored by Europeans. Thousands of years before the explorers Abel Tasman and James Cook sailed into the South Pacific, the Aborigines had crossed the land bridge from Asia formed by the Malay Archipelago and had spread throughout the mainland and Tasmania. They remained, however, a sparse, primitive, and nomadic people. When Capt. Arthur Phillip of the British Royal Navy landed with the First Fleet at Botany Bay in 1788, the event that marks the true beginning of modern Australia, there were probably not more than 300,000 Aborigines altogether, and there was no man-made structure on the continent more solid than a bark hut.

The most striking characteristics of the vast, 3,000,000-square-mile (8,000,000-square-kilometre) landmass are its isolation, its low relief, and the aridity of much of its surface. Its isolation from other continents explains much of the strangeness of Australian plant and animal life; its low relief results from the long and extensive erosive action of the forces of wind, rain, and the heat of the Sun during the great periods of geological time when the continental mass was elevated well above sea level.

A member of the Commonwealth of Nations, the Commonwealth of Australia is a prosperous, independent nation united under one government. Australians are in many respects fortunate in that they do not share their continent—which is only a little smaller than the United States—with any other nation. Their nearest neighbours are Papua New Guinea and Indonesia to the north, the Polynesians and Melanesians of the Pacific Islands to the east, and New Zealand, which, like Australia, is a member of the Commonwealth of Nations, to the southeast. On the other hand, Australia is extremely remote from its two principal allies: it is 12,000 miles (19,000 kilometres) from Australia to Great Britain via the Indian Ocean and the Suez Canal; it is 7,000 miles across the Pacific to the West Coast of the United States.

Like Canada and the United States, contemporary Australia is a political federation with a central government (the Commonwealth) and six constituent states (New South Wales, Victoria, Queensland, South Australia, Western Australia, and Tasmania), each of which has its own government enjoying a limited sovereignty. There are also two internal territories: the Northern Territory was established as a self-governing territory in 1978, and the Australian Capital Territory, seat of the federal capital city Canberra, is administered directly by the Commonwealth, which also governs the external territories of Norfolk Island, Cocos (Keeling) Islands, Christmas Island, Ashmore and Cartier Islands, Coral Sea Islands, and Heard and McDonald Islands and claims the Australian Antarctic Territory. The Cocos Islands was a non-self-governing territory until 1984, when it was integrated with Australia following an act of self-determination approved by the Cocos Malay people. Papua New Guinea, which formerly was an Australian external territory, became an independent nation in 1975.

The article is divided into the following sections:

PHYSICAL AND HUMAN GEOGRAPHY

Geological history

A popular misconception about the observed peculiarities of Australia is that it is the oldest of continents. The statement is, in fact, unjustified: the cores of all of the continents are of approximately the same age, and rocks of all geological ages, of sedimentary as well as volcanic origin, are found in Australia. The misconception about its age stems partly from its relative paucity of surface relief, the result of relative stability and the absence of recent strong movements of the Earth's crust, rather than the old age of its rocks. The geologically young mountain-building forces, such as those that formed the Himalayas in Asia, the Alps in Europe, the Rocky Mountains in North America, and the Andes in South America, have bypassed Australia.

Australia resembles the continents of Eurasia and Africa in being strikingly different from the landmass of North America, which consists basically of an old core, the Canadian Shield, surrounded by mountain zones formed from belts of folded rocks of varying ages, all younger than the central core. Australia is strikingly asymmetric, with a folded belt of rocks bordering its east coast only and extending westward to no more than one-fourth of the continent's width.

Like the other continental masses, Australia is built of three types of structural units. In the west is the Western Australian Shield (Western Craton), a comparatively stable and resistant block consisting mainly of ancient rocks that have not been folded during at least the last 1,000,000,000 years. In the east is an ancient fold mountain belt, greatly eroded but relatively uplifted again. Between these two lies the platform-like area occupied by the central plains, a region where crystalline or folded rocks are overlain by relatively thin sequences of flat-lying or gently deformed sediments. Each of these major units is complex. The Western Australian Shield, for example, includes, in addition to extensive outcrops of crystalline rocks in the core area itself, blocks of sedimentary strata deposited in troughs and basins within this structural unit. The central platform consists of several basin structures. The Eastern Uplands include both upfaulted and depressed structural blocks, along with large, exposed, formerly molten intrusions (batholiths) of granite formed within the past 50,000,000 years, and extensive lava flows formed during the Cenozoic Era.

The outline of the Australian continent, unlike those of South America and Africa, and the Indian subcontinent, is not roughly triangular with an apex pointing southward, though some resemblance to this distinctive configuration of the other continents and subcontinents can be seen in the shape and position of Tasmania. This island is part of the continent if the base of the shallow peripheral submarine area known as the continental shelf is taken as its outer boundary. The continental shelf and the slope separating it from the ocean depths are moderately wide in the south and southwest and spectacularly wide in the northwest from Exmouth Gulf to the Timor Sea, where the distance of the slope base from the shore exceeds 300 miles (470 kilometres), and the line marking depths that exceed 650 feet (200 metres) is more than 250 miles from the shore. To the north of Australia, the continental shelf extends to the island of New Guinea. The shelf is narrow along the east coast, where the Great Barrier Reef marks its edge. A trough, with depths of about 3,000 feet (900 metres), separates the Great Barrier Reef from the largely submerged Queensland Plateau, where only a few small islands and reef tops reach sea level. Off Brisbane and Sydney, the continental shelf is only some 30 miles (48 kilometres) wide, and the slope plunges steeply to the floor of the Tasman Sea.

The discovery of vast and varied mineral deposits in Australia stimulated intensive studies of its geological history and structure. At the same time, new developments in general geological theory, including the concepts of continental drift and ocean-floor spreading, have led to renewed speculation regarding possible former land connections between Australia and other continents, particularly Antarctica. As a result, the answer to the problem of how the history of the Australian continent fits into the picture of the development of the Earth's crust as a whole will have to come from a deeper understanding not only of the geology of the continent but also of that of the ocean surrounding it.

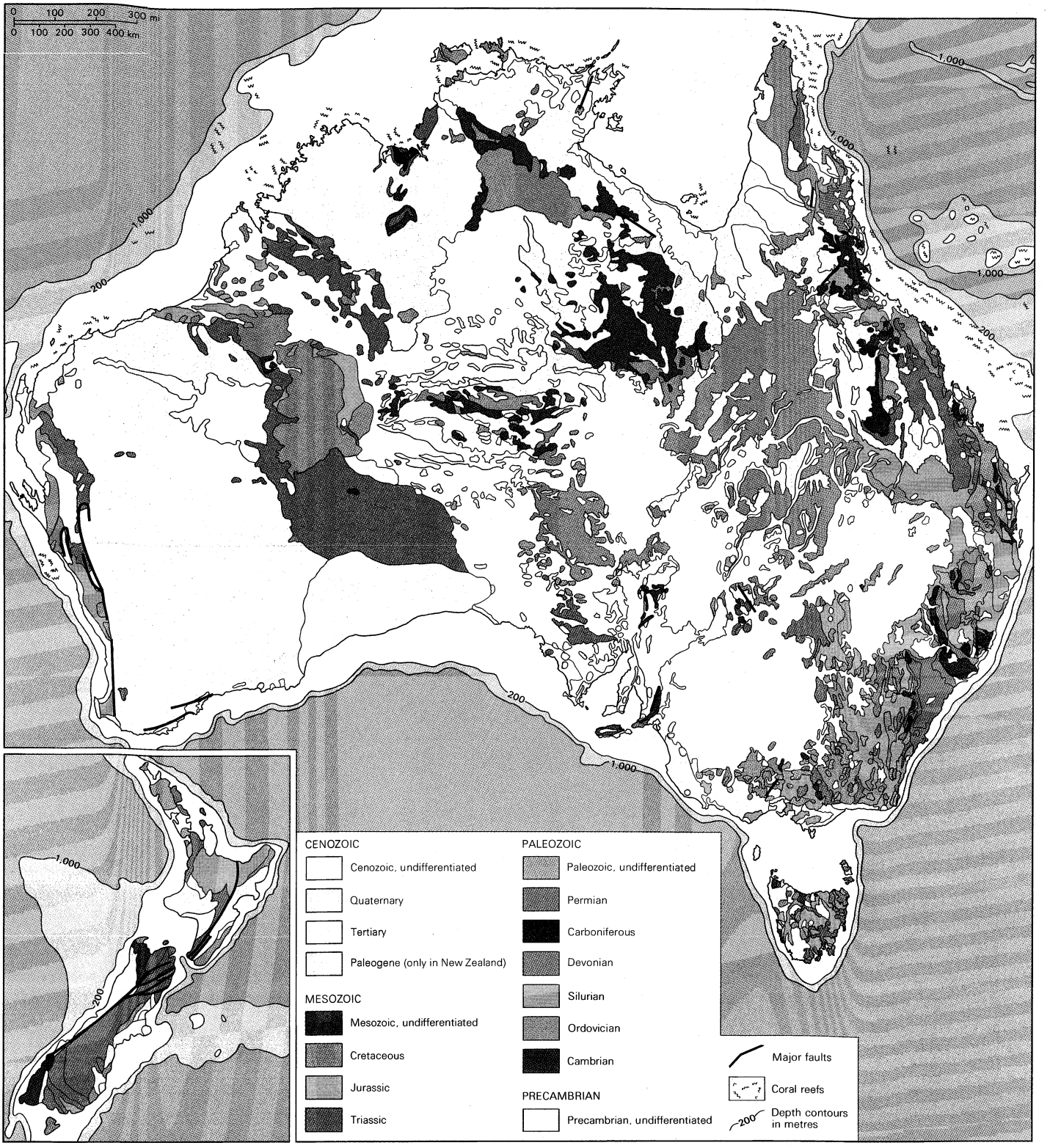
THE OLDEST ROCKS: THE PRECAMBRIAN

The oldest rocks of the continent occur in the southwest, where they form part of the Western Australian Shield, an area underlain by rocks that have not been folded since early Precambrian (Archaean) time. These oldest rocks are now thought to be about 3,000,000,000 years old. The Yilgarn nucleus of the Western Australian Shield extends over 230,000 square miles (600,000 square kilometres) and consists mostly of granite, with belts of altered sedimentary rocks (metasediments) and greenstones, the latter including ancient rocks rich in economically important mineral deposits, as well as basic intrusions of formerly molten rocks. Similar to metal-bearing greenstone rock belts occurring in the nuclei of the ancient shield areas of other continents, these Australian formations are being studied in the hope that the origin of these rock belts and their ore bodies will be explained in terms of the state of the Earth's crust and mantle at the beginning of geological time. The conditions under which the continental crust was formed and then reworked are not yet known, but geologists believe that some of the peculiarities of ancient rocks, their structure, and their composition can be explained only by assuming that the Earth's crust was originally much thinner than it is today.

Formation of the continental crust

The Yilgarn nucleus does not extend as far as the west and south coasts of Western Australia, where younger Precambrian rocks occur. Those on the west coast are separated by the Darling Fault, which forms the western margin of the shield near Perth, and by the Perth Basin. The Frazer Fault apparently marks the boundary between rocks of the main shield area that were metamorphosed by heat and pressure more than 2,500,000,000 years ago and others to the southeast that were similarly affected less than 2,000,000,000 years ago. Each of these major fault lines is about 600 miles (960 kilometres) long and is considered as a major lineament in the continental structure. The older rocks appear again in the Pilbara nucleus in the northwest, where some granites are more than 3,000,000,000 years old and where altered volcanic, greenstone, and sedimentary rocks also occur.

Large platforms and basins separate the Yilgarn and Pilbara nuclei of the Western Australian Shield from other occurrences of ancient rocks in the northwest and north of the continent and in central Australia. It is not yet known whether a major part of these other instances of

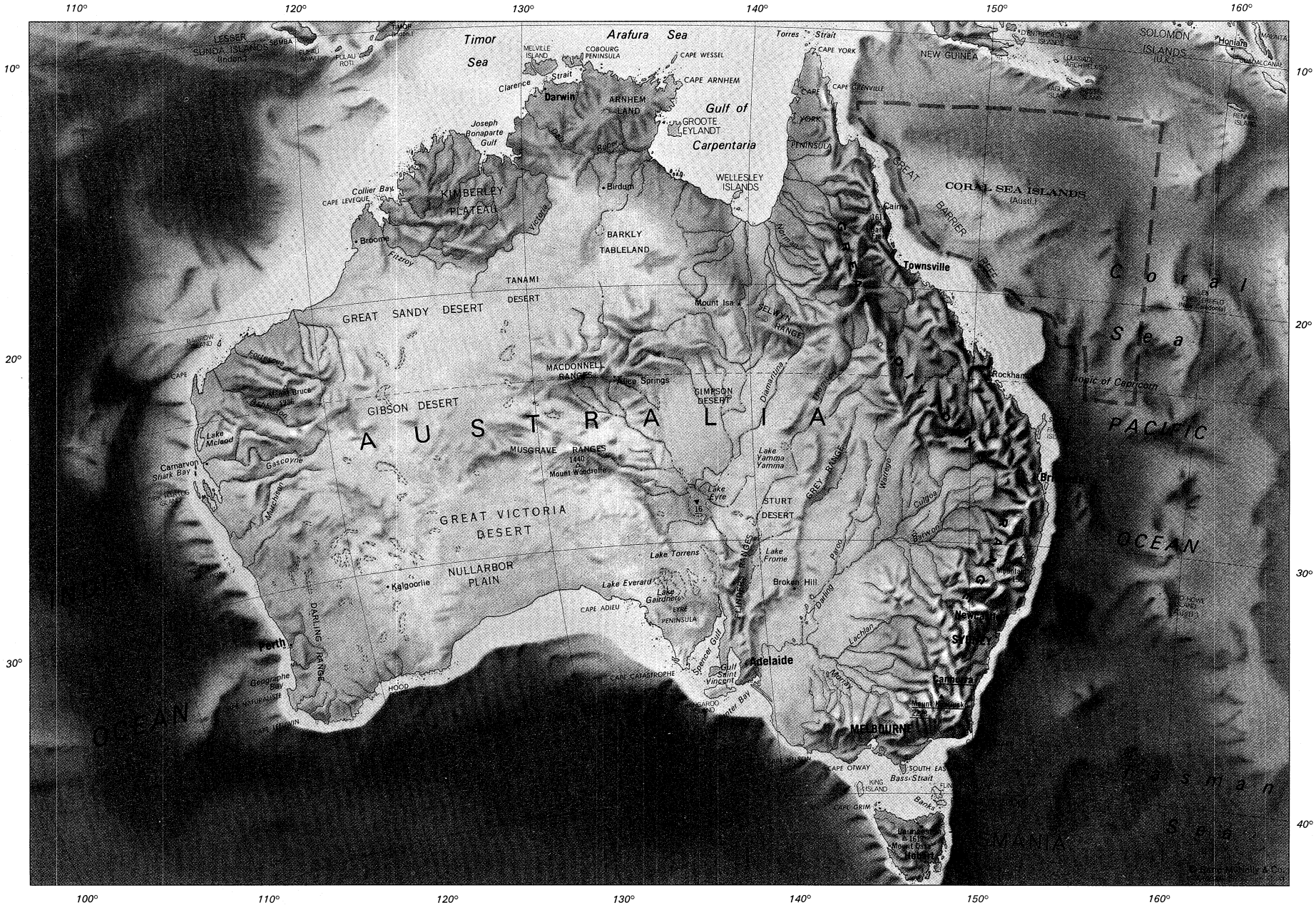


Geological structure of Australia and (inset) New Zealand.
By courtesy of the Department of National Development, Canberra

ancient rocks are as old as those that have been dated as Archaean in the west. Therefore, the earliest geographic configuration of the Australian continent cannot be reliably reconstructed; it could have extended northward and eastward far beyond the limits of the present Western Australian Shield.

Old (Archaean) rocks on the margins of the Kimberley Block form its basement; they occur also in the Katherine–Darwin area of the Northern Territory, in the Rum Jungle Complex. No definite datings of rocks in the east indicate ages as great as those of the oldest shield rocks, although part of the Arunta Complex, north of Alice Springs in the centre of Australia, may be approximately 2,250,000,000 years old.

The post-Archaean geological history of Australia can be seen in a clearer framework; the general terms Early, Middle, and Late Proterozoic, or the regional terms Nullaginian, Carpentarian, and Adelaidean, are used to identify intervals in the time scale of this period and are documented by sedimentary rock sequences.



AUSTRALIA

Size of symbol indicates relative size of town • ⊙

Elevations and depths in metres

0 200 400 600 800 km

0 200 400 600 mi

Post-Archaean regional rock units

The Nullaginian System comprises platform cover south of the Pilbara nucleus, including the vast banded iron ore formations of the Hamersley Ranges (notably the Mt. Bruce supergroup, which also contains volcanic rocks dated at 2,200,000,000 to 1,950,000,000 years ago) and the rocks in the east Kimberley area. The Carpentarian System comprises a belt of sedimentary rocks—the McArthur Basin, southwest of the Gulf of Carpentaria—that may be 1,500,000,000 to 1,800,000,000 years old; younger granites in the Katherine–Darwin region; and the sediments and intrusive granites of the Mt. Isa–Cloncurry region. The rocks of these regions are rich in mineral deposits. In South Australia, the basement rocks (gneisses) of Eyre Peninsula in the west and the Mt. Painter Complex in the northeast, and extending to Broken Hill in New South Wales (Willyama Block), may be of the same age as those of the Carpentarian System, although the iron formations of the Middleback ranges have been considered to be as old as the early Proterozoic iron formations found in the Pilbara region.

The Adelaidean System is made up of rocks of Late Precambrian age. Consolidation of the Musgrave Block in central Australia and intrusion of various molten rocks took place during the time interval 1,400,000,000 to 1,000,000,000 years ago, forming an extension of the Western Australian Shield. At the same time, sedimentary rocks were laid down in the Adelaide Geosyncline, a vast downwarping of the Earth's surface lying to the southeast. Sedimentation continued into mid-Cambrian time in a shallow marine trough extending from the coast south of Adelaide at least 400 miles (640 kilometres) northward, through the Mt. Lofty and Flinders ranges, and measuring 640 miles (1,025 kilometres) from west to east, from Lake Torrens to beyond Broken Hill. These rocks, with a total maximum thickness of more than 10 miles (16 kilometres), formed in an intermittently sinking trough in which the accumulating sediments were subjected to moderately strong folding and faulting in Late Precambrian and Early Paleozoic time. This Adelaidean sedimentary sequence has characteristic features of outstanding importance for Earth history: most of the sediments, which have remained unaltered, contain definite evidence of a major glacial period; they also contain the oldest known rich and varied assemblage of fossil animals, all of which were essentially soft-bodied. Rocks of similar age also occur in the Peake and Denison ranges west of Lake Eyre, in the Amadeus Basin, and on the margins of the Kimberley Block. In the area there is also clear evidence of a Precambrian glaciation. Its earlier phase has been dated at about 740,000,000 years, while the later phase was little older than 660,000,000 years. This is the likely time range of glaciation in the Adelaide Geosyncline and probably also of related glaciations in other continents: it was probably one of the most extensive periods of cold in the Earth's history. No definite Precambrian animal remains are known from rocks that have been identified as older than the onset of this glaciation; on the other hand, primitive aquatic plant fossils have been found abundantly in earlier Precambrian strata.

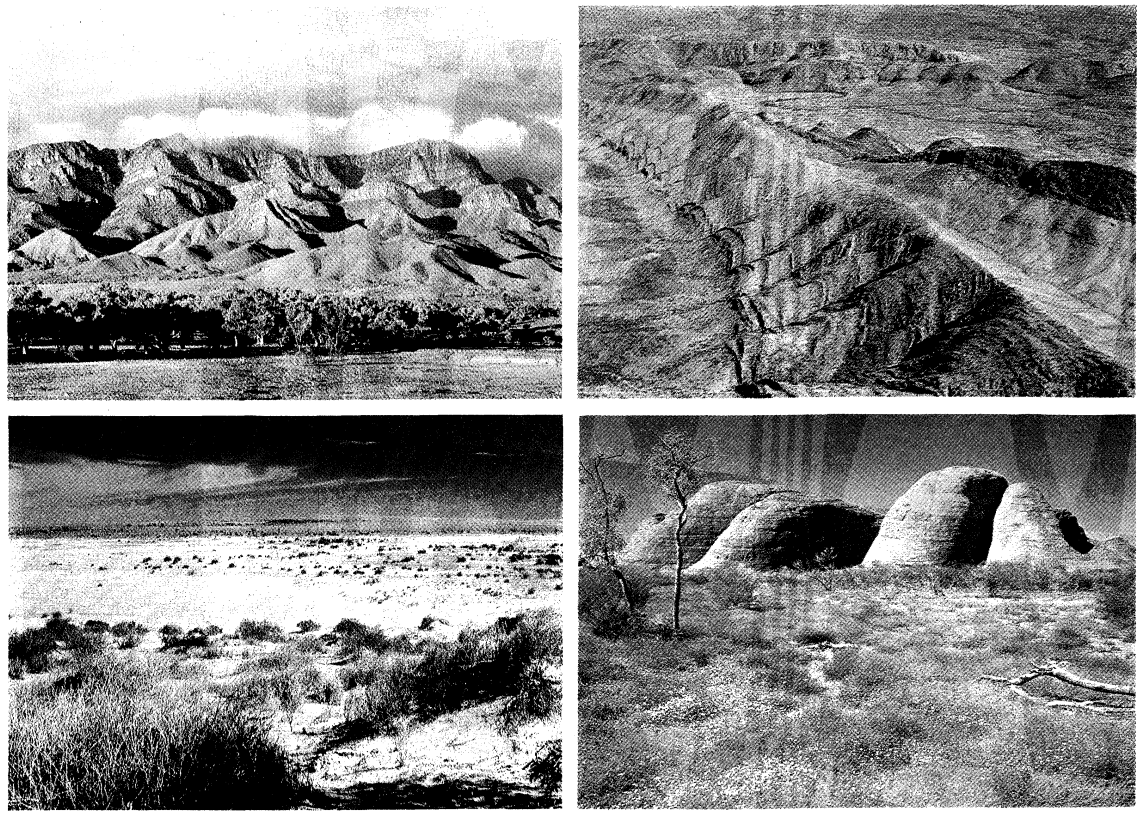
The Adelaide Geosyncline

PALEOZOIC HISTORY OF EASTERN AUSTRALIA

The Adelaide Geosyncline is often considered as a precursor of the more extensive fold belts of eastern Australia. These form another great downwarp known as the Tasman Geosyncline. While the rocks of the Adelaide Geosyncline were folded in Late Cambrian time, about 500,000,000 years ago, and elevated so that sedimentation ceased, the Tasman Geosyncline began to form in Cambrian time, the first period of the Paleozoic Era, which was to last for the next 300,000,000 years.

The Tasman Geosyncline

(Top left, top right) Photographic Library of Australia, (bottom left, bottom right) Shosta



Australian landscapes.
(Top left) St. Mary Peak, Flinders Range, South Australia. Mineral traces are evident in upper crests. (Top right) Macdonnell Ranges, west of Alice Springs, Northern Territory. The low parallel ridges, cut in folded strata, run practically unbroken for 150 miles. (Bottom left) Simpson Desert, central Australia, one of the most desolate areas of the continent, contains neither water nor habitation. In the foreground are sand humps covered with spinifex, the scattered vegetation characteristic of the desert. (Bottom right) Mt. Olga, a huge red sandstone monolith, rises 1,500 feet above the plain, south of Lake Amadeus in the Northern Territory.

Cambrian strata. Cambrian strata, containing marine fossils indicating their age, are known only from a few areas in eastern Australia: Tasmania, central Victoria, and northeast of Broken Hill in western New South Wales. It is possible that Cambrian rocks outcrop also on the coast of southern New South Wales. These few occurrences are insufficient to support any well-founded reconstruction of the geography of the Australian continent in Cambrian time, but they do indicate a fundamental difference in conditions of sedimentation in the east, where the rocks are similar to those deposited in deep geosynclinal troughs generally, and the large areas in southern, central, and northern Australia, including the Amadeus and Georgina basins, where the Cambrian rocks were mostly formed in shallow water. These differences are the expression of the relative stability of the western areas as compared with the greater geological mobility of eastern Australia during Paleozoic time, although any boundary between these two areas is likely to be gradational rather than a definite line.

An attempt at a generalized western margin of the Tasman Geosyncline might be drawn from the north Queensland coast, at about longitude 144° E and latitude 14° S, around the Precambrian Georgetown area to Charters Towers, south-southwestward to near Wilcannia on the Darling River, and thence either due south or southwestward to Kangaroo Island. The discontinuous folded and granite-intruded Paleozoic rocks east of this somewhat hypothetical boundary are now separated by generally flat-lying younger rocks deposited in separate sedimentary basins. Because there is no obvious continuity of folded zones and of belts of contemporaneous folding from the south coast to the north of eastern Australia, it has been suggested by some that a Lachlan Geosyncline in the southeast (mainly in Victoria and south central New South Wales) should be distinguished from the New England Geosyncline, the western boundary of which extends from Newcastle to south of Townsville. They are two component parts of the Tasman geosynclinal belt and are similar in rock types, but they are different in age and geographic position.

Ordovician rocks. In contrast with the paucity of information on Cambrian Australia, the geography of eastern Australia in Ordovician time, from 500,000,000 to 430,000,000 years ago, is sufficiently well known to permit some general statements. The Lachlan Geosyncline was the site of sedimentation that took place in a series of troughs separated by rising upwarped ridges. No western shoreline can be defined, but the characteristics of Ordovician rocks in the Broken Hill region, and the evidence of bores in northeastern South Australia, indicate that these localities lay outside the geosynclinal belt in platform areas where subsidence was less intense. The geosynclinal nature of most of the rocks of Victoria, and southern and central New South Wales, is indicated by the great thicknesses of marine rocks deposited partly in deep quiet water (as in the case of black shales) and partly by mud-laden currents flowing from higher areas of the sea floor, producing the muddy sandstones known as graywackes. The rocks in the central part of the geosyncline were altered by heat and pressure, and volcanic rocks are widely distributed. In many parts of the Lachlan Geosyncline, folding movements (known as the Benambran orogeny, or mountain-building movement) occurred at the end of Ordovician and in early to mid-Silurian time. Accompanied by granite intrusions, these movements are thought to have caused the changes in composition and texture of the Ordovician rocks in the Snowy Mountains.

The Lachlan Geosyncline

Silurian rocks. The Benambran mountain-building movements changed the geographic framework of the sedimentation that occurred during Silurian times (from 430,000,000 to 395,000,000 years ago). The eastern margin of the Lachlan Geosyncline now extended from the Melbourne area northward through Cobar to the Adavale Trough. In the Melbourne Trough there are thick Silurian graywackes and mudstones with graptolites and other fossils, and the area of sedimentation extended southward into Tasmania, where shallow-water rocks were laid down. Farther east, in southern New South Wales, the Yass Shelf was an area of volcanic rocks and sedimentary rocks, including limestones and shales that are now famous for their fossils. Silurian rocks are not well known in the New England Geosyncline and in southern Queensland, but farther north there is another peripheral shelf development, with Silurian limestones occurring along the western margin of the Tasman geosynclinal belt and the Chillagoe Shelf and its southward extension.

Devonian rocks. Deposition in the Lachlan Geosyncline generally continued from Silurian into Lower and Middle Devonian Periods, from 400,000,000 to 350,000,000 years ago, although the earth-movements of the Bowning orogeny occurred at the end of the Silurian and through early Devonian. Evidence of this can be seen as an unconformity, or break, in the rock layers of the sedimentary series of the Canberra–Yass area: molten rocks were intruded into the sedimentary rocks during this period. This activity again changed the geography of sedimentary areas in the Lachlan Geosyncline. In particular, a shallow-water platform with limestone deposition extended from eastern Victoria (Buchan Caves) to the Yass–Taemas area. Geosynclinal deposition during early and mid-Devonian time was widespread in the New England Geosyncline and its extension in southeastern Queensland, but only corals preserved in isolated limestones permit precise dating of the rock layers.

The Tabberabberan earth movements

The Tabberabberan period of earth movements, which occurred at the end of mid-Devonian time, had a profound effect on the geographic development of eastern Australia. Marine sedimentation in the Lachlan Geosyncline ended, and nonmarine rocks, in part red sandstone beds resembling similar sandstones found in northern Europe, were deposited. Discoveries of abundant fossil remains of armoured fishes—notably the widely distributed *Bothriolepis,* which is also found in central Australia, North America, Eurasia, and Antarctica—have been made in western New South Wales. Remains of the scale-tree *Leptophloeum* are also common. Erosion of the rocks of uplifted areas in central Australia, Victoria, and most of New South Wales must have produced the material for extensive Late Devonian and Early Carboniferous lake and river deposits. These include sandstones of the Grampian Mountains in western Victoria.

Carboniferous and Permian rocks. In the eastern troughs, from New England to eastern Queensland, geosynclinal deposition continued under the influence of volcanic activity. In the Carboniferous Period, about 300,000,000 years ago, nonmarine rock deposition occurred in the south and west, while marine sediments accumulated in the east. The central part of New England was uplifted. The earliest FIleozoic glaciations in eastern Australia are of Late Carboniferous age and have been considered as the result of mountain glaciation.

The youngest rocks in the New England Geosyncline that were deposited before final deformation are of Permian age (about 280,000,000 to 225,000,000 years ago) and of marine origin. In Middle and Late Permian time the great New England batholith was emplaced in the older rocks, which were heavily folded toward the west. A line extending from the mouth of the Hunter River near Newcastle in the south to the eastern margin of the Bowen Basin in the Rockhampton–Bowen area marks the western limit of the Hunter–Bowen orogeny, which was the last of the major mountain-building phases in the Tasman geosynclinal belt and which terminated its existence as a belt of troughs of sedimentation.

The Hunter–Bowen earth movements

SEDIMENTARY BASINS AND MESOZOIC HISTORY

Quite distinct from the history of the successive, more or less elongated, highly mobile (in geological terms) troughs discussed above is the development of sedimentary basins over the Australian continent. Some of these developed at various times on the Precambrian basement of the western and central parts of the continent, while in other areas this type of sedimentation followed the consolidation of folded belts of Paleozoic rocks and interrupted their continuity. Others are superimposed on differently shaped preexisting basins.

Like the Perth Basin, which adjoins it to the south, the Carnarvon Basin of Western Australia began to form in Silurian time and contains Late Paleozoic deposits, including Early Permian glacial sediments, followed by Lower Triassic and, in places, nearly 13,000 feet (4,000 metres) of Jurassic sediments, dating from about 180,000,000 to 140,000,000 years ago. During the Upper Cretaceous Period, about 100,000,000 years ago, sedimentation in the Carnarvon Basin changed from sands and shales to limestones and chalks, and this may be related to a change to warmer water conditions, which continued in the north until Miocene time, less than 25,000,000 years ago. The west coast of the continent was, to varying extents over the course of the last 400,000,000 years, a geologically mobile shelf area; it exhibits an alternation of marine and nonmarine deposits representing successive transgressions and regressions of the sea.

In northwestern Australia the shallower Canning Basin with the adjoining deeper Fitzroy Trough formed a large subsidence area between the Western Australian Shield and the Kimberley Block. The Joseph Bonaparte Gulf and Ord basins extend along its northern and eastern margins on the West Australian–Northern Territory border. These basins date back to Ordovician time. The Canning Basin may have been linked eastward, through the Officer and Amadeus basins of central Australia, with the Tasman geosynclinal belt in the east, and, again in Permian time it may have extended southeastward in a similar manner. Such temporary configurations are exceptional in ancient Australian geography. As a rule, the structural independence of the basins is supported by differences in their fossil remains. An outstanding development of Devonian reef complexes is known from the northern edge of the Canning–Fitzroy basins, while Permian glacial deposits occur on its southern margin and continue into South Australia, Victoria, and New South Wales.

The Amadeus Basin. The Amadeus Basin in central Australia is a large depression extending east–west for 450 miles (720 kilometres) and north–south for no more than 160 miles (260 kilometres) and filled with many varieties of ancient sedimentary rocks. It is situated between the so-called Musgrave structural block in the south and the Arunta structural complex in the north, and both have moved toward, and folded sediments of, the basin at different times. Some sediments were deposited during the folding of the Macdonnell Ranges near Alice Springs, which later erosion shaped into spectacular scenery. The isolated sandstone masses of Ayers Rock and Mt. Olga are now considered to be of Cambrian age.

The Canning Basin continued to exist as a depositional trough during Jurassic time, when only its northwestern margin received marine sediments, and also through the early Cretaceous Period. During the early Cretaceous a great marine transgression joined it again southeastward with the Eucla and Officer basins and eastward with the Great Artesian Basin.

The Eucla and Officer basins. The Eucla Basin is a large but shallow depression lying between the Western Australian Shield and the basement rocks of Eyre Peninsula. Permian and Cretaceous rocks are known to exist from the evidence of bores. Above them lie Tertiary limestones, which are no more than a few hundred feet thick, perfectly flat-lying, with their surface forming the arid Nullarbor Plain. They are riddled with caves, many of which remain unexplored. The Officer Basin, south of the Musgrave Block, is filled with late Precambrian and Cambro-Ordovician rocks.

The Great Artesian Basin. The Great Artesian Basin is a large (670,000 square miles [1,735,000 square kilometres]) and complex, generally slowly subsiding area, in which, by Jurassic time, several previously separate basins (such as the Cooper Basin, with its rich gas-bearing Permian sands) had been incorporated. The Jurassic nonmarine sands of the Great Artesian Basin, which spread from the eastern highlands to the western margins of the basin, provide its most important sources of water. By mid-Cretaceous time the sea invaded the basin from the northeast and southwest, but the early Upper Cretaceous is represented by marine fossil-bearing deposits only near the northern coasts. The northernmost part of the Great Artesian Basin is separated from its main body by a basement ridge that hardly reaches the surface; north of it only Cretaceous and younger rocks are found. The southern boundary is formed by extensive basement outcrops in the Broken Hill–Wilcannia–Cobar areas, and the lowlands to the south are considered a separate unit, the Murray Basin. Permian and Mesozoic rocks play only a minor part in its filling. Its rocks are mostly of Tertiary age, marine in the west and mainly nonmarine in the east. A range of granitic hills in the south, the Padthaway Ridge, extending southeastward from Murray Bridge, forms a persistent boundary between the Murray and Otway basins. The Otway Basin, which extends into western Victoria and the continental shelf offshore, contains much greater thicknesses than the Murray Basin, of Cretaceous and Tertiary rocks in alternating marine and nonmarine development. Oil exploration in the Murray and Otway basins had not been commercially successful by the late 20th century, but the adjoining submarine Bass Basin between Victoria and Tasmania had become a major petroleum-producing region along with the early Tertiary age rocks of the offshore Gippsland Basin, which was Australia's largest petroleum-producing region.

The Murray and Otway basins

The Sydney Basin. The Sydney Basin was superimposed on a part of the folded rocks of the Tasman geosynclinal belt in Permian time and was filled with a considerable thickness of Permian and Triassic marine and nonmarine rocks. Overlapped by the Great Artesian Basin in the northwest, it continues southeastward under the continental shelf. The Permian begins with shallow-water marine sediments with erratic blocks indicating continued glaciation of the adjoining land, and volcanic material. The plant fossils are typical of the Permian of the southern continents. Coal deposition followed, and after another marine interlude, the Permian Period ended with the formation of further coal-bearing sediments, while volcanic activity and some earth movements continued. The Triassic sequence of the Sydney Basin consists of the Narrabeen Group of sandstones and shales, the Wianamatta Group, and the Hawkesbury Sandstone, which forms conspicuous cliffs around Sydney Harbour and is much used as a building stone.

The Clarence–Moreton and Maryborough basins. On the border between New South Wales and Queensland, from south of Grafton to north of Brisbane, the Clarence–Moreton Basin developed in Triassic time as an offshoot of the Great Artesian Basin. It contains basic volcanic rocks and coal measures that were laid down in several ages. The Maryborough Basin is the easternmost of the Australian sedimentary basins. It is filled with nonmarine Triassic and coal-bearing Jurassic strata, followed in order by volcanic rocks, Lower Cretaceous marine sediments, and coal measures. These beds are the most strongly folded post-Paleozoic rocks in eastern Australia. It is now known that in other basins the Mesozoic and Tertiary rocks are slightly folded, but it is believed that the observed folds, such as those in the central part of the Great Artesian Basin, are due to movements of the underlying older rocks. The more intense folding on the eastern seaboard has been thought to be related to intrusions of Jurassic granites and to a late structural phase in the development of eastern Australia.

Tasmania. The most important Mesozoic event in Tasmania was the large-scale (2,000 cubic miles) intrusion of dark, layered volcanic rocks in Early to mid-Jurassic time. This process produced rocks similar to those in the Karroo region of South Africa, in Brazil, in Antarctica, and also in the Palisades of the New York area, in the Northern Hemisphere.

CENOZOIC HISTORY OF THE AUSTRALIAN CONTINENT

During the Cenozoic Era, occupying the last 60,000,000 years of geological time down to the present, true mountain-building movements were confined to the island arcs in the north and east. Nonetheless, the Australian continent was not altogether stable. On the mainland, movements of the basic structural blocks produced some folding in the sedimentary cover of the basins, and uplifts

ate
olcanic
ctivity

and downwarps restricted or extended their boundaries. Faulting in the southeast and east led to outpouring of large fields of basalts in Victoria in mid-Cenozoic and Late Cenozoic time, and also to scattered but important volcanicity along the east coast. The mid-Miocene (20,000,000 years old) volcanoes of the Warrumbungle Mountains and the Tweed River areas in northern New South Wales and the spectacular Glass House Mountains in southeastern Queensland are the most important examples of this activity. In Victoria and parts of New South Wales, basaltic lava flows followed ancient valleys and buried their stream-deposited gravels. Known as "deep leads," they are locally gold bearing.

The relative ages of the geological events that took place during Cenozoic time are still under study. There is no doubt about the Late Cenozoic age of the uplift of the eastern highlands and also of the Flinders Ranges in South Australia, and the corresponding retreat of the sea from the Murray Basin. This process seems to have occurred intermittently throughout Tertiary time, perhaps culminating in a Pliocene–Pleistocene phase that has been named the Kosciusko Epoch, after the highest mountain on the continent. During the Pleistocene ice age, an area of some 400 square miles around this mountain and some surrounding valleys were glaciated, as were large areas of the Tasmanian highlands. This weak but obvious expression of a worldwide climatic change is only one phase in a sequence of changes that affected Australia during Cenozoic time. Like the sequence of structural movements with which it is likely to be connected, the history of climatic changes is still disputed. The evidence of plant and animal fossils, together with determinations of the temperatures of the period, indicate that the climate in Early Tertiary (Eocene) time, some 50,000,000 years ago, was temperate in the south to tropical in the north. It was also humid, with rivers flowing to the south coast and coal-forming swamps existing in the areas of Eyre Peninsula and south Gippsland. Currents warmer than at present reached the south coast in early Miocene time (some 25,000,000 years ago), while during the late Miocene and Pliocene epochs there was a fall in temperature and a change in the plant life. This came to be dominated by eucalypts and acacias, replacing the southern beech (*Nothofagus*) that had been dominant in Early and mid-Tertiary time. Climatic change is also reflected by other evidence, for example by the large extent of lateritic weathering, causing the formation of red, iron-rich residual deposits as the result of alternating wet and dry seasons. Whether the change to the present arid conditions of the inland regions was gradual and progressive is uncertain. Some soil scientists and biologists have postulated a period of greater aridity than the present, dated a few thousand years ago, but this is disputed by others. There is no doubt about changes of sea level as well as earth movements during the Pleistocene ice age: the eastern highlands were uplifted by some 2,600 feet, as was the western plateau, and there were corresponding changes of wind and ocean water-current patterns. These factors combined in an intricate manner with worldwide climatic changes to alter the climate during the last few million years. The last of the great hippopotamus-like marsupial herbivores (*Diprotodon, Nothotherium*) and giant flightless birds (*Genyornis*) died out when the great lakes in the centre of the continent (Lake Eyre, Lake Callabonna, etc.) turned into salt pans only a few thousand years ago.

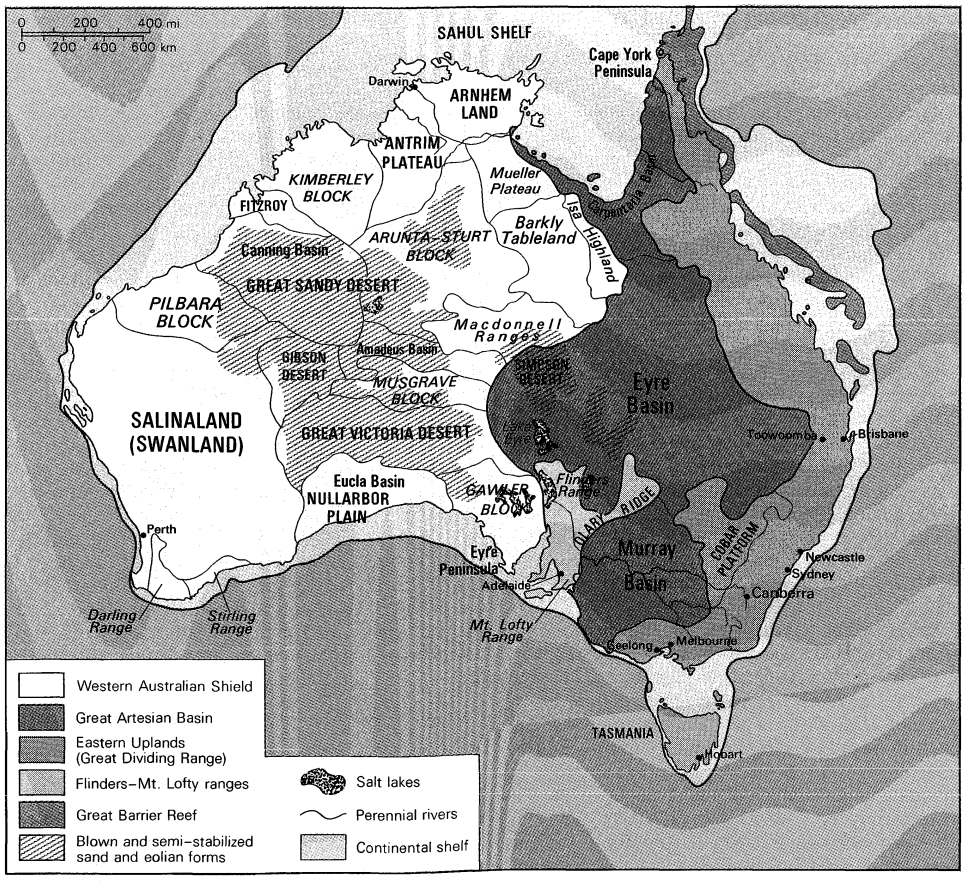
Recent climatic changes

(M.F.G.)

Physical geography

THE LAND

Australia is both the flattest continent and the driest. Seen from the air it is hard to believe that its vast plains, sometimes the colour of dried blood, more often tawny like a lion's skin, are not one huge desert. One can fly the 1,900 miles to Sydney from Darwin in the north, or the 2,000 miles from Perth in the west, without seeing a single town or anything but the most scattered and minute signs of human habitation. A good deal of the central depression and western plateau is indeed desert. Yet appearances can be deceptive. The red and blacksoil plains of Queensland



Physiographic regions of Australia.

and New South Wales have long supported the world's greatest wool industry, while recent discoveries have revealed that some of the most arid and forbidding areas of Australia conceal great mineral wealth.

Moreover, the coastal rim is, almost everywhere, an exception to these rules. In particular the east coast, where European settlement began and where the majority of Australians now live, is hilly, well watered, and fertile.

Inland from the coast runs a chain of highlands known as the Great Dividing Range, from Cape York in northern Queensland to the southern seaboard of Tasmania. From the coast itself this range, which may be anything from 20 miles (32 kilometres) to 200 miles (320 kilometres) distant, often appears as a bold range of mountains, though few of its peaks exceed 5,000 feet (1,500 metres). In fact, it is more like the escarpment of a giant plateau, formed of gently rolling hills, which then slopes imperceptibly down to the western plains. There are similar, though smaller, stretches of hilly, well-watered land all around the rim of the continent except on the south coast where the Nullarbor Plain stretches to the sea; but everywhere the rainfall diminishes rapidly as one penetrates further from the coast.

The "Outback"

To Australians the land beyond the Great Dividing Range and the coastal rim is the Inland, or the "Outback." For them it still retains some of the mythical quality it had for the first explorers searching for inland seas and great rivers. It is their Frontier: the land of hope and adventure. Yet in fact it is still very sparsely populated and perhaps always will be. The real heart of Australia lies in the industrial cities of the east and west coasts.

In this huge continent there are wide variations in scenery and climate. The thickly wooded ranges of the Great Divide have little in common with the treeless, sun-dried plains of the Inland. There is a vast difference between the red rocks and monumental hills of central Australia and the tropical rain forests and sugar plantations of north Queensland. Yet visitors to Australia usually detect a certain uniformity created, perhaps, more by the red earth, the brilliant light, and the drab, olive-coloured leaves of the ubiquitous eucalyptus than by any real resemblance. And if visitors from the Northern Hemisphere are at first repelled, as the English novelist D.H. Lawrence was, by "the vast, uninhabited land and by the grey charred bush . . . so phantom-like, so ghostly, with its tall, pale trees and many dead trees, like corpses," they should remember that to Australians born in the country the bush is friendly and familiar. Australia is not a pretty country, but it has a unique and haunting beauty that exerts a powerful fascination on those who get to know it.

If there is no real uniformity in the Australian landscape, there is certainly uniformity among the Australian people. No marked regional differences have emerged in the 200 years of European development, and steadily improving means of transport and communication have constantly worked to erase such differences as did exist. Today there is a strong similarity in the speech, manners, and customs of all Australian states, and everywhere the culture of white Australia is immediately recognizable as characteristic of Anglo-Saxon culture in Britain and North America.

Relief. *Overall characteristics.* Australia is a land of great plains. The island continent is almost 3,000,000 square miles in area, but of this, only 6 percent is above 2,000 feet elevation. Its highest peak, Mt. Kosciusko, rises to only 7,310 feet (2,228 metres). This situation stems in part from Australia's position at the edge of a zone of significant and recent earth movement and in part from the long periods of geological time during which Australia has been subject to weathering and erosion.

Patterns of faulting and folding in large measure control the distribution and attitude of rocks and thus play a significant part in determining the shape of the land surface. But the nature and intensity of the processes at work at and near the land surface also give rise to characteristic assemblages of forms. Australia is an arid continent; fully one-third of its area is occupied by desert, another third is steppe or semidesert, and only in the north, east, and southeast is rainfall adequate to support a vegetation that significantly protects the land surface.

The Western Australian Shield. The ancient west Australian core area, known geologically as a shield, or craton, is subdivided, most obviously in the north and west, by long, straight (or only gently arcuate) fractures called lineaments. These fractures delineate prominent rectangular or rhomboidal blocks, some of which have been raised to form uplands, others of which have been depressed to form lowlands or topographic basins. The lineaments display strong northwest–southeast and northeast–southwest trends in the northern, northwestern, and southeastern parts of the shield, but east–west alignments are prominent in the centre, and major structural lines are more nearly meridional in the west and southwest. In all areas, however, trends other than those locally dominant can be discerned.

The role of lineaments

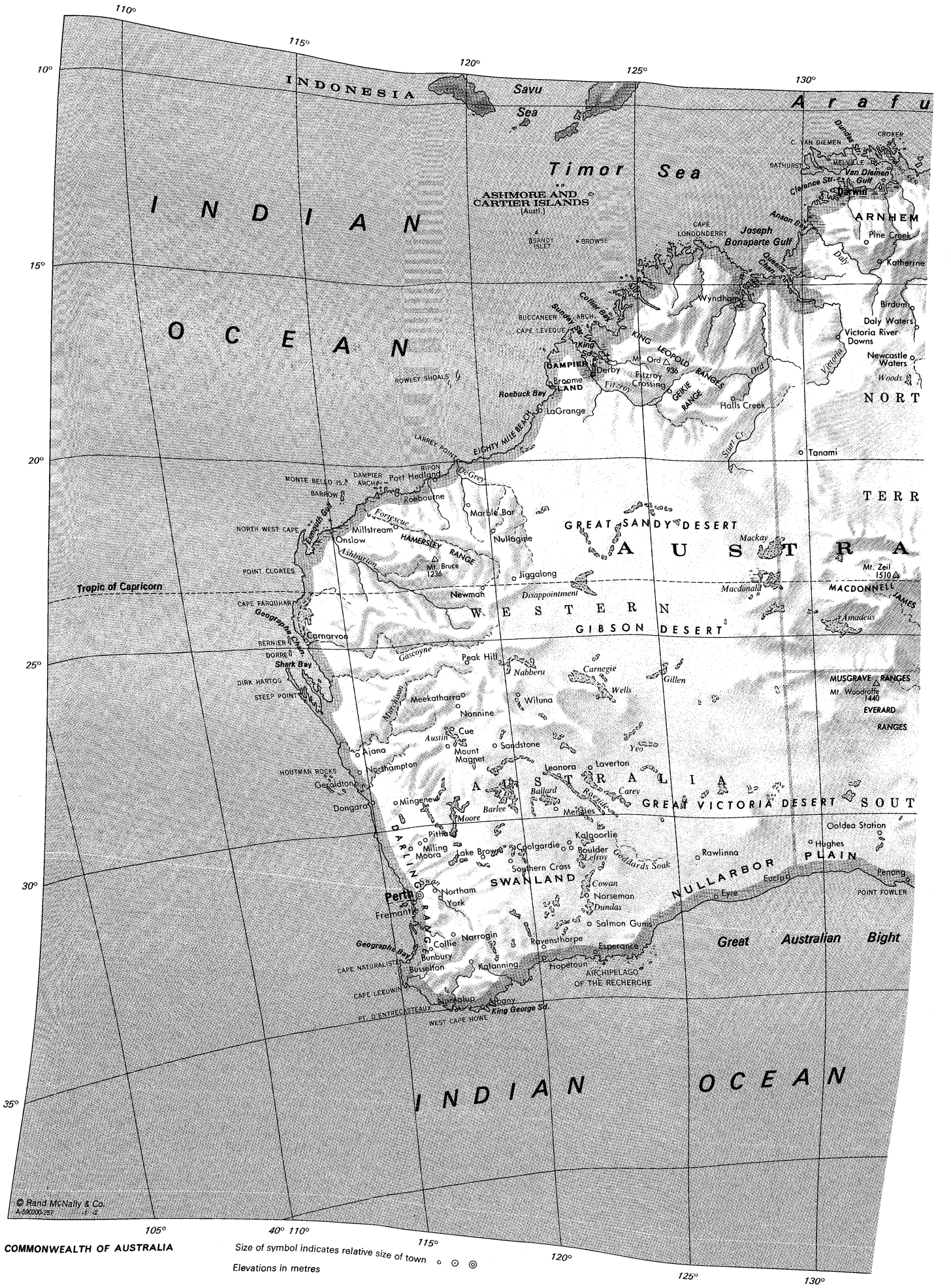
Within such structurally defined areas as the Kimberleys, the Isa highlands, and the Pilbara, the nature of the land surface varies according to the type and disposition of the rock outcrops. In the Kimberleys and the Mueller Plateau there are extensive outcrops of flat-lying massive sandstone that have been dissected to give rise to the striking isolated rock features known variously as plateau, mesa, and butte. Under these circumstances, local joints and bedding planes in the rocks, combined with the permeable nature of the bedrock, control the local landforms. Similar plateau forms dominate the Pilbara and Arnhem Land, though in the former region horizontally bedded or only gently warped massive ironstone formations, together with massive sandstones, give rise to prominent bluffs bordering the plateau assemblages; and in the latter karst landforms (greatly eroded by solution) are developed where limestone occurs at the surface. At the margins of the Kimberleys (in the Fitzroy region and in the Durack Range) and in the southern part of the Pilbara, in the Ophthalmia Range, dipping rock strata have been differentially eroded to form ridges and valleys. Such features are also extensively and well developed in the uplands of central Australia (the Macdonnell, James, and Kirchauff ranges), in the Isa highlands, and in the Stirling Range of the southwest. In all of these areas it is the sandstones and quartzites that underlie the upstanding ridges, the intervening valleys being eroded in siltstones or shales; and in all these areas the pattern in plan of ridge and valley reflects the pattern of folding in the underlying rocks.

In the far southwest, the Darling Range forms an upfaulted block underlain mainly by granite but capped by laterite, a reddish, iron-rich product of weathering rock. The Gawler Block, in the southeast, is complex. There are crystalline and sandstone uplands in the east, sandstone plateaus in the northeast, and, in the centre and north, the rounded Gawler Ranges built of Precambrian lava flows. Much of Eyre Peninsula is occupied by a rolling plain traversed by fixed sand dunes, but in the northwest numerous low isolated granite rocks of spectacular appearance, called inselbergs, stand above the plain. These epitomize the isolated ranges and hills widely developed in the northwest of South Australia, in the Musgrave, Everard, Birksgate, Mann, and Tompkinson ranges.

Isolated inselbergs

The lowlands between these raised blocks also display varied topography. The so-called Barkly Tableland is in reality a high plain of remarkable flatness, partly eroded in Cambrian sedimentary rocks and partly underlain by Tertiary swamp deposits. The Nullarbor Plain is approximately coincident with the Eucla Basin. A vast area of the southwest of Western Australia is occupied by Salinaland, an extensive high plain traversed by elongate ribbons encrusted with salt, the desiccated and disrupted remnants of former river courses. The Gibson Desert consists in large part of a laterite-capped plain, but huge areas of the plains of central and northern Australia are occupied by active sand dunes, and large areas of southern South Australia and Western Australia are covered by fields of fixed dunes.

Actively developing and moving sand ridges occupy the Canning Basin (the Desert Basin), the Great Victoria Desert, the Amadeus depression and large areas of the Arunta–Sturt Complex. The dune fields extend to the east into the Great Artesian Basin, where the dunes constitute the well-known Simpson Desert. These dune deserts reflect the



COMMONWEALTH OF AUSTRALIA

Size of symbol indicates relative size of town ○ ⊙ ◎

Elevations in metres


Coral Sea
PACIFIC OCEAN
Tasman Sea
PAPUA NEW GUINEA
CORAL SEA ISLANDS TERRITORY
QUEENSLAND
NEW SOUTH WALES
VICTORIA
TASMANIA
GREAT ARTESIAN BASIN
GREAT DIVIDING RANGE
GREAT BARRIER REEF
Gulf of Carpentaria
CAPE YORK PENINSULA
Brisbane
SYDNEY
MELBOURNE
Canberra
Adelaide
Hobart
Newcastle
Wollongong
Townsville
Rockhampton
Cairns
Launceston
Bass Strait
MURRAY RIVERINA REGION
FLINDERS RANGES
SIMPSON DESERT
SOUTH AUSTRALIA
AUSTL. CAP. TER.
SOLOMON ISLANDS (U.K.)

MAP INDEX

prevailing aridity of most of Australia, and the dune trend displays a huge swirl around the centre of the continent.

But even in these most arid areas—the area around Lake Eyre averages less than five inches of precipitation per year—rain falls from time to time, and the rivers run occasionally. Because of the scarcity of vegetational protection, and because of the common development of impermeable rock layers of various types, runoff in the arid lands tends to be rapid and to achieve dramatic and significant results. Hillslopes are scoured and washed bare of weathered debris; streams erode gullies and transport large volumes of sediment from the uplands to the plains; broad braided river channels are developed; and extensive alluvial plains are formed. It is the alluvium, carried to the lowlands by rivers and deposited on the plains, that is, in large measure, the source of the sand out of which the desert dunes are molded by the wind.

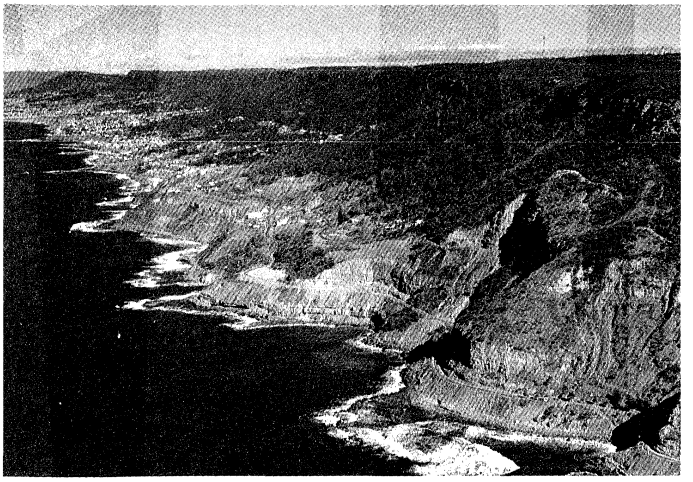
Effects of rapid water runoff

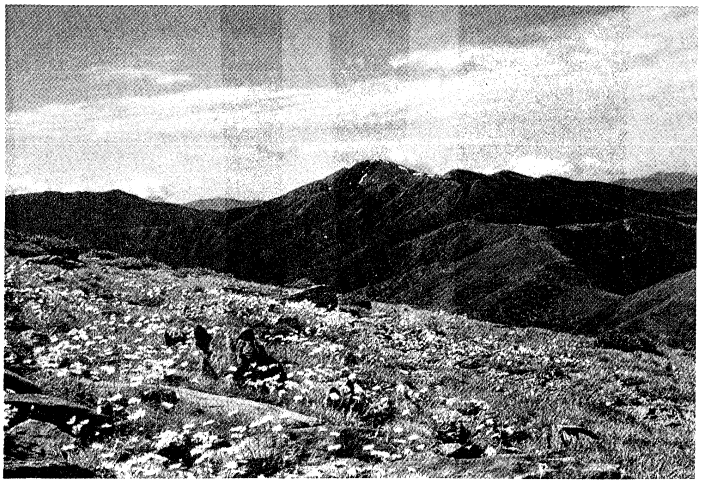
In the far southwest of the shield, and especially in the northern areas, rainfall is sufficient to support a considerable vegetation and is regular enough for streams to flow seasonally. Here the work of rivers in shaping the land surface is more obvious and widespread; the landscape consists essentially of valleys and intervening divides, the precise form of each depending on local structure. But in such areas the rate of landscape change is more rapid than in the arid zones.

Many of the landforms of the shield are inherited from the past, when different climatic conditions obtained. Remnants of laterite are widespread in many parts of Australia: the Darling Range, Salinaland, the Isa highlands and Mueller Plateau, the Darwin area, southern Eyre Peninsula. Clearly, at some time or times in the Tertiary Period, these areas had been reduced to low relief, and humid tropical climates prevailed, for laterite is at present forming only under such conditions in such areas as Southeast Asia and the Congo Basin. The disrupted former drainage system of Salinaland has already been referred to, and remnants of similar old stream networks occur in the Amadeus depression, on the Nullarbor Plain, and in the Great Victoria Desert. A large swamp formerly occupied the south of the Barkly Tableland; and Lake Woods, near Newcastle Waters, is now dry, with a bed of some 70 square miles in extent, but shorelines indicate that the lake formerly occupied some 1,100 square miles. Fossil remains also suggest wetter climates in the past in many parts of Australia, and subsequent deterioration toward aridity. But in the south the occurrence of dunes now fixed by vegetation shows that the climate there has recently become moister.

Finally, in several parts of the shield remnants of eroded surfaces, planed off and covered with hard, silicified crusts of weathered rock, cut across local bedrock and are either preserved high in the relief or buried beneath later sedimentary deposits. They attest to changes in the disposition of the land surface (either base-level changes or regional warping or faulting) and also indicate that, in the past, surfaces of low relief similar to present ones were widely developed. Reference has already been made to the distribution of the laterite surface. At the eastern margin of the shield there are remnants of a still older surface, of late or middle Mesozoic age, which has been warped by subsequent earth movements and now disappears beneath the sediments of the Great Artesian and similar basins. Evidence of the existence of this surface has been forthcoming from northwestern Queensland, central Australia, and South Australia.

The Flinders–Mt. Lofty ranges. Adjacent to the southeastern extremity of the shield, these uplands occupy the





Eastern Highlands (Great Dividing Range).
(Left) South of Sydney, where they shape the coastline near the towns of Bulli and Wollongong in left background; and (right) at Mt. Feathertop, inland Victoria.
(Left) G.R. Roberts, (right) John R. Brownlie—Photo Researchers

site of the Adelaide Geosyncline, or downwarp in the Earth's surface. These sediments were folded and faulted, principally in the early Paleozoic, though recurrently since. The Flinders Ranges are a much-eroded fold mountain belt characterized by ridge and valley forms in which sandstone ridges and bluffs are dominant. The Willochra Plain occupies an elongate intermontane basin excavated from a major upwarped structure and achieved through the erosion of some 20,000 feet of sediments. There are remnants of old land surfaces of low relief, and, in the north, extremely rugged relief developed on a much-shattered granite outcrop.

'he Villochra 'lain

To the south, the Mt. Lofty Ranges are a much dissected and complex horst, or ancient uplifted structural block. Bounded on both east and west by meridional or gently arcuate fault scarps, which developed initially in the Early Paleozoic but which have suffered recurrent movements since (and which indeed are still active), the ranges are surmounted in many areas by the remnants of a lateritic plain. In many other areas, such a hard capping of rock, if ever present, has been eliminated by stream erosion. Sandstones again form prominent ridges and residuals (isolated relief features), like Mt. Lofty itself; small granite outcrops give rise to boulder-strewn surfaces; and exposures of gneiss form slablike blocks known as tombstones, monk stones, or penitent rocks.

Between the Mt. Lofty and the Flinders ranges is the midnorth, a region of broad simple folds in which the sandstone ridges run for the most part north–south and in which the broad open valleys were in some instances occupied by lakes during the Tertiary Period, some 50,000,000 years ago. Similar upland areas of low relief, but with domes of crystalline rock standing above the general level, dominate the Olary Arc, which swings northeastward from the midnorth region.

The Great Artesian Basin. This platform area consists of three major basins, the Carpentaria Basin, the Eyre Basin, and the Murray Basin. Between the Carpentaria and Eyre basins the Euroka–Kynuna Plateau barely rises to reach the surface in such minute residual relief elements as Mt. Brown and Mt. Fort Bowen, in northwest Queensland. The Wilcannia threshold separates the Eyre and Murray basins, and the latter is separated from the Otway Basin and the Southern Ocean by the Padthaway Ridge. The two southern basins are entirely terrestrial, but the Carpentaria is partly inundated by the sea.

The Carpentaria plains, occupying the basin of the same name, form a narrow lowland corridor between the Isa highlands and the Einasleigh uplands (part of the Eastern Uplands). They are drained by the Leichhardt, Flinders, and Gilbert rivers, and in the south take the form of broadly rolling plains underlain by heavy gray lime-enriched (pedocalcic) soils and known as the Rolling Downs. In the north, however, there are extensive flat depositional plains, some of them related to Pleistocene swamps, some associated with the present floodplains of the braided river systems. Standing above the plains, for example around Normanton, are considerable plateau and mesa remnants of the Tertiary laterite surface.

Similar rolling plains with laterite residuals standing above them occur in the Eyre Basin, particularly around the headwaters of the Diamantina, near Kynuna. But to the south, toward the more arid interior, the plains become flatter and are protected by a veneer of stones—the well-known stony desert with its mantle of gibber (hammada, serir, desert armour). In many parts of southwestern Queensland, northeastern South Australia, and northwestern New South Wales there are plateau and related relief remnants similar to those found in other parts of the lowlands, although these are capped and protected not by laterite but silcrete, another hard rock residue. This region is folded in places, and the subsequent dissection by erosive forces has brought about disintegration of the silcrete, which is of Middle Tertiary age and which formerly extended over vast areas of central Australia. This process provided much stony debris for the gibber plains so characteristic of much of central Australia and particularly of the Lake Eyre depression.

Lake Eyre

The catchment of Lake Eyre extends over some half million square miles of central and northern Australia. It occupies the lowest point of the Australian continent (46 feet [14 metres] below sea level) and many large river systems drain into it. These rivers drain the driest part of the continent. But no desert is rainless, and floodwaters cover the bed of Lake Eyre about twice each century, the waters deriving not only from central Australia but also from the higher rainfall areas drained by the headwaters of the Georgina, Diamantina, Thomson, Barcoo, and similar rivers. It is now clear that during the late Pleistocene period, more than 10,000 years ago, the rainfall of central Australia was heavier than it is now; as a consequence, these rivers, past and present, have brought vast quantities of sediment and salt to the interior drainage basin. As a result, there is ample source material for the Simpson Desert dunes, and many of the normally dry lake beds, including all the large ones, are encrusted with salt. Most of the very large salinas, or salt pans (Eyre, Frome, Torrens, Gregory, Blanche), are, at least in part, of structural origin, having been formed by downfaulted blocks. Torrens and Gregory are surfaced mainly by gypsum, but the remainder carry a crust of sodium chloride, common salt. Around the major salinas there are extensive alluvial plains.

Under the prevailing arid conditions, fine dust is winnowed from the surface sediments and can be carried high into the air in dust storms. Some is carried long distances, even reaching New Zealand from time to time. The sand of the alluvium is molded into dune ridges.

Sand dunes also occupy large areas of the Murray Basin.

They are, by contrast, fixed (or "fossil") dunes, which developed at some time in the recent past and have since been stabilized by higher rainfall conditions. In the east of the basin, near the foothills of the Eastern Uplands, there is evidence of former higher rainfalls in the numerous abandoned river channels of the Riverina. But the western Murray plains are a stony as well as a climatic desert. The plains are underlain by Miocene limestones and, in many areas, by calcrete, a calcareous soil accumulation. There are instances of water-dissolved sinkholes and enclosed depressions, and lack of surface drainage characteristic of this type of topography. Only the Murray, which originates outside the area in a different environment, crosses the basin, flowing in a narrow trench in its lower reaches.

In the east of this region there are extensive alluvial plains associated with major tributaries of the Murray. One feature of interest is the diversion of the Murray, near Echuca, by a rising structural block bounded by fault zones and known as the Cadell Fault Block.

The Eastern Uplands. This complex series of high ridges, high plains, plateau, and basins extends from Cape York Peninsula in the north to Bass Strait in the south, with a southerly extension into Tasmania and one extending westward into western Victoria. The uplands are the eroded remnants of an ancient mountain range recently rejuvenated by block faulting. They occupy the site of the Tasman geosynclinal belt, the sediments of which were folded and faulted in late Paleozoic times. Granite batholiths were intruded into this region, and during the Cenozoic Era lavas appeared extensively in areas as far apart as north Queensland and Tasmania. Characteristic features associated with this process were lava fields, with stony rises, soil-filled depressions, and lava caves. Extinct cones and craters survive in southeastern Queensland, in the Monaro district of New South Wales, and in western Victoria.

Volcanic landforms

The landforms, in considerable measure, reflect these various geological events. Uplifted structural blocks, many of them trending north to south, are common in some areas, while straight river courses reflect the control exercised by fault zones. Ridge and valley forms, as found in the Grampians of Victoria, reflect the differential erosion of broken and folded rock strata. On the exposed granitic batholiths, massive domes or clusters of boulders are common. The lava plains and plateaus display stony rises, shallow alluvial depressions, and volcanic vents and plugs of various types and ages.

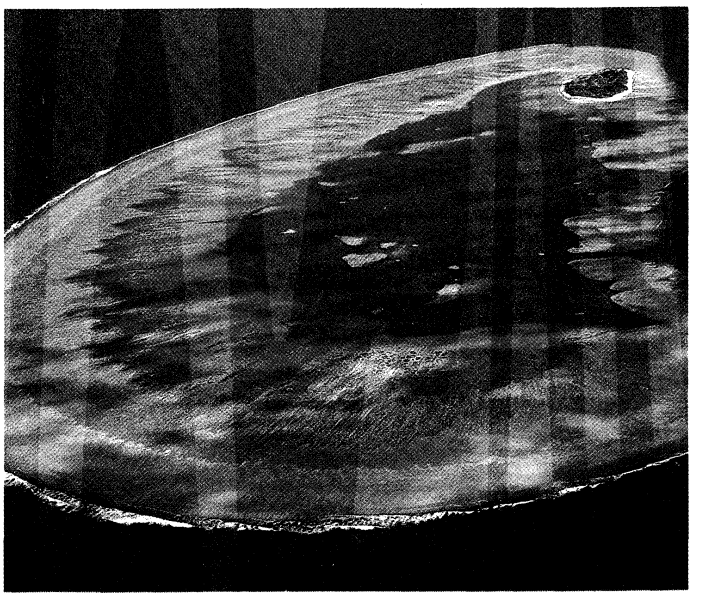
Other features reflect the erosional history of the region. Wide areas of the upland had been reduced to a uniform low relief by the time of the later Mesozoic Era, about 100,000,000 years ago, and many remnants of this ancient surface, exhumed by erosive action from beneath a Later Cretaceous cover, survive in the landscape, notably in north Queensland. The Middle Tertiary leaching of rocks by weathering in humid climates with the formation of iron-rich residuals (laterization) also affected the uplands, from northern Queensland to Tasmania.

Lastly, during the Pleistocene Epoch, some 2,000,000 years ago, small glaciers developed in the Kosciusko area of New South Wales and the central plateau of Tasmania. Small, ice-scoured hollows and small moraines (ridges of glacial debris) attest to these events, while over rather wider areas frost shattering of rocks and resulting downslope flowage of soil (solifluction) have helped shape the surface. No snow normally survives through summer in either of these areas now, but in winter the snowfields of the Kosciusko area alone are more extensive than those of all of Switzerland.

The Great Barrier Reef influence

The Great Barrier Reef is related in important respects to the Eastern Uplands. Lying off the Queensland coast, this great system of coral reefs and atolls owes its origin in part to Pleistocene changes in sea level but in most part to long-continued subsidence, related to faulting, of the offshore region. This slow subsidence has enabled a great thickness of coral to develop, and it is on this basement that the present reefs and coral atolls have grown in the clear warm waters of the Coral Sea.

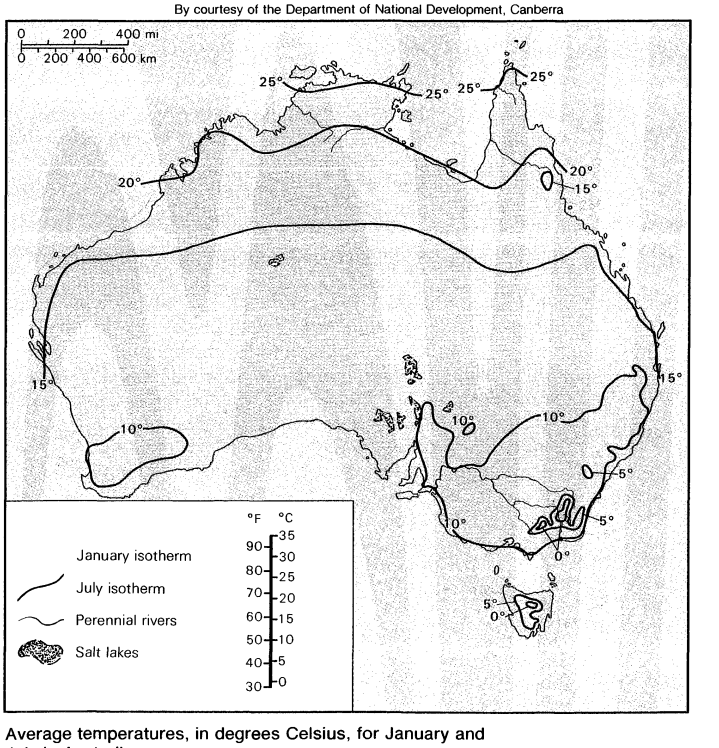
The coming of man. Though neither aboriginal man nor the later European settlers have been long in Australia, they have already achieved widespread, and in most ways deleterious, effects on the landscape. European man in particular has been responsible for initially minor, but later significant and widespread, changes, notably considerable soil erosion. The clearing of vegetation for agricultural purposes, overgrazing, the introduction of exotic plants and animals, the making of tracks and roads, even the clearing of stones from paddocks—all have rendered the land surface more susceptible to soil erosion. Man has set in train his own great cycle of erosion, similar to that which beset many parts of western Europe in the 18th century and which has assailed many parts of the American West since late in the 19th century.



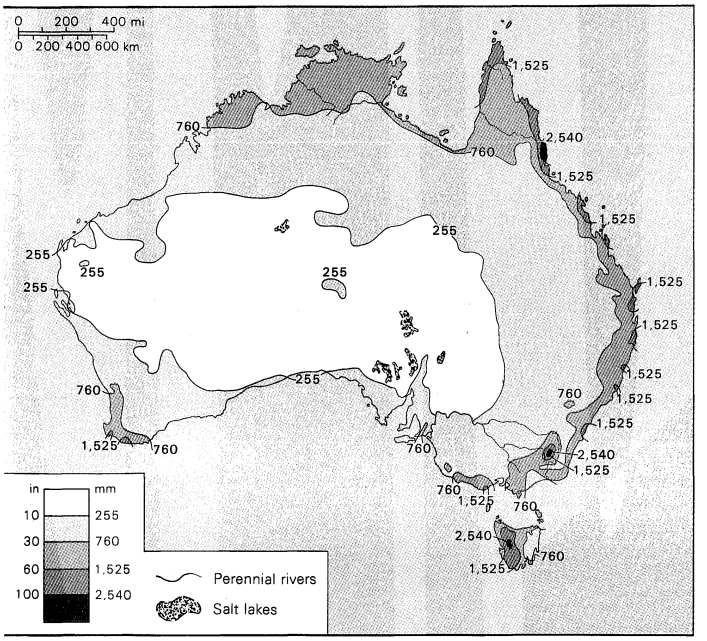
A portion of the Great Barrier Reef, lying off the coast of northeastern Australia and forming the largest coral structure in the world.
Douglass Baglin

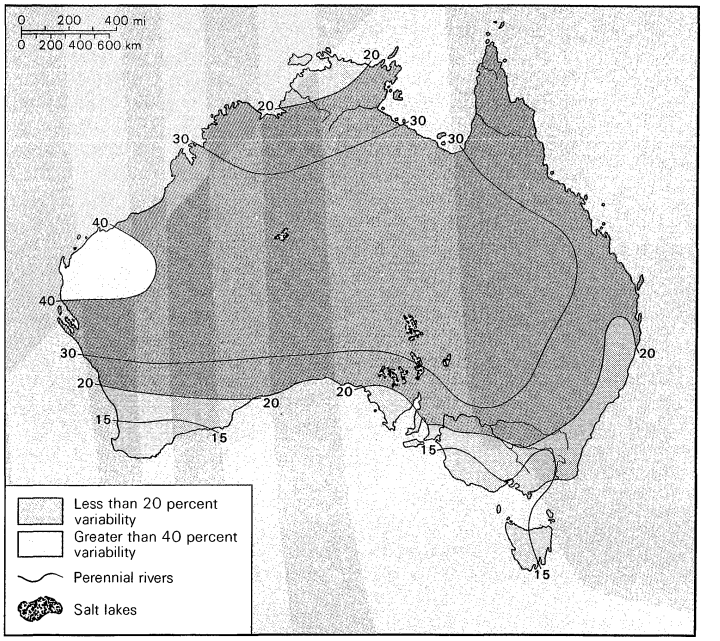
Climate. Australia is the arid continent. Over two-thirds of its landmass rainfall per annum averages less than 20 inches (500 millimetres) and over one-third of it is less

By courtesy of the Department of National Development, Canberra



Average temperatures, in degrees Celsius, for January and July in Australia.





Rainfall patterns in Australia.
(Left) Average annual rainfall in Australia (measurement figures on the map are in millimetres).
(Right) Percentage mean variability from annual mean rainfall.
(Right) From G. Leeper (ed.), *The Australian Environment;* Melbourne University Press

than 10 inches. Only just over 10 percent receives more than 40 inches per annum. As has been noted, in winter the snowfields of Tasmania and the Kosciusko area can be far more extensive than those of Switzerland, but on the whole Australia is a very hot country with a high incidence of heat waves, in consequence of which evaporation losses are high and the effectiveness of the rainfall received is reduced. In addition, the severity of climate, the predominance of the outdoors in the minds and lives of many, and the national importance of agricultural and pastoral pursuits, all make Australians perhaps more climate-conscious than most. In no country of comparable development do climate, the weather map, and the latest forecast loom so large in the lives and conversation of the common people.

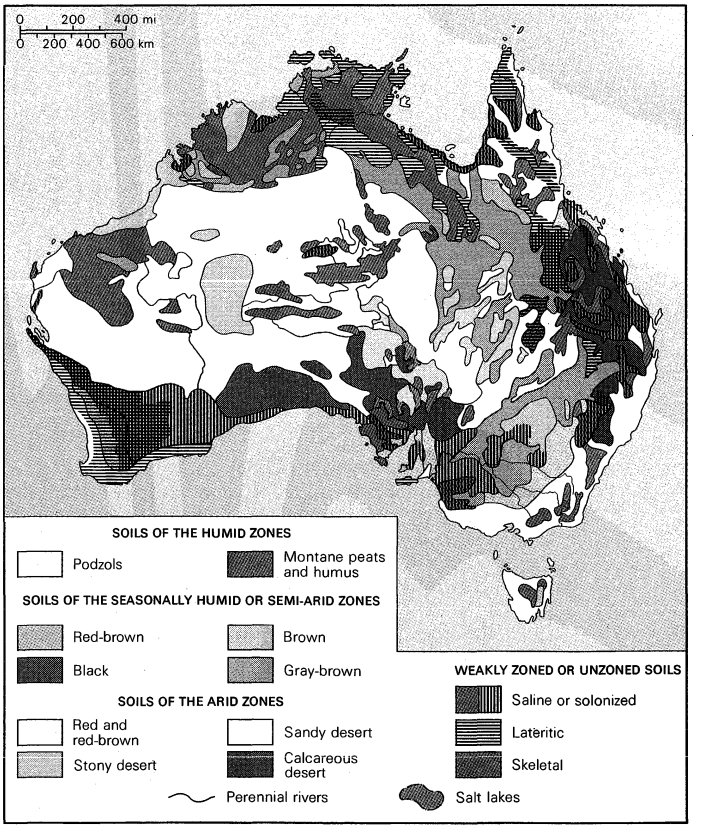
Major influences on climate

The principal features of Australia's climate stem from its position, shape, and size. Australia is a compact tropical and near-tropical continent located between latitudes 10°41′ S and 43°39′ S. No major arms or embayments of the sea penetrate far into the landmass. The only extensive uplands occur near the east coast, and even they are not, by world standards, very high.

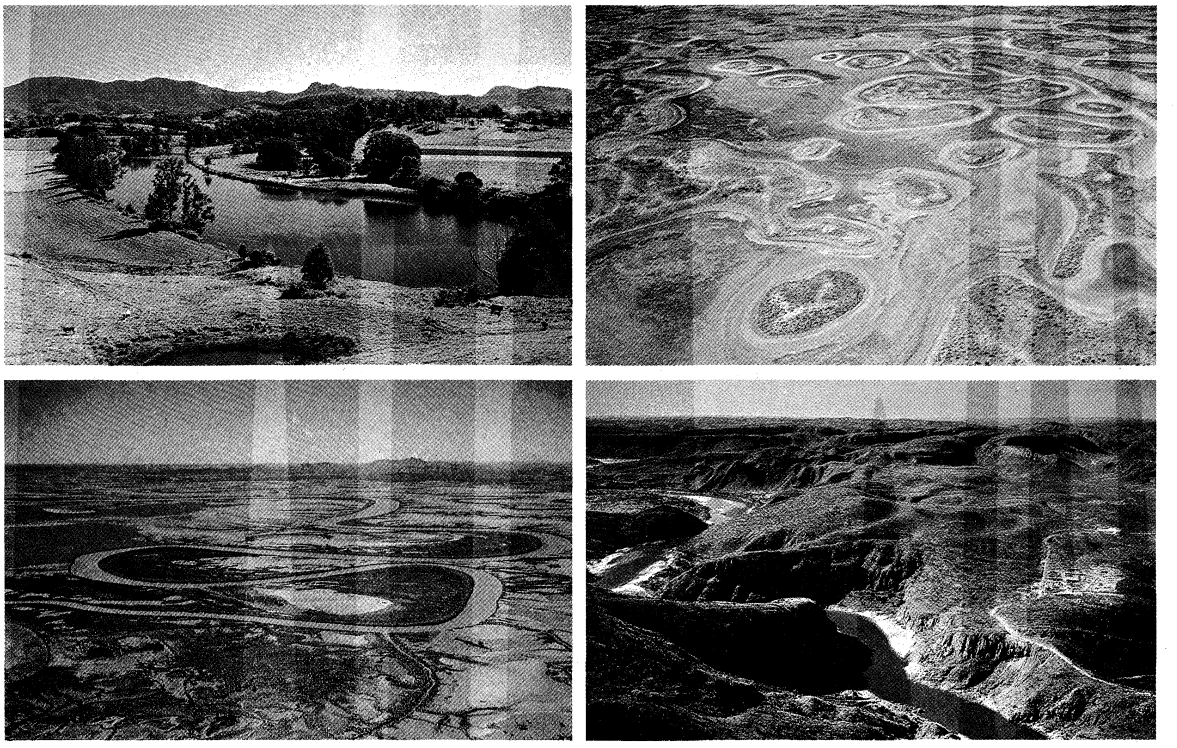
In summer, when the sun is directly overhead in northern Australia, temperatures are extremely high. The sea exerts little moderating influence, and the uplands are not sufficiently extensive or high to have more than local effects. Because of the lack of cloud over most of the interior, however, while temperatures commonly soar over the 100° F (38° C) mark, there is considerable radiation loss at night, and daily temperature ranges are quite high. But, on the whole, high temperatures dominate the Australian summers in all but Tasmania. Heat waves are common, and though the highest amounts of radiation are received in northern South Australia, the highest temperatures and longest heat waves are recorded in the northwest of Western Australia. Marble Bar, for instance, has recorded a maximum temperature of 100° F or more on 162 consecutive days. Temperatures in winter remain moderate except in the uplands of Tasmania and southeastern Australia, where snow is common. Night frosts are common in winter throughout southern Australia and in the interior.

Because of its latitudinal position, Australia comes under the influence of the southeast trade winds in the north and the westerlies in the south. Northern Australia is affected by a northerly monsoon, partly because of the latitude and the seasonal migration of planetary wind zones and partly because of the summer heating of the continent and the indrawing of surface winds. The monsoon brings summer rains to the northern coastal area and extends inland for variable distances. These summer rains are all the more important because most of northern Australia is in the sheltered rain shadow of the Eastern Uplands so far as the southeast trades, which are dominant in winter, are concerned. The trades, forced to rise by the uplands, bring heavy rains to the Pacific coast of Queensland and of northern New South Wales. These areas, which are also

By courtesy of the Department of National Development, Canberra



Types of soils in Australia.



Australian drainage patterns.
(Top left) Tweed River near Murwillumbah, New South Wales. (Top right) Salt lakes and "islands," Lake Amadeus, central Australia. (Bottom left) Fitzroy River system near Rockhampton, Queensland. (Bottom right) The Ord River, in the Kimberly district of Western Australia, is an important source of irrigation.

By courtesy of (bottom right) Australian Information Service; photographs (top left, top right, bottom left) Photographic Library of Australia

affected by tropical cyclones, receive the heaviest rains of any part of Australia, and, within this coastal fringe, the north Queensland area around Cairns is the wettest.

Southern Australia receives winter rains from depressions associated with the west-wind zone. Again there are local topographic controls, with uplands receiving more than the adjacent plains. Parts of the southern Mt. Lofty Ranges, in South Australia, average more than 40 inches of rainfall per annum, but Adelaide, to the west, averages only 21 inches, while the Murray plains, in the rain shadows of the ranges, receive 15 inches or less rainfall annually.

The dry interior

In the great mass of the interior of Australia, rainfall averages less than 20 inches per annum, and over vast areas the total is less than 10 inches; the Lake Eyre region averages less than 5 inches. Rainfall in these areas is unreliable and capricious, with long droughts broken by damaging rains and floods. Over Australia as a whole, rainfall is indeed extremely variable. Only in the far north, around Darwin, in the southwest of Western Australia, in southern South Australia and Victoria, in Tasmania, and in eastern New South Wales is the recorded annual rainfall not more than 10 percent above or below the long-term average in different years.

Drainage. Permanently flowing rivers are found only in eastern Australia and in Tasmania. The major exception is the Murray, a stream that rises in the Kosciusko region in the Eastern Uplands and is fed by melting snows. As a result, it acquires a volume sufficient to survive the passage across the arid and semi-arid plains that bear its name and to reach the Southern Ocean southeast of Adelaide. All other rivers in Australia are seasonal or intermittent in their flow, and those of the arid interior are episodic.

Many areas—notably the Nullarbor Plain, underlain by limestone, and the sand ridge deserts—are without surface drainage. A map of Australia can be misleading; though many "lakes" are depicted in the interiors, the fact is that many of them are now salt lakes that contain no water for years on end (see also below, *Natural resources: water resources*).

Soils. In broad view, the continental pattern of soils is closely related to climatic factors; mineral or skeletal soils exist over much of arid Australia, with virtually no organic content and little development to any depth. They may consist merely of a rough mantle of weathered rock. Gypsum is present in many of the desert loams and arid red earths. The soils of the semi-arid regions (where annual rainfall is from 8 to 15 inches) are also alkaline, with gypsum and lime a common feature. The organic content of the soils is again low in the solonized (salt-enriched) brown soils and the gray and brown soils of heavy texture that are common in these areas.

In both the arid and semi-arid regions gilgai—patterns of swells and depressions due to the alternate swelling and contraction following wetting and drying of clay soils—have developed. They are especially well represented in areas of seasonal rainfall. In areas with 15 to 25 inches of annual rainfall, black earths, brown soils, and red-brown earths are the most common soils. In the wetter areas the leaching out of minerals is a prominent feature of the soils. Podsols—sandy, with much humus at the surface and acid throughout—are the characteristic soil types. In the alpine regions humus soils—surface peats over a mineral—are noteworthy.

Local soil variations

Superimposed on these broad, climatically determined, soil patterns are local variations due to topography, groundwater conditions, and parent materials. For example, red soils of one kind (krasnozems) are developed on the basalt outcrops so common in eastern Australia, and those of different composition (terra rossas and rendzinas) on calcareous bedrock. In addition, laterite and silcrete originated in remote geological times, when conditions were very different from those of today. Laterite is represented in every state, including Tasmania, though it is forming nowhere in Australia at the present time, while silcrete is restricted to arid Australia and parts of subhumid Western Australia, South Australia, and Queensland.

(C.R.T./J.D.Pr.)

PLANT AND ANIMAL LIFE

Overall characteristics. Less than 200 years ago, Australian vegetation was still in its primal stage; the older plants in its present woodlands and forests began their

lives before the continent was invaded by Europeans. So close is the continent's prehistory that many a eucalypt still bears the great scar of a canoe or shield cut from its bark by Aborigines. Others can be found bearing blazes and inscriptions dating back to the first years of exploration and settlement.

In the short history of modern Australia, vast changes have been wrought in the continent's vegetation. Agricultural expansion stripped whole regions, substituting introduced crops, pastures, and plantations, while uncleared areas near the new settlements, ranging from the densest forest to the sparsest woodland, were cut through for timber. Enormous central and northern areas too arid for agriculture or too remote for timber getting were stocked with millions of sheep and cattle. In addition, many weeds were introduced, along with rabbits, other herbivorous vermin, and frequent bush fires. The native vegetation is still in that moment of destructive upheaval that marks the passing from aboriginal man to technological man and domestic flocks. Reserves and wilderness areas hold some native vegetation as national heritage, although it is doubtful if these are adequate or sufficiently protected.

Plant life. Australian federal and state governments maintain institutions for the scientific collection and study of the kinds (or taxa) of plants. Cumulative knowledge of Australia's flora stems mainly from these endeavours and is partly available in handbooks (Floras) listing species, together with appropriate "keys" for their recognition. G. Bentham's *Flora Australiensis* (1863–78), based on 19th-century collections sent to Europe, remains—although much outdated—the only comprehensive survey of Australian flora. More up-to-date information, scattered in the botanical literature, is not easily accessible except to professionals.

Australia's phanerogamic (seed plant) flora is estimated at 15,000 to 20,000 species, believed to be a blend of elements from various original sources and the outcome of a long and complicated history. One contribution to the present plant patterns is thought to have originated, in the remote past, from some southern landmass that then linked all the present southern continents. There is evidence to support this view: Australia shares with South Africa, Madagascar, New Zealand, and South America such plants as the southern beech (*Nothofagus*) and characteristic genera in other families, which constitute an Antarctic element.

Plant origins

Australia also shares many more kinds of plants with its northern neighbours. Also, typically Australian genera such as *Callitris* (native pine), *Banksia,* and *Eucalyptus* (gum tree), extend into the Indo-Malaysian area, some as far as the Celebes. This evidence has given rise to the idea of an Indo-Malaysian element in the Australian plant life, in a process that involved some two-way exchanges. These, and other links with plant life elsewhere, have prompted various hypotheses invoking the development of species, migrations, and extinctions throughout geological time, along with changes in the disposition of land, sea, and climate.

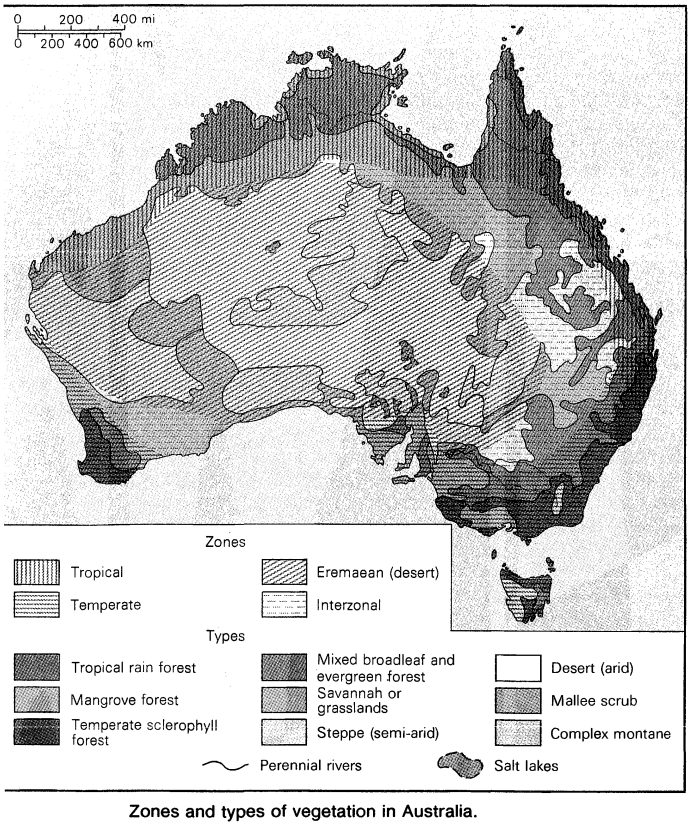
That characteristic part of Australia's plant life that is not much shared with other lands, together with those specialized characteristics that apparently originated in the continent long ago, form what has been designated as an Australian (or autochthonous) element. It includes many of the plants lending character to typical Australian vegetation scenery. It also shows a marked tendency to sclerophylly (formation of hard leaves) and wide-ranging adaptation in many genera, extending their species throughout the whole range of the continent's environmental habitats. Most obvious to the visitor are the *Eucalyptus,* which are represented by over 600 species, ranging in size from diminutive mallees, smaller than a man, to forest giants matching in bulk and height the world's biggest plants. Their habitat is similarly varied, ranging from rain forest to snowfield to hot desert fringe. *Acacia* has undergone similar adaptive radiation; its species range from mulga and myall—the dominant trees of vast areas—to small leafless blades at ground level. The banksias and hakeas (Proteaceae), the grass trees and blackboys (Xanthorrhoeaceae), and the kangaroo-paws (Haemodoraceae) are examples of the many other characteristically Australian plants.

Vegetation. Vegetation, as opposed to plant life, implies the structure and communal relations of the landscape's plant cover, whether it be forest, grassland, or marsh. There is no standard, or worldwide, classification system (such as exists for describing flora) for this aspect of the environment. Initial attempts to apply European and American classification ideas and methods to Australia were not particularly satisfactory, as a result of the peculiarities of the continent's vegetation and environment. For example, climatic control of local vegetation zones was often found insufficient to support vegetation analysis; on the contrary, soil patterns and geological history quite override climatic control in many localities. Similarly, structural descriptive schemes useful for Northern Hemisphere coniferous and deciduous vegetation proved quite inappropriate when confronted by the great variety of evergreen vegetation—notably mallees and scrubs—found in Australia. The mapping of Australian vegetation is based largely on factual descriptive features, and by this means such comprehensive and detailed accounts and maps as those contained in the *Atlas of Australian Resources* and *The Australian Environment* (see *Bibliography*) have been produced. Scientific vegetation analysis, not to be confused with mapping, is also well advanced in Australia, contributing modern ideas equal to those anywhere else in the world.

Classification difficulties

Modern Australian plant life is distributed in three main zones—the Tropical, the Temperate, and the Eremian—which reflect overall climatic conditions. The first zone, arced east and west across the north margin of the continent and extending halfway down the eastern seaboard, has a mainly dry monsoonal climate, with very wet patches. The second, with a cool to warm temperate to subtropical climate and mostly winter or nonseasonal rainfall, is arced across the southern margin, embracing Tasmania and extending up the eastern seaboard to overlap slightly with the Tropical Zone. The third zone covers the whole of central Australia, through to the central west coast; its climatic characteristic is aridity.

The major structural units comprising this geographical distribution are rain forest, sclerophyll forest (dominated by hard-leaved plants such as eucalypts), woodland,



Zones and types of vegetation in Australia.

scrub, savanna, and grassland forms, each with a range of subforms. The bulk of the Tropical Zone carries mixed deciduous woodland and sclerophyll low-tree savanna, with areas of tussock grassland, coastal mangrove complexes, and tropical rain forest. There is a high occurrence of exotic vegetation, particularly in the northeastern part of Cape York Peninsula, in Queensland, and a strong Indo-Malaysian influence occurs throughout the entire zone. The rain forests—with large trees with stem buttresses and multiple vegetation layers with interlaced canopies of lianas and epiphytes growing parasitically on the trees—qualify technically as jungles.

The Temperate Zone

The Temperate Zone is characterized by dry and wet sclerophyll forests, temperate mixed woodlands, savannah woodlands, mallees, and scrubs, with areas of Alpine vegetational complexes, temperate rain forest, and sclerophyll heath. Native plants form a much higher proportion of the vegetation cover, much of which is typically and recognizably Australian. Within this zone the southwest corner of Western Australia is outstanding, both for the high proportion of Australian plants and for the richness of the plant life, while the vegetation of Tasmania is notable for its southern beech forests and for its links with New Zealand and South America. In marked contrast to the tropical rain forests, the predominant trees throughout the majority of the Temperate Zone communities are either *Eucalyptus* or *Acacia.* Much of the Temperate Zone vegetation has been cleared for agricultural purposes, leaving only the vegetation communities of infertile or inaccessible localities.

The vegetation of the Eremian Zone ranges from barely vegetated desert sand hills through a variety of semiarid shrub savannas, shrub steppes, semiarid tussock grassland, and sclerophyll hummock grasslands. Many shrubs have adapted themselves similarly to the arid conditions so that in their vegetative state representatives of numerous families look alike. *Acacia, Eremophila,* and *Casuarina* are examples of genera that tend to displace *Eucalyptus* as the dominant tree shrub. Much of this vegetation is badly degraded. (R.T.La.)

Animal life. Human activities had a modifying and generally destructive influence on Australian animal life, or fauna. The arrival, about 30,000 years ago, of the Aborigines was in this respect less significant than European occupation, which has had profound effects over a period of only two centuries. The Aborigines brought only the dingo, a placental carnivore that must have affected the native marsupials. The Aboriginal hunters and food gatherers lived in ecological balance with their environment, and their demands were not large enough nor their technology sufficiently developed to upset it. There is even less scientific evidence to link them with the extinction of the large Pleistocene mammals than there is in other continents. All this changed with the advent of the Europeans: their cats, rats, foxes, rabbits, cattle, and sheep and their technology modified and largely destroyed important habitats of the Australian fauna. The inhabited coastal region (and even part of the arid inland regions) of the continent now contain a very much modified fauna compared with the indigenous life of only 200 years ago, which cannot now be reconstructed.

Destructive effects of Europeans

The Australian fauna is markedly different from that of the nearest land areas, the islands of Indonesia. A 19th-century biologist, Alfred Russel Wallace, designated an Australian zoogeographic region in 1876 and drew the boundary separating it from the Oriental region of Southeast Asia between Bali and Lombok and between Borneo and Celebes. This division, which became known as Wallace's line, is still recognized in modern biogeography as the boundary of a transitional zone, across which animals spread according to their ability to cross narrow seaways. These passages were much narrower during the Pleistocene glacial periods of the last 600,000 years, when so much oceanic water was frozen at the poles that the sea level fell to or beyond the edges of the continental shelf that now fringes Australia. At such times, the continent had land connections with New Guinea and Tasmania, while remaining separated from the Indonesian Archipelago during all of Cenozoic time (*i.e.,* for the last 50,000,000 years). The history and origin of the distinctive Australian fauna has led to much controversial speculation and searching for relationships among the southern continents.

A published count indicates the existence of 108 placental mammal species, 119 marsupials (125 according to a later count), 2 monotremes (egg-laying primitive mammals), 520 birds, about 380 reptiles, 122 frogs, and 180 freshwater fish. There are also more than 54,000 species of insects (Mackerras, 1970; see *Bibliography*) and about 750 species of mollusks. The placental mammals (apart from those introduced by humans) belong to groups that can swim (seals), fly (bats), or drift on logs (rodents).

The marsupials are considered distinctive Australian animals because they are more abundant and more differentiated on the Australian continent than in America, where few remain, and in Europe, where none has survived. They are not closely related to any fossil or living species found elsewhere. Their history in Australia has been traced back to fossil remains not more than 25,000,000 years old; that is, for less than one-half of the time of their presumed existence in isolation in Australia. There, isolation from placental predators and competitors gave them time to differentiate. The kangaroos, herbivores of the open woodlands and grasslands (occupying the habitat of horses or antelopes elsewhere); the tree-dwelling cuscuses (*Phalanger*) and flying phalangers, or gliders (*Petaurus*), resembling flying squirrels; the "native bear," or koala (*Phascolarctos*), which is specialized to live on eucalypt leaves; the burrowing wombat (*Vombatus*); the native cats (*Dasyurus*); the marsupial mice (*Sminthopsis* and *Antechinus*); the numbat, or banded anteater (*Myrmecobius*); the dog-sized Tasmanian wolf (*Thylacinus*); and, finally, the marsupial mole (*Notoryctes*) are examples of the adaptive differentiation of these mammals.

The role of marsupials

There are two types of monotreme extant, and both are wholly protected by Australian law. The duck-billed platypus (*Ornithorhynchus anatinus*) is common in streams and lakes throughout eastern Australia. The short-nosed echidna (*Tachyglossus aculeatus*) is also common but is widely distributed throughout Australia, including Tasmania. (A related anteater, the long-nosed echidna [*Zaglossus bruijni*], is found only in New Guinea.)

There has been much concern among nature lovers and conservationists in Australia about the obvious losses suffered by the marsupial fauna. These losses have been due both to increases in rangeland agriculture and to the largely uncontrolled exploitation of the kangaroos. According to one observer,

> The lands of arid Australia have been grazed by sheep, cattle and rabbits for periods varying from a few years in the most recently occupied areas to a little over a century in the oldest settled parts, a period short in comparison with most other arid lands of the world. In that short period we have caused far more degeneration of land resources than the aborigines caused in twenty to thirty millennia. Australia is fortunate indeed that the period has been so short. (From R.O. Slatyer and R.A. Perry [eds.], *Arid Lands of Australia;* Australian National University Press, 1969).

To this progressive destruction of habitat has to be added the effect of loosely controlled, or uncontrolled, shooting of kangaroos for meat and fur. Though it was banned from 1973 to 1975, the export of kangaroo products has become increasingly profitable. On the other hand, popular opinion had forced the government to take measures for the strict protection of the koalas in the 1920s. Many small marsupials are also endangered, and some have been eliminated in their original habitats by marauding cats and foxes. Conservationists hope that the increasing number of national parks and reserves will assist in the preservation of the native fauna, though supervision in the vast, almost uninhabited inland areas, where natural conditions remain least disturbed, is a serious problem.

Australian birds

Dangers similar to those threatening the marsupials also affect the birds of Australia, many of which are also unique to the continent and, therefore, of great scientific interest. The best known typically Australian birds are the flightless emu (*Dromaeus*), the mallee fowl (*Leipoa,* which builds a mound-shaped nest for hatching its eggs and actively controls the mound's temperature), the spectacularly

abundant cockatoo (*Cacatuinae*), the lyrebird (*Menura*), the fairy wren (*Malurus*), the honey-eaters (Meliphagidae), the Australian magpie (Cracticidae), and the bowerbird (Ptilonorhynchidae).

Australia has one freshwater crocodile (*Crocodilus johnstoni*), which lives in the tropical north but is also represented in rock engravings made by Aborigines in South Australia. There are 10 freshwater tortoises belonging to a family (Chelyidae) that is also known from South America. Lizards include geckos, skinks, legless lizards, and goannas, or monitor lizards (*Varanus*), which have relatives in Southeast Asia and Africa. There are many poisonous snakes, of which the taipan (*Oxyuranus scutellatus*), the tiger snake (*Notechis scutatus*), the death adder (*Acanthophis antarcticus*), and the brown snake (*Pseudonaja textilis*) are most dangerous to man.

The isolation and predominant aridity of Australia makes its freshwater fish fauna an interesting object for study. The Queensland lungfish (*Neoceratodus*) has lived in Australia without change for millions of years and is very much like its European ancestors of 200,000,000 years ago; it is very different from African and South American lungfishes. European fishes introduced recently into some Australian rivers and streams seem to be a danger to the survival of native species.

Insects have been as successful in Australia as in other continents, showing many adaptations to hot and dry conditions, particularly by adjusting the timing of drought-resistant developmental stages. Leaf-eating insects, including locusts, may be plagues in pasture regions. Other insects attack timber, and the destructive, as well as the constructive, activities of the widespread termites are well known. Bloodsucking insects attack cattle and sheep, and some are disease carriers. The larvae of blowflies (Calliphoridae) attack the living tissues of sheep and continue to cause losses of millions of dollars. (M.F.G.)

SETTLEMENT PATTERNS

Australia did not yield easily to development by Europeans. Even on the relatively favoured eastern seaboard, the first settlers found it hard to cross the rugged mountains of the Great Divide. When they did so they had to fight an endless battle against savage droughts, sudden floods, and fierce bush fires. Though they had little to fear from the Aborigines or from the gentle and harmless Australian animals, the land itself often appeared as a harsh and implacable foe. The long struggle by these settlers to tame the Australian Outback helped to form the tough and independent character of modern Australians just as the struggles of the pioneers and frontiersmen helped to mold the national character of the United States.

By the second half of the 20th century the automobile, the airplane, and the radio had greatly eased the harshness of life in country districts. Although many families still live in astonishing isolation by the standards of most other countries—it is nothing for a grazier to live 50 miles from the nearest country town and perhaps 10 or 15 miles from the nearest neighbour—most of them are linked by telephone and by well-surfaced roads. Even those who live on the great sheep and cattle stations beyond the Darling in New South Wales or in northwestern Queensland are linked by radio and can call the radio doctor and the flying ambulance in emergency. Many of the bigger stations own their own airplanes.

heep nd cattle ations

In Australia a sharp distinction was made in the past between the grazier, who runs sheep or cattle on his "station" or "property," and the small farmer who grows wheat or fruit or raises dairy cattle on the coast or in the irrigation areas. That distinction has been blurred since World War II by the subdivision of the biggest properties and by the fact that many graziers, faced by rising costs and the falling price of wool, have been forced to grow wheat or other crops to supplement their income. But Australia is still one of the last countries where one can find huge properties of 100,000 acres (40,500 hectares) or more where sheep and cattle graze over land unbroken by the plough. Life on a big sheep station is still a privileged one where a comfortable income, the pride of owning wide acres, and the spacious dignity of life on the sunburnt plains more than compensates for the loneliness and the inevitable hazards of drought and bush fires.

The life of the small farmer in Australia is less attractive and, consequently, less envied. Most of them struggle against the same hazards of drought and fire without the reassurance of broad acres to fall back on. Very often their homes are modest, and their cash income may not equal that of an industrial worker on the basic wage. Many of them could not survive or compete with imported produce without generous government subsidies.

Because of the vast distances and sparse settlement, nothing like the European village has ever developed in Australia. Instead, there are country towns that serve a wide area and vary a great deal in size and amenities. Many of the Outback towns are dusty little settlements with one wide main street, a store, two or three hotels, and not much else. But in the areas of closer settlement nearer the coast, many substantial country towns have grown to a point where they offer excellent medical and educational services and first-class shopping. Very few towns in the Outback, however, have more than 25,000 inhabitants.

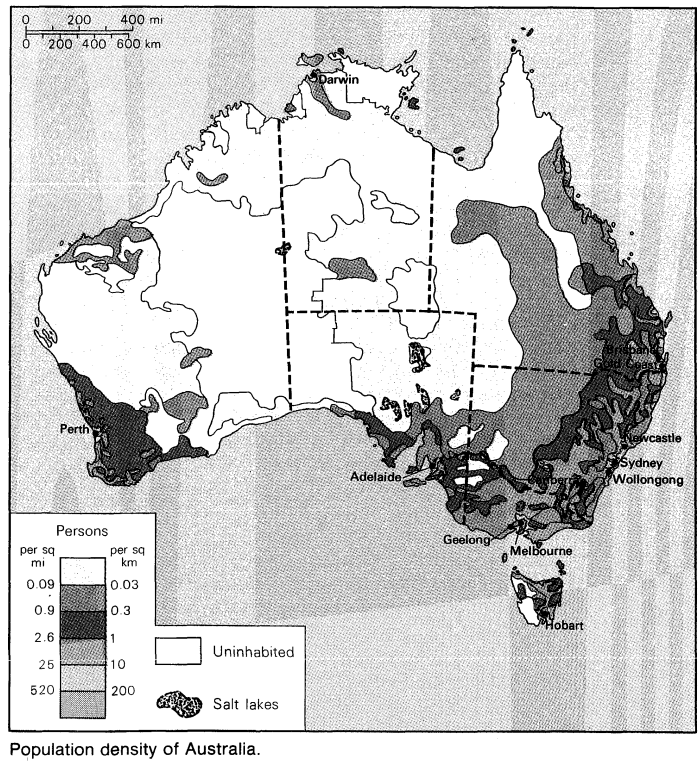
The great paradox of Australia, however, is that in this huge continent with its small population, relatively few people live in the country at all. The majority of the population lives in the seven capital cities. Many Australians regard this balance of the population as unhealthy, and most politicians pay lip service to the principle of decentralization from time to time; but historical, climatic, and economic reasons all suggest that this situation will not change much in the future.

Urbanization

Yet there are compensations in this imbalance. Both Sydney and Melbourne are large, modern, sophisticated cities that can compare in the services and amenities they offer with any city in the world except, perhaps, London, New York City, and Paris. Although neither is the federal capital, they are still the most important cities in Australia, the natural centres of both business and the arts. After Sydney and Melbourne the largest cities are the state capitals of Queensland (Brisbane), South Australia (Adelaide), and Western Australia (Perth).

All of these cities grew naturally as centres of local trade and commerce. Canberra, on the other hand, was artificially created as the federal capital of Australia. Although building did not begin until 1913, by the late 20th century it had become the most rapidly growing city in Australia.

The original plan for Canberra was designed by Wal-



Population density of Australia.

ter Burley Griffin, an American architect of the Chicago School, as the result of a competition held for the purpose. Though Griffin's plan was later modified, Canberra is still a conspicuously well-planned city, with broad avenues lined with trees and an artificial lake in the centre of the city. Some of the most important public buildings include the National Library, the High Court, the National Gallery, and the Parliament Building.

For all its advantages of site and plan, Canberra has not quite escaped the main criticism levelled against Australian cities—that they have allowed suburbia to run riot. Because of the availability of land and the determination of the average Australian to own his own house and garden, all the main cities stretch for miles around their centres, putting an inevitable strain on public transport and services. The rising price of land, particularly in Sydney, has slowly brought a change to apartment dwelling and higher density housing, but most Australians still live in suburbs, and some critics have found in this an explanation for a certain suburban outlook in contemporary Australia.

Human geography

THE PEOPLE

The population of Australia is remarkably homogeneous. The largest percentage is of British heritage, while other Europeans make up most of the balance, and only a small percent of the total population is nonwhite. This includes more than 150,000 Aborigines, of whom only one-third are full bloods. There is also a significant Asian minority, including many refugees from Vietnam and a few thousand Chinese Australians, the descendants of Chinese coolies and diggers who entered Australia during the gold rush of 1851 before the White Australia policy was adopted.

Until the end of World War II almost all immigrants to Australia came from the British Isles. The ancestry of present-day Australians has been estimated as about half English, one-fifth Irish, and one-tenth Scottish, with only a few Welsh. (The remainder are of non-British origin, largely from continental Europe.)

After World War II it was decided to try to increase the population of Australia by encouraging migrants from other European countries, including refugees who had been rendered homeless by the war. Missions were sent to many of these countries, and government aid was extended to European as well as British immigrants. As a result, a steady stream of migrants arrived from Europe after 1945, about half of them from Britain.

Australia's immigration program has been remarkably successful. On the whole, European immigrants have been easily assimilated, and most apply for Australian citizenship after the statutory five years residence. Few return to their homelands, though, curiously enough, it has been estimated that one of every six British families who came to Australia after World War II returned to the United Kingdom.

The White Australia policy

This homogeneity, of course, was achieved only by enforcing a White Australia policy. Strictly speaking, there was no such policy. The phrase did not appear in any act of Parliament. In fact, from 1901 to 1966 all Australian governments saw to it that black, yellow, and brown immigrants (politely called non-Europeans) were not admitted to permanent residence in Australia. From 1966 onward, however, a limited number of non-Europeans and people of mixed descent were admitted.

In 1973 the Labor government of Gough Whitlam declared an end to all racial discrimination in immigration policy, though the actual numbers of non-Europeans admitted each year remained about the same. Since then a large proportion of immigrants have come from Asia, many from Indochina, the Philippines, and Malaysia. On the whole this historic change in Australia's traditional immigration policy, carried out by successive governments somewhat furtively and without much public debate, has been accepted by the majority of Australians.

Australia has one nonwhite minority that is too often forgotten. That is the Aborigines, who have survived two centuries of white persecution and indifference. (There is no true definition of an Aborigine, but anyone with a recognizable amount of Aboriginal blood tends to be counted as an Aborigine.)

Aborigines

These people have had little share in Australia's growth and prosperity. Most of the full bloods are to be found in the Northern Territory, the north of Queensland, and Western Australia, where some of them still live in tribal societies and a few—a very few—still follow the nomadic, hunting life of their ancestors. Others find appropriate employment as stockmen on the cattle stations. But most of them live in Aboriginal communities, outstations, mission stations, or government reserves, depending on social security or earning a little money as casual workers. The same is largely true of the part-Aborigines in the southern states, although they have had greater opportunities for employment. There are few Aborigines in Tasmania, where they were largely exterminated early in the 19th century.

Official federal policy is to help the Aborigines assume responsibility for their own affairs while improving the services available to them. So far, however, they have been held back by white prejudice or indifference, by their own deep suspicions, and, most of all, by the vicious circle of poverty, ignorance, and disease in which they are trapped. Their birth rate is high, but infant mortality during the late 20th century rose as high as five times that for white Australian children. Those who survive tend to suffer from malnutrition and other diseases, which gravely handicap their development at school and afterward.

Several measures that give some hope of improvement have been taken by the federal and state governments. All acts discriminating against Aborigines have been repealed except in Queensland; Aborigines now have the right to vote in all federal and state elections. Moreover, in 1976 the federal government for the first time granted the Aborigines living in the Northern Territory freehold over part of the territory's land area. Since then South Australia has granted the Pitjantjatjara people land rights over part of that state, and other states, some more reluctantly than others, have been considering legislation granting Aborigines the right to own the reserves on which most of them live.

The Aborigines themselves, whose numbers are increasing rapidly in spite of their high rate of infant mortality, are becoming more politically aware and more outspoken in their demands for land rights and full equality. While there is much sympathy for them among urban white Australians, there have also been some signs of reaction among white Australians living in country towns with a high proportion of Aborigines. It can no longer be claimed that there is no racial conflict in Australia.

Major religious groups

If Australia is largely white and predominantly Anglo-Saxon, it is far from being overwhelmingly Protestant. Many of the European immigrants have been Roman Catholic. Sectarian feeling between Roman Catholics and Protestants has at times played an important part in Australian political and social life, and it cannot yet be said to be entirely extinct. From the beginning, the Roman Catholic Church has played an important part in Australia, and its adherents make up about one-third of the total population, almost as large a portion as the membership of the Anglican Communion. There are also Jewish and Muslim communities in Australia, and in some areas of the country the Aboriginal religion survives. (J.D.Pr./Ed.)

THE ECONOMY

Australia's traditional world reputation as a predominantly farming country has changed. Its meteoric rise after World War II to a major mining nation has attracted much world attention. It has come to be realized that beneath this vast continent and its offshore areas lies a wealth of mineral riches. Australia has become one of the chief sources of minerals for the leading industries of Japan.

The minerals boom, the impressive progress of industrialization, and the growth of service industries and other features of a sophisticated economy have added new dimensions of economic strength and transformed the nation's traditional image of a predominantly agricultural and pastoral country with vast open spaces into that of

a bustling industrial nation. In the Pacific Basin area, Australia plays a vital role in trade, commerce, and investment. By the late 20th century, Australian living standards were comparable to those of the advanced economies.

Australia exports to all the important world markets and it imports goods from most countries, particularly the leading industrialized nations. (E.I.U.)

Resources. *Mineral resources.* The first settlers, who arrived in Australia at the beginning of the Industrial Revolution in Europe, soon reported the existence of coal near Sydney and elsewhere. Discoveries of alluvial gold west of Sydney, later in Victoria, and finally in Western Australia led to great movements of population across the continent and to extensive immigration. Prospectors discovered many rich mining areas throughout the continent, particularly in the west, the area around Adelaide, Victoria, Tasmania, western and central New South Wales, northern New England, coastal and western Queensland, and the Northern Territory. In 1949, Australia began a period of almost continuous mineral discoveries. This period continued into the late 20th century, the rate of discoveries and, more importantly, of mineral production, rapidly increasing.

Australia is endowed with resources of all the major minerals, particularly coal, oil, natural gas, iron ore, lead, zinc, copper, bauxite, uranium, tin, gold, silver, nickel, and beach-sand minerals. Australia's iron ore and bauxite reserves are among the world's largest. Its reserves of lead, zinc, copper, and the rare minerals rutile, zircon, and ilmenite are also very large by world standards.

Australia's significant mineral reserves are located in Western Australia (iron ore, nickel, bauxite, diamonds, and gold), Queensland (bauxite, lead, zinc, and silver), New South Wales (coal, lead, zinc, silver, and beach-sand minerals), and off the coast of Victoria (oil and natural gas).

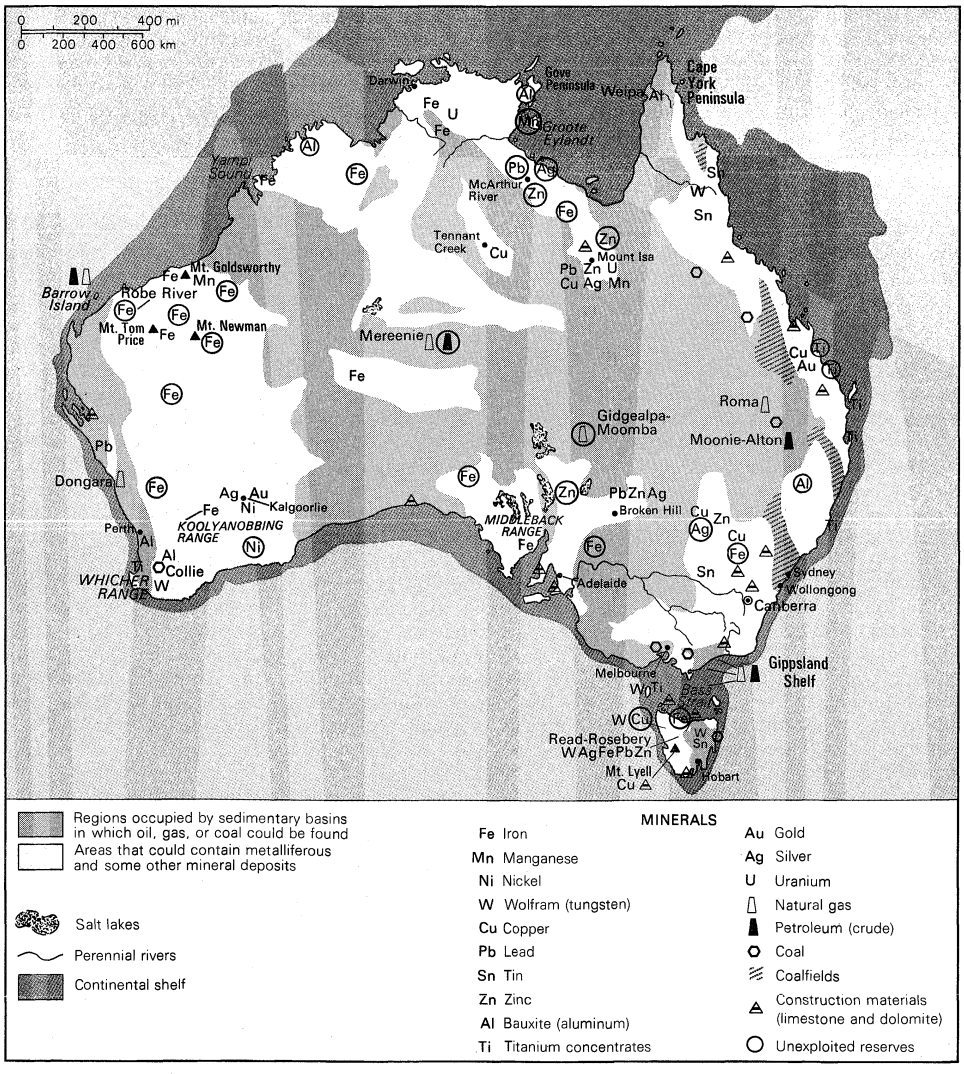
The coal resources of eastern Australia have attracted much interest from overseas buyers, and new mining ventures have been undertaken. Black coal is found in the Permian sediments—some 250,000,000 years old—in the Sydney Basin and in Queensland. Western Australia has a small Permian coal basin at Collie, and South Australia has a Triassic coalfield—dating from 200,000,000 years ago—at Leigh Creek. Large brown-coal deposits are exploited for electricity production at Yallourn and Morwell in Victoria, and west of Melbourne.

Commercial gas fields

The fuel supply situation in Australia changed fundamentally with the discovery of commercial gas fields and the opening of pipelines to capital cities in 1969. The Roma area of Queensland supplies gas to Brisbane; the Gippsland offshore area supplies Melbourne; gas from the Gidgealpa–Moomba area is brought to Adelaide and Sydney; while the Dongara field in the Perth Basin and the offshore Barrow Island fields on the northwestern coast are connected by pipeline with Perth. The Mereenie and Palm Valley fields of the Northern Territory are large gas reserves, located in the Ordovician sandstone (some 450,000,000 years old). The Palm Valley field is linked by pipeline to Alice Springs.

The first commercial oil discoveries in Australia were made, in 1961 and 1964, respectively, at Moonie and Alton, in the Surat Basin. By the late 20th century indige-

By courtesy of the Department of National Development, Canberra



Mineral resources of Australia.

nous production was meeting more than two-thirds of the continent's crude oil requirements.

Uranium. Post-World War II exploration efforts, linked to the world demands of a growing atomic energy industry, led to the discovery and opening of uranium mines in northwest Queensland, South Australia, and the Northern Territory.

Major uranium deposits

During the early 1980s the Mary Kathleen uraninite deposit, situated in northwestern Queensland between Mt. Isa and Cloncurry, closed down, leaving the Ben Lemond deposit, near Townsville on the eastern coast, as the country's major producer of uranium. The Radium Hill mine, located near Olary in South Australia, was based on a deposit discovered in 1906 and worked from 1954 to 1960, producing uranium in the form of davidite. In the Northern Territory four large deposits in the Alligator River area were discovered: the Jabiluka; Ranger; Koongara; and Narbleck. The first large discovery of a sedimentary uranium deposit in Australia was made at a shallow depth, under the Lake Frome plains in South Australia, and is clearly derived from the uranium-bearing Mt. Painter Complex in the adjoining northern Flinders Ranges. In Western Australia the Yeerlirrie deposit has sizable reserves.

Iron ore. Iron ore is produced in Australia for iron and steel production and for export; the physical and chemical properties of the iron ores control the value of the ore for modern technological processes. These processes facilitate the use of otherwise valueless fine-grained ore, although strict control of phosphorus and nickel content is required.

The old ore-producing areas are in the Middleback Range on Eyre Peninsula in South Australia, and in Yampi Sound and Koolyanobbing Range in Western Australia. The Middleback Range contains numerous ore bodies of banded hematite-jaspilite iron formation.

Iron ore discoveries

Discoveries are on a vast scale: the newly developed Hamersley iron province contains thousands of millions of tons of ore in iron formations. The largest high-grade ore deposits are Mt. Tom Price, with reserves totalling hundreds of millions of tons of ore; Mt. Whaleback; and Mt. Newman. Other ores occur in vast quantities in the Robe River area; reserves are estimated at a few billion tons, with an iron content of 56–57 percent. The Savage River iron deposits, in Tasmania, were also developed during the late 20th century.

Ferroalloy metals. Tungsten has been produced in Australia since the end of the 19th century, and in the late 20th century its contribution to world output has been substantial, generally exceeding one-twentieth of world production. It is produced from wolframite and scheelite found primarily on King Island (in Bass Strait) and in Queensland. Manganese has been mined from many small deposits, but between one and two million tons of ore are produced annually from the deposits on Groote Eylandt in the Gulf of Carpentaria. One of the most significant of the many economic mineral occurrences discovered in Australia is that of nickel. The Kambalda deposits, the country's largest producer of nickel, lying 35 miles south-southeast of Kalgoorlie, were discovered in 1964. Many other similar deposits in the old goldfields area of the Western Australian Shield were also explored. Later finds are located in the Windarra area, in the northern part of the nickel belt. The Greenvale operation in Queensland is the second largest producer of nickel. In addition, there are large areas of lower grade nickel ores in the Musgrave Block, near the borders of Western Australia, South Australia, and the Northern Territory.

Nonferrous base metals. The Australian continent has long been one of the world's principal producers of lead and zinc. The Broken Hill lode, in western New South Wales, discovered in 1883, was, a century later, still producing roughly half of the country's annual production of lead and zinc. At Mt. Isa, in western Queensland, important lead–zinc and copper ore bodies were discovered in 1923. During the late 20th century another lead–zinc deposit was developed in the McArthur River area of northern Australia, as was a deposit in Tasmania.

Copper production

Copper is widely distributed in the Precambrian and Paleozoic rocks of the continent, although most occurrences are small. Among Australia's principal producing copper mines in the late 20th century, Mt. Isa accounted for roughly three-quarters of the total output. Mt. Morgan in Queensland, Mt. Lyell in Tasmania, Tennant Creek in the Northern Territory, and Cobar in New South Wales produced most of the remaining copper ore. The South Australian copper mines, at Wallaroo–Moonta and Burra, although of great importance to the early economic development of that state, were maintained as reserve mines in the late 20th century. Vast resources of bauxite for aluminum production have been discovered at Weipa (on the Gulf of Carpentaria), at Gove (Arnhem Land) in northern Australia, and at the Darling Range in Western Australia. Tin is produced in eastern and western Australia and in Tasmania; it occurs in lodes connected with granites and as alluvial deposits. Rutile (titanium oxide) and zircon are heavy minerals that are extensively mined from beach sands, mainly on the east coast.

Precious metals. Gold discoveries—which first occurred near Orange, in New South Wales, in the mid-19th century, and later in Victoria, north Queensland, and Western Australia—were important stimulants for the growth of population in Australia. From a peak production of nearly 4,000,000 fine ounces in 1904, annual output had declined by the late 20th century to about one-seventh of that level, most of it coming from the Kalgoorlie–Norseman area of Western Australia. Silver occurs in the rich lead–zinc ores, most of it found in Broken Hill and Mt. Isa, and small amounts of platinum and palladium have been found during the search for nickel.

Nonmetallic deposits. As can be expected from its size, the Australian continent has abundant deposits of such industrial minerals as clays, mica, salt, dolomite, building materials of all kinds, refractories, abrasives, talc, and asbestos. An intensive search for phosphates to offset the declining production of Nauru and Ocean Island led to the discovery of large deposits in the Cloncurry–Mt. Isa area, but it has not been economical to develop these deposits.

Mining of gems

Gem minerals occur in many localities, and mechanized industrial prospecting and mining is in operation. The Australian white and "black" opals, mainly from Andamooka and Coober Pedy in South Australia and White Cliffs and Lightning Ridge in New South Wales, are world famous. The sapphires and topaz from Queensland and the New England district are also well known. In 1979 a vast deposit of diamonds was discovered in the Kimberley region of Western Australia.

Water resources. If not the oldest continent, Australia is certainly the driest. The average annual rainfall is less than 20 inches, with less than 10 inches falling in more than one-third of the continental area. The more significant index is of average variability from average annual rainfall, which is more than 20 percent, and reliable rainfall occurs only in the southeast, southwest, and in the north around Darwin. In the interior, the low and unreliable rainfall, the high evaporation rate, and the flat topography combine to reduce the streamflow. The catchment area of the Murray River system, the largest in Australia, is almost a third that of the Nile, but the average annual flow is only one-sixth (18,000,000 acre-feet). The control of water runoff and the general development of storage dams are of great importance in the management of water resources in Australia. An elaborate development scheme for the waters of the southeastern highlands for storage, irrigation, and power, the Snowy Mountains hydroelectric scheme, was completed in 1974; it is one of the greatest projects of this kind ever undertaken. Some of the waters of the Snowy River, which takes the meltwaters of snow from the eastern highlands straight to the sea, are turned into the Murray Basin to be used for irrigation. The total storage capacity of the many large dams of the Snowy Mountains scheme is more than 6,000,000 acre-feet.

The Snowy Mountains scheme

In the Great Artesian Basin there are about 25,000 recorded water bores. The quantity of water drawn from the basin apparently exceeds the rate of recharge, but the need for further investigations and reassessment of data on the hydrology of the Great Artesian Basin is recognized. The underground water resources of the Murray Basin and of other sedimentary basins are significant but insufficient-

ly known. Considerable underground water supplies are available for the development of Alice Springs in central Australia, and drilling has also supplied sufficient water for the mining developments in the northwest. An important problem—at least until desalination becomes economically feasible—is the quality of underground water; the water from the Great Artesian Basin bores is used only for livestock. The groundwater at Kalgoorlie, in Western Australia, is as salty as the sea, and at Norseman the salinity is twice as high. Investigations are in progress to promote the better use of Australia's water resources as knowledge of their limited amount increases.

Biological resources. The natural biological resources of the Australian continent, as distinct from introduced plants and animals, are limited. Native timber resources in the form of eucalypts are used for papermaking, and the jarrah and karri forests of southwestern Australia provide valuable sawmilling timber. In many parts of the country the virgin native plants still provide the principal pasture component, but in the more arid parts of the country overstocking has destroyed much of this resource and damaged the soil. The arid part of the continent carries only one-third of the total livestock, but introduction of types of cattle more suited to the country and a greater awareness of the limits of its carrying capacity have improved the utilization of these regions. In the zones of better climate, introduced pasture plants such as lucerne and clover, together with the use of superphosphate as fertilizer and of trace elements for soil improvement, have greatly improved Australian pasture resources.

Hydroelectric and other power resources. By the late 20th century Tasmania was one of the few states with sufficient water resources to permit the continuous operation of large hydropower stations. Its hydroelectric potential has been estimated to be half of Australia's total. Although the use of hydropower is expected to grow, the country's potential for this form of energy is small because of generally flat topography and low rainfall.

Petroleum and black coal are Australia's most important energy sources. About three-quarters of the total installed electric capacity comes from thermal power equipment; well over three-quarters of the electric power actually generated is thermal in origin.

Livestock, agriculture, forestry, and fisheries. Australia is the world's largest wool producer and a major supplier of cereals, dairy products, meat, sugar, and fruit. The vast majority of the wool and over a third of most of the other rural products are exported. In contrast with most other countries, by the last quarter of the 20th century more than 90 percent of the utilized land area in Australia remained in its natural state or was capable of very limited improvement; *i.e.,* was land used solely for rough grazing. The area cultivated for agriculture and intensive grazing is about 10 percent of all the utilized land. The main limiting factor in this respect is lack of water, but unsuitable soil and topography are also important determinants. Australia's rural holdings have a combined land area equal to about two-thirds of the total land area. (A rural holding in Australia is defined as a piece of land of one acre or more in extent used for the production of agricultural products or for the raising of livestock and the production of livestock products.)

Livestock. The sheep industry accounts for about one-half of Australia's total farming area. By the late 20th century Australian sheep made up about one-sixth of the world's wooled sheep, producing more than one-fourth of the world's wool supplies. Merino sheep form the basis of the Australian wool industry, making up about three-quarters of total sheep numbers. The Merino, which produces a fine quality wool, is well adapted to the extreme vagaries of climate in Australia's wool-producing areas. Australia's other types of sheep are crossbred and comeback varieties and dual-purpose Australian and British breeds, raised mainly for fat lamb production.

Sheep

A dramatic rise in producton since 1950 is attributable to a number of factors: success of a myxomatosis campaign; pastureland flock improvement; improved methods of pest and disease control; better management of sheep farms; and improved breeding techniques.

Sheep are run throughout Australia under a wide range of climatic conditions and environments, but about a third of Australia's sheep are reared in those areas with an annual rainfall of less than 15 inches (380 millimetres), generally known as the pastoral zone, where lack of water and fodder usually limits development. Dry conditions in the pastoral zone make the adaptable Merino the main breed. In areas of higher rainfall, up to 25 inches, sheep are often reared in conjunction with wheat and other cereals. About 40 percent of Australia's flock are found in these wheat–sheep zones, and breeds other than Merinos constitute a high proportion of the flocks. The rest of Australia's sheep are reared in the areas of relatively high and reliable rainfall, which produce most of the country's superfine wool. Victoria is the most important producer of mutton and lamb, using crossbred sheep raised in areas of high rainfall and fertility. Most of this type of grazing in Victoria is part of a mixed farming operation, where wheat is often the other principal product.

The breeding and fattening of cattle are usually carried out in different climatic zones. The better quality pastures in the relatively high rainfall areas are usually reserved for the fattening of cattle, which are often purchased seasonally and moved to rich natural pastures. Most of Australia's beef cattle are raised in Queensland and New South Wales in regions of warm and coarse pastures. In the northern parts of Australia herds of the more recently introduced zebu and Brahman cattle are thriving and increasing; they are well adapted to the generally poor tropical pastures and resistant to heat and insects. An introduction from the United States, the Santa Gertrudis breed, also offers good prospects of development. Australia's dairy industry is mainly located in the temperate zones of high rainfall in coastal New South Wales and Victoria.

Cattle

Agriculture. Wheat is the most important grain crop grown in Australia; other grain crops are barley and oats. Wheat is usually grown in the medium rainfall belt in all states, and it has become increasingly integrated with sheep grazing and cultivation of other crops. Wheat, barley, and oats are often grown on the same farm for grain and green fodder or hay for livestock. Most of these cereal crops are grown for grain. Sugar, Australia's most important crop after wheat, is grown from cane in coastal areas of Queensland and in northern New South Wales. The former area accounts for nearly all of the Australian sugar crop. The sugar industry is based on small farms.

Other important crops grown in Australia include tobacco, citrus fruit, grapes, apples, cotton, bananas, potatoes, and sorghum.

Forest and fisheries. Australia is estimated to have several millions of acres of commercial or potentially commercial areas of forest, in addition to large regions of low-grade forest suitable only for the production of small quantities of forest products for use in the immediate vicinity.

The main commercial forests are in the areas of high rainfall on the coast or near the coastal highlands of southeastern and eastern Australia, Tasmania, and the southwestern coast of Western Australia. The main type of tree in Australian forests is the eucalyptus, a broad-leaved genus that provides timbers of great strength and durability for building and packaging. The rain forests of the wetter regions, with many species of broad-leaved trees, are also important.

It has been estimated that there are about 2,000 species of fish (including freshwater species) in Australia and the waters surrounding it. The resources include mullet, cod, bream, perch, tuna, snapper, whiting, flathead, abalone, mackerel, and Australian salmon. New South Wales and Western Australia are the most important producing states, the latter being famous for its crayfish. There has not been much change in the types of fish caught in Australian coastal waters over the years, but lobster and crayfishing have been expanded. In 1978 Australia established a 200-mile fishing zone that encompassed the continent and the External Territories. By establishing this zone the Commonwealth assumed responsibility for the management of fisheries and other living resources within the region.

(E.I.U./M.F.G./Ed.)

Depression in rural industry

Industry. *Rural industry.* Although the rural sector still makes a substantial contribution to Australia's national income, its importance has declined sharply. By the late 20th century the gross annual value of rural production totalled about 5 percent of gross national product, a substantial drop compared with a 15-percent contribution to GNP during the middle of the century. Despite a declining rural labour force, agricultural output has increased by nearly 80 percent since World War II. One of the main reasons for this improvement is the use of fertilizers. Wool production accounts for almost 10 percent of the total export earnings. Because of recurrent droughts, wheat production has fluctuated. As a result, the value of beef production has exceeded that of wheat, and Australia has become one of the world's largest exporters of meat. Wheat is Australia's most important rural crop in terms of area and production. Forestry and fishing contribute only an insignificant proportion, less than 1 percent, of Australia's gross national product.

Mining and quarrying. The story of Australia's mineral industry is one of strong expansion and development. Australia has achieved self-sufficiency in most minerals of economic importance except for petroleum and natural gas. Domestic supplies of petroleum and natural gas, however, fulfill much of Australia's needs. There also are adequate reserves and production to allow exports of many minerals, and the contribution to national income and exports from the nation's mineral industry has increased at a spectacular rate. The two most important contributors to mineral output are the coal and iron-ore industries. Other minerals of major economic significance, as mentioned above, are silver, lead, zinc, the construction materials group (road metal, gravel, clays, limestone, and building stone), copper, gold, tin, mineral sands, nickel, and bauxite. Australia is one of the world's largest exporters of iron ore along with Brazil, Canada, and the Soviet Union. Coal and iron ore account for about three-quarters of the total value of primary mineral exports.

Manufacturing. A vital sector of the Australian economy, manufacturing, has grown spectacularly since 1950. Australia's limited home market will not support certain types of industry, such as the manufacture of highly sophisticated machines, heavy fabricating, certain types of electrical equipment, some industrial chemicals, and highly specialized scientific instruments and equipment. With the exception of these special classes of goods, however, the domestic market is well supplied with products of local manufacturers, which include the full range of items required in a modern country. Most of the industries are soundly based; only a few are carried out at an excessive cost, and require high tariff protection.

About one-fifth of the nation's workforce is employed in Australia's manufacturing industries. Since World War II the expansion of manufacturing industries has been widely spread throughout the industrial structure, but it has been particularly strong in the engineering, vehicle, oil-refining, petrochemical, and construction materials industries.

The motor vehicle industry is a vital sector of the Australian economy, directly employing more than 5 percent of the manufacturing workforce, and numerous other industries are dependent upon demand created by its level of activity. Its importance is underlined by the fact that Australia has one of the world's highest per capita motor vehicle ownership, while most of these vehicles are locally produced or assembled.

The motor vehicle industry

Iron and steel production is a virtual monopoly in the hands of one company, the Broken Hill Proprietary Co. Ltd. Other major manufacturing industries include oil refining, chemicals, textiles, and domestic appliances.

Energy. The main source of power used in Australian homes, factories, and other buildings is electricity. There has been a strong rise in the domestic use of electricity because of the increased use of appliances, extension of electricity supply into rural areas, and the high rate of home building. Thermoelectric power, based on either brown or black coal, is the main source of electrical energy in Australia.

Finance. Australia has a well-developed banking system broadly similar to that operating in Britain. The system is made up of the Reserve Bank of Australia (the central bank), the trading (commercial) banks, the savings banks, the Commonwealth Development Bank of Australia, the Australian Resources Development Bank, and the Primary Industry Bank of Australia.

The government-owned Commonwealth Banking Corporation operates a trading bank, a savings bank, and a development bank. Some of the savings banks are owned by state governments. These savings banks operate within the borders of a particular state only, and some offer checking-account facilities. The major banks in Australia operate under a branch-banking system, and most of them have branches throughout Australia.

The Australian Resources Development Bank was set up to provide finance mainly for the development of Australia's mineral projects. One of its aims is to provide finance in order to maintain a substantial proportion of Australian equity in these mineral enterprises. The Primary Industry Bank of Australia provides loans to primary producers for longer terms than are otherwise generally available.

The banking system in Australia is an important repository for savings, the most important source of finance, and the medium through which most commercial and financial transactions are settled. The trading banks provide two types of facilities: current accounts operated on by checks and payable on demand (similar to checking accounts of U.S. banks), and fixed deposits, which are lodged for specific periods.

The Reserve Bank of Australia applies its monetary policy to regulate the economy through the banking system, which usually bears the brunt of monetary and credit restraints during periods when the authorities consider it necessary to dampen inflationary pressures in the economy.

Another important source of finance in Australia is finance companies, which grew strongly after the 1950s, largely as a result of the restraining influence of official monetary policy on the banking system. All of the major Australian banks have substantial shareholdings in finance companies, which in some instances are subsidiaries of the banks. A number of foreign banks also have acquired interests in finance companies in order to enter the Australian financial market, because Australian government policy prohibits the establishment of foreign banks in Australia, the exceptions being the long-established ones.

Finance companies

Australia has a well-developed short-term money market. There are also several merchant banks performing such functions as underwriting, company flotations, mergers and takeovers of companies, portfolio management, lending, and general financial and allied services.

Other important institutions in the financial system are the insurance companies (fire, marine, and general), life insurance companies, credit unions, pastoral finance companies, and the building societies.

Each state capital has its independent stock exchange, but a nationwide coverage is achieved by reciprocal arrangements with brokers in other capitals and agents in country centres.

The system is well developed and sophisticated, and the stock exchanges provide a ready market for shares of listed companies and government securities. Quotations are distributed widely and quickly by special teleprinter channels and by radio and newspaper services.

Trade. By the late 20th century Australia's exports made up about one-seventh of the gross national product. Minerals contributed about one-fifth of Australia's export income; other important sources of export income were wool, meat, wheat, and manufactured goods. Australia's principal imports were machinery, transport equipment, chemical and petroleum products, foods and beverages, and crude oil and gas. Japan was Australia's most important export customer and a principal import supplier. Other important customers and suppliers were the United States, the European Economic Community countries (particularly the United Kingdom), New Zealand, and Asia, as a whole.

Administration of the economy. The Australian economy is essentially one based on private enterprise, although government plays an important role in influencing economic conditions.

Most capital expenditure in Australia is undertaken by private business enterprises and persons. There is only limited public ownership of business enterprises. Major enterprises owned and run either directly by federal or state governments or by government corporations include the post office, Australia's international airline, one of the two major domestic airlines, the Commonwealth Banking Corporation, the Australian Shipping Commission, and the electricity and gas distribution organizations in the various states.

The government regulates economic activity through fiscal policies embodied in its annual budget and through monetary policy. The Reserve Bank and the federal government's Treasury Department are the government's chief source of economic advice.

Taxation

In Australia taxes are levied by federal, state, and local governments. The main taxation authority is the federal government, which levies income taxes, customs and excise duties, and sales taxes. There are also a number of minor taxes imposed for specific purposes. There is no general duplication of taxes by the three tiers of government. The states impose a wide variety of taxes that include stamp duties, taxes on motor vehicles, payroll taxes, land taxes, liquor taxes, and probate duties.

The trade-union movement is well organized in Australia. Most workers are members of trade unions; in a number of instances membership is virtually mandatory. Industrial disputes are common and often disruptive in terms of loss of production. Australia has a unique arbitration system for settling disputes, but there have been signs of this system breaking down, and the trend is increasingly toward direct negotiations with employers. (E.I.U.)

Transportation. Because of the great size and small population of Australia, transport has always been costly and has absorbed an unduly high proportion of the total work force. Moreover, the main lines of road and rail transport were laid down in the second half of the 19th century,

Fragmentation of transport system

when Australia was a collection of separate colonies, each of which looked to Britain for most of its trade. The transport system was therefore designed to maintain this trade, with roads and railways radiating from the main ports. Little thought was given to internal transport between the separate colonies. An unfortunate relic of this is the existence of three different railway gauges in the continent. It was not until 1970 that it became possible to go by train from Sydney on the east coast to Perth in the west without changing trains.

In spite of these historical and geographical handicaps, Australia is fairly well equipped with roads and railways, especially in the southeastern states. It is one of the most highly motorized countries in the world.

Australia is almost entirely lacking in internal waterways. For a short period during the 19th century the Murray–Darling river system was widely used to transport produce (mostly wool) from the country districts of New South Wales and Victoria to the coast, but the variable flow of water in these rivers always made this method hazardous and unreliable. With the coming of the railways it was quickly abandoned.

On the other hand, Australia is in many ways ideally suited to air transport. The great distances, lack of mountains, and fine prevailing weather make flying safe and economical.

For the historical reasons already outlined, roads and railways are mainly the responsibility of the six state governments, although the federal government has taken over railway systems in some states. Shipping and air transport between the states are the responsibility of the federal government. Since 1946 the Australian Transport Advisory Council, including the federal minister for shipping and transport and the six state ministers for transport, has provided machinery for coordinating transport on national lines.

Component systems. As has been noted above, the main road networks in Australia radiate from the ports, and especially from the capital cities of Sydney, Melbourne, Brisbane, Adelaide, and Perth. Most of the states, aided by federal grants, have also made progress in sealing thousands of miles of country roads.

The most serious lack is still adequate highways between the state capitals. Both the Hume Highway between Sydney and Melbourne and the Pacific Highway between Sydney and Brisbane are far too narrow for the heavy traffic they carry. A federal government bicentennial program has been successful in making significant improvements to major highways.

The provision of adequate expressways and throughways in the great conurbations of Sydney and Melbourne has also fallen behind. Like the industrialized nations of the West, Australia is finding it hard to keep up with the insatiable demands of the automobile.

Railways

Five of the six state governments own and operate a railway system, and the federal government operates the Australian National Railways. Private rail companies operate in each state in order to serve mining, agricultural, and industrial complexes. The largest private railway operations are those serving iron-ore mines in northwestern Western Australia.

Australian railways have a hard struggle to pay their way against the competition of road and air services, but rail transport still plays an important part in the development of the Australian economy. New capital is being spent in providing standard-gauge lines between the different state capitals and industrial centres and in building new branch lines where mineral discoveries justify it, as in the case of northwestern Australia. Sydney and Melbourne have major electrified underground and suburban railway systems.

There are about 70 ports of commercial significance in Australia, of which the great majority are on the east coast. The rest of Australia is notably lacking in good natural harbours. The most important port is Sydney, which has one of the finest harbours in the world. Sydney (with Botany Bay) discharges and ships the largest amount of freight tons annually, followed by Port Hedland, Melbourne, Fremantle, Newcastle, Brisbane, Hay Point, Port Walcott, Gladstone, Port Kembla, and Port Adelaide.

Shipping is still the lifeblood of Australia. Most of this shipping is in foreign hands, although Australian companies have a monopoly of interstate trade around the Australian coast.

Air transport

Australia has one international airline, Qantas Airways, Ltd., which is owned by the federal government. It is one of the world's leading air carriers. Many foreign-owned international airlines operate regularly in and out of Australia. Sydney has long been the site of the main international airport, and a new terminal was opened there in 1970. The same year Melbourne opened a new international airport at Tullamarine.

Several airlines operate domestic services throughout Australia. The two major companies are Trans-Australia Airlines (TAA), owned by the federal government, and Ansett-Airlines of Australia (Ansett), owned by private enterprise. Competition between the two is strictly controlled and regulated by federal legislation. Australia is well served by its internal airways, which offer a service that compares favourably in safety, cheapness, and extent with any in the world.

ADMINISTRATIVE AND SOCIAL CONDITIONS

Government. The constitution of Australia may be described crudely as an amalgam of the constitutional forms of Great Britain and the United States. Like Britain it is a monarchy, and the king or queen of Britain is also the king or queen of Australia. Like Britain, too, the governments both of the Commonwealth and of the states are chosen from the majority party in Parliament. Like the United States, on the other hand, it is a federation, and the duties of the federal government and the division of powers between the Commonwealth and the states are laid down in a written constitution that can be altered only by a referendum that gains the consent of a majority of all the electors and a majority in at least four out of the six states.

Any disputes arising out of the constitution are decided by the High Court of Australia.

Although the monarch of Britain is also monarch of Australia, Australia is wholly independent. Except when the monarch is in Australia, his or her functions, which are

largely formal and decorative, are exercised by a governor general, who resides in the federal capital in Canberra, and by governors in each of the six states. Though formally the governor general and the governors are appointed by the monarch, they are invariably recommended by the Australian governments concerned, and in recent years there has been a growing tendency to choose Australians. By convention, the prime minister (the leader of the party or coalition of parties victorious in the general election) is the nation's chief executive.

Due to a past constitutional crisis that developed between a prime minister and the governor general, there have been proposals for various constitutional amendments that would define, and limit, the powers and duties of the governor general. Some Australians, although they are clearly in a minority, favour severing all remaining ties with Great Britain and declaring Australia a republic, thus abolishing the post of governor general altogether.

The constitution defines the form and duties of the federal government in some detail. The most important of these are defense, foreign policy, immigration, customs and excise, and the post office. Those powers not given to the federal government in the constitution—the residual powers—are left to the states. The state governments are the direct descendants of the colonial governments of the 19th century. Originally they owed their powers to an act of the British Parliament, but these acts are short documents that lay down only the form that the parliaments shall take, not the sort of legislation they may pass.

State and territorial government

In theory the states are sovereign, with real and important powers. They are responsible for justice, education, health, internal transport, and, indeed, for most of the things that most closely concern the citizen. In practice, however, their powers have been greatly weakened by the federal government's decision in 1942, at the most dangerous point of World War II, to collect all income taxes itself and then to reimburse the states on an agreed formula. This temporary arrangement has become permanent, in spite of many political and legal challenges, so that the states, deprived of any equivalent "growth tax," have become largely dependent on federal grants for their revenues.

The Northern Territory was established as a self-governing territory in 1978 when the Commonwealth government transferred most of its powers to the government of the territory. The government of the territory has powers similar to the governments of the states but has an administrator instead of a governor and a chief minister instead of a prime minister. Members of the Legislative Assembly are elected for a period of four years.

Australia is a true parliamentary democracy. Both the federal upper house (the Senate) and the lower house (the House of Representatives) are directly elected by universal adult suffrage, with a voting age of 18. All state lower houses are similarly elected.

The political process

Preferential voting (placing the candidates in order of preference on the ballot paper) is used for all parliamentary elections. In elections for the House of Representatives the alternative vote system is used; the 64 senators (10 from each state and two from each of the territories) are elected by proportional representation. Voting in both federal and state elections is compulsory.

Since federation the political struggle in Australia has been between the Australian Labor Party (ALP) and a coalition of anti-Labor parties under different names. For most of the years after 1949 the federal government was formed by a coalition between the Liberal Party, which broadly represents the interests of private enterprise, and the National Party (formerly the Country Party), which represents the interests of the farmers and graziers in rural constituencies. The ALP is a typical Western social-democratic party, based firmly on the support of the trade unions, and normally preferring practical reforms to socialist theories. It has, however, always had a left wing retaining a belief in Marxist principles.

Another political party of some significance, the Australian Democrats Party, was formed by a former Liberal minister, Donald L. Chipp, in 1977. The Australian Democrats Party appeals to many middle-class voters with liberal or radical views.

The Communist Party of Australia, split into several different factions, is insignificant in numbers. It has retained a disproportionate influence in some trade unions.

Voting in state elections does not necessarily follow voting in federal elections. During the long rule of the Liberal–Country Party coalition in the federal Parliament after 1949, Labor governments at different times held office in five of the six states. But with a few minor exceptions the lines of battle are the same, though local issues play a larger part. State Liberal and Labor governments, once elected, though on different sides politically, often find themselves in alliance against the Commonwealth either in defending states' rights or in seeking more money from the federal government.

Justice. The law of Australia is based on the common law of England, and many laws are identical with those laid down in acts of the British Parliament. The administration of the law is largely in the hands of the states, each of which has a series of courts culminating in a supreme court. Between them these courts have a comprehensive jurisdiction that extends to all matters of state, and to most matters of federal, jurisdiction. The states are also responsible for the police, although there is a federal police force that performs general police duties in the Australian Capital Territory and is the principal agency for the enforcement of federal laws.

The High Court of Australia, established by the constitution, is the federal supreme court. It has a general appellate jurisdiction over all other federal and state courts, and has a special duty to decide disputes involving the interpretation of the federal constitution and acts of the federal Parliament. The High Court of Australia has a high reputation among legal authorities both inside and outside Australia.

The federal Court of Australia was created to exercise the jurisdictions formerly exercised by the Federal Court of Bankruptcy and the Australian Industrial Court.

Armed forces. Australia has a proud military tradition dating from the landing of the first Australian Imperial Force on the Gallipoli Peninsula during the Dardanelles campaign of World War I. Since then Australian armed forces have served with distinction in World Wars I and II and in Malaysia, Korea, and Vietnam. The Royal Australian Navy, the Army, and the Royal Australian Air Force are separate services with distinctive uniforms and separate commands.

The Royal Australian Navy is a small but compact, flexible force with an emphasis on antisubmarine warfare.

The first aim of the army is to provide a highly trained field force for limited warfare in Southeast Asia. Compulsory national service was introduced in 1964 but was abandoned in 1972.

The Royal Australian Air Force has five operational elements: strike/reconnaissance, tactical fighter, air transport, tactical air support, and maritime forces.

Health, welfare, and education. Australia may be described as a modified welfare state. The federal and state governments have not gone so far as some European countries in providing free health and other services for its citizens but have gone further than the United States. Government benefits include pensions for the aged, invalids, widows, and single parents; unemployment and sickness benefits; health insurance provisions; maternity allowances; and child endowment. Benefits are paid from the National Welfare Fund, which is financed from consolidated revenue.

On the other hand, there is a strong tradition in Australia that one should buy or build one's own house. There are relatively few state or municipal houses and apartments for rent, though there are various federal and state financial schemes to help citizens purchase their own homes.

Education

Responsibility for education lies primarily with the states. Free education at government schools is provided at the primary and secondary levels, though there is also a considerable private sector formed by schools run by the churches (mostly Roman Catholic) and a few grammar schools for children of the wealthy. The principle of giving some state aid to Roman Catholic schools has now been accepted by all the main political parties.

Since 1974 tertiary education has been free both at universities and at colleges of advanced education that are oriented toward practical training and technology. Universities are mainly, and colleges of advanced education wholly, financed by the Commonwealth government. Australia has 19 universities, most of them located in the capital cities, and some colleges of advanced education, some of the latter quite small.

Health services are everywhere excellent. The method of financing these services has been changed several times. The states provide free hospital treatment in public hospitals, but patients may also insure to cover the cost of better accommodations, choice of doctor, and other variable costs.

Australia's most interesting contribution to the welfare of the ordinary citizen is the arbitration system, which has roused much interest in other countries. In brief this is an attempt, unique to Australia and New Zealand, to fix wages and working conditions by law.

Wage fixing

The constitution gives the federal government the right to undertake conciliation and arbitration in industrial disputes. Armed with this power, the first federal government set up the Commonwealth (now Australian) Conciliation and Arbitration Commission, which began by establishing a basic wage necessary to keep in "frugal comfort" the average family of an unskilled worker. From this beginning an elaborate system has been built up, involving both federal and state arbitration courts, conciliation commissioners, and wage boards, the aim of which is to prevent or mitigate industrial disputes.

If a dispute cannot be solved by collective bargaining or conciliation, then either the employers or the trade union concerned may take the dispute to the relevant court for a judicial decision that has the force of law. Strikes are not forbidden, but a union that strikes in defiance of a judicial award may be held to be in contempt of court and fined accordingly. In practice, therefore, the judges on the Australian Conciliation and Arbitration Commission, after hearing argument from both sides, fix minimum wages and conditions over a large section of Australian industry.

This system has frequently been criticized as excessively cumbrous and, more recently, complaints have been heard that the judges who make the awards are not always equipped to appreciate the economic consequences of their decisions. On the whole, however, it has worked reasonably well over the years though it has certainly not prevented industrial unrest in Australia. It is still supported, with reservations, by both employers and trade unions.

Both Australia's welfare services and arbitration system spring from a deep concern for the common man. From the earliest days Australians have been strongly egalitarian in outlook, quick to resent any claims to privilege either by a class or an individual. This has not prevented class distinctions in Australia—there is a stronger class system than is often admitted—or wide differences in wealth, but it has greatly eased the conflicts inherent in a capitalist society. Australian trade unions are as militant as any in the world in the pursuit of higher wages and better conditions, but the Australian working man is well paid, well cared for, and living in a prosperous, democratic, and expanding country. He is widely judged to be no wage slave but rather the master of his own fate.

CULTURAL LIFE

The cultural milieu. For the first 100 years of white settlement the arts were naturally neglected. Settlers were too busy exploring and developing this harsh land to have much time or energy left for the graces of life. This neglect was never total. There were always Australians who wrote poems or novels or who painted the landscape of their adopted country, but for the most part they were content to take their art and culture secondhand from England.

Early developments in the arts

The first sign of an Australian consciousness in the arts was the emergence of a small group of writers and artists in the 1890s associated with the Sydney *Bulletin*. Most of them—though not all—tended to express the radical, egalitarian, and nationalist views that were then beginning to stir political life in the colonies. Of these writers the most important were Joseph Furphy (1843–1912), author—under the pseudonym Tom Collins—of the long novel *Such Is Life* (1903), and Henry Lawson (1867–1922), whose short stories are still worth reading. The *Bulletin* also provided an opportunity for a remarkable collection of artists working in black and white, among whom Norman Lindsay (1879–1969) was the best known.

But Australia's great leap forward in the arts did not take place until World War II, when the growing wealth and sophistication of the cities, and isolation from Britain and Europe, provided a powerful stimulus. This movement began with painting and poetry and later spread to all the arts in turn. Since 1954 the federal and state governments have subsidized the arts on an increasing scale.

The arts. The first of the arts to attract attention outside Australia was painting. Starting in Melbourne during World War II, a group of painters—of whom George Russell Drysdale, Sidney Nolan, and Arthur Boyd are the best known—developed an original school of Australian painting by means of adapting contemporary techniques to depict Australian myths and Australian landscape. Since that time the influence of that school in Australia itself has dwindled, and young Australian painters prefer to paint in the international styles of their contemporaries in Paris, London, and New York City; in spite of this, their work is distinguished by a freshness and vigour that are characteristic of their young nation.

Australian poets have also revealed a new maturity and sophistication, though, unlike painters, they have tended to turn their backs on the more revolutionary developments overseas and to prefer traditional metrical forms. This may be due to the powerful influence of A.D. Hope and James McAuley (1917–76), both professors of literature as well as poets, who consciously set and maintained classical standards in their own work. More recognizably "modern" in their very different ways are Geoffrey Lehmann and the remarkably gifted Les A. Murray, who almost alone has managed to find an authentic Australian idiom. But many other younger Australian poets, such as Bruce Dawe, Bruce Beaver, Peter Porter, and David Malouf, deserve mention.

Australian novelists have, on the whole, been less original and creative, though Patrick White, who won the Nobel Prize for Literature in 1973, has received acclaim worldwide for his lyric and visionary contemporary novels *The Tree of Man* (1955), *Voss* (1957), and *Riders in the Chariot* (1961). Thomas Keneally has enjoyed international success with his historical novels, *Gossip from the Forest* (1975), *Confederates* (1979), and *Schindler's Ark* (1981). Christina Stead (1902–83) and Shirley Hazzard, two other distinguished novelists, have not always been recognized as Australian because of their long residence in the United States.

The performing arts are more dependent on financial support and suffer from the fact that outstanding artists can always go overseas to Britain or the United States where a wealthier market offers greater rewards. This is particularly true of actors and singers, many of whom, from Dame Nellie Melba to Dame Joan Sutherland, have won international fame. In recent years, however, increasing financial support, the growing wealth and population of the larger cities, and the building of Sydney's spectacular Opera House (designed by the Danish architect Jørn Uzton) and new arts centres in Melbourne and Adelaide have led to a remarkable revival of the theatre in Australia, which has provided opportunities for dramatists, such as David Williamson, as well as for actors, singers, and dancers. Australia now has first-class national opera and ballet companies, and the Australian Broadcasting Commission (ABC) maintains symphony orchestras in all the capital cities.

Even more remarkable has been the revival of the moribund Australian film industry, which has been turning out a number of excellent films that have won international recognition. Australia's young film industry has also produced one director, Peter Weir, of outstanding talent. His best known films include "Picnic at Hanging Rock" (1976), "Gallipoli" (1980), and "The Year of Living Dangerously" (1983).

The Australian popular arts follow too closely develop-

ments in Britain and America to deserve special mention, and the remarkable Aboriginal arts, notably dancing and painting on bark, must be considered dead though efforts are being made to preserve and revive them.

Cultural institutions. The first serious attempt to organize the arts in Australia was the formation of the Elizabethan Theatre Trust in 1954. It was this body that formed the Australian Opera and Australian Ballet companies. The trust still administers the opera and ballet orchestras, but its main functions have been taken over by the Australia Council, established in 1975, which advises the Commonwealth government on the arts and allots government funds to various bodies. The Arts Council of Australia, a separate and independent body, has a division in each state that takes the performing arts to school and adult audiences in rural areas and coordinates community arts activities.

The Australia Council

Australia has many festivals devoted solely or largely to the arts, of which Perth's annual festival and Adelaide's biennial festival attract both overseas and Australian artists and companies.

The state governments have long provided museums and art galleries in their respective states. In 1982 the National Gallery was opened in Canberra to house the national collection. Since it is accepted that it is now impossible for such a gallery to compete with the great galleries of Europe and America in collecting European painting and sculpture from the past, this gallery concentrates on Australian art past and present, art of the Pacific and Asian areas, and international art of the 20th century.

Press and broadcasting. The press of Australia is free, independent, competitive, and vigorous. It has long reached high standards in such papers as *The Sydney Morning Herald* and the Melbourne *Age,* both of which date from the mid-19th century, though Australia has also always had and still has popular papers that are more noted for sensationalism than for sober reporting or informed comment. All of these papers circulate only in the states in which they are published. In 1964 *The Australian* was first published with the aim of becoming a national paper. It is printed in several states and has won a small but influential readership. The Australian *Financial Review* (daily) and the *National Times* (weekly) also have a national readership.

Broadcasting and television are shared between the ABC, a national service dependent on federal government grants, and a number of commercial radio and television stations operating under licenses granted by the postmaster general and dependent on advertising for their revenues. In general it may be said that while the ABC provides a high quality service, especially in the broadcasting of music, it attracts only a small part of the available audience. The commercial stations, on the other hand, attract large audiences with rather second-rate programs. Commercial television in Australia is heavily dependent on material imported from Britain and the United States. (J.D.Pr.)

HISTORY

Aboriginal Australia

The ancestors of the Aborigines of Australia arrived on the continent some 25,000 years ago. These people, whether in one or several periods of migration, probably travelled by way of the now submerged Sahul Shelf or, where land connections were absent, by rafts or canoes. They came from Southeast Asia, but whether they derived from one or more racial stocks is still unknown. They brought with them the dingo, a species of dog. Before the arrival of Europeans in Australia, Aborigines occupied the northern and eastern coasts of the continent, the Murray River Valley, and part of Tasmania. The estimate of their numbers at that time is about 300,000. Population density was highest in more fertile riverine and coastal areas; in arid zones, the exigencies of gaining a livelihood made distribution over much larger territories necessary. These broad environmental differences had direct sociocultural implications for mobility and for control over natural resources. Aborigines in the drier inland moved camp more frequently, had a narrower margin of survival in bad seasons, and relied more directly on ritual supplication as a necessary aid to everyday living.

There were approximately 500 tribes and subtribes—a tribe being a constellation of interacting groups, the members of which acknowledge certain features in common. Most tribes had specific names for themselves; other names, sometimes less flattering, were applied to them by their neighbours. They occupied recognized territories and were distinguished by features in culture or language.

Early writers on the subject of Aboriginal Australia spoke of "nations," usually referring to a group of tribes identifying themselves by the same word for "man." Although there did exist interaction between neighbours, there was no direct social communication across the continent, and the Aborigines were certainly not a nation in the modern sense of the word: they had no overarching political system and no overall name.

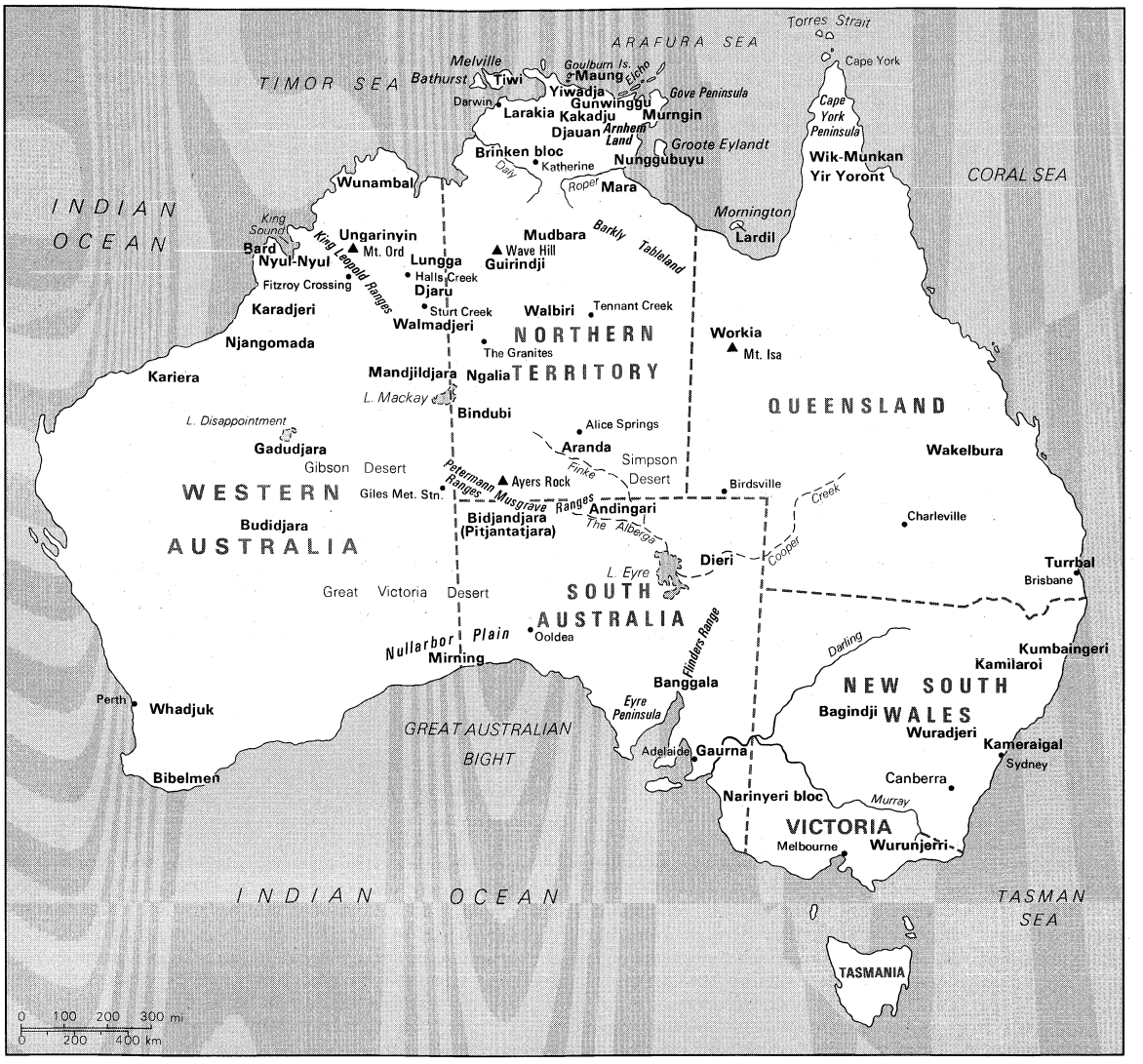
TRADITIONAL SOCIOCULTURAL PATTERNS

All Aborigines, coastal and inland, were directly dependent on their natural environment. They were seminomadic, wandering across limited stretches of the country, hunting, and collecting food. This basic economic circumstance had several concomitants: (1) food-gathering groups were fairly small and usually scattered, and all members of a tribe rarely came together, even for ritual occasions; (2) cooperation was essential; (3) over and above particular techniques and skills, Aborigines believed that religion was required for survival.

The Aborigines' view of their place in the world was summarized in their concept of "the Dreaming," or "eternal Dreamtime." This does not refer to dreams in an ordinary sense. In one of its meanings the Dreaming was the formative or creative period at the dawn of time, when mythic beings shaped the land, brought various species into life, and established human life and culture. Although these mythic beings died or were transformed, they lived on eternally in spirit. They left tangible evidence of their physical presence on earth and were actually identified with or manifested through particular species and natural elements. The Aborigines also believed that these spirits were manifested through and in human beings, that the mythic creatures of the Dreaming were not basically different from humans, and that humans, supernatural beings, and natural species were interdependent and mutually sustaining. Hence, Aborigines were never isolated; they saw themselves as acting with others, and the bonds of kinship were extended outward, embracing the nonhuman and nonempirical world.

Aboriginal cosmogony

Social groups and categories. The principal issues in Aboriginal social life were religion and economics; these elements were regarded as interdependent and were reflected in the two kinds of social units to which all Aborigines traditionally belonged. One of these social units was the local descent group, which was always patrilineal (descent through the male line), exogamous (marrying outside the group), and associated with a particular stretch of territory—the "estate." The adult males of a local descent group were responsible for the upkeep of the traditional sacred sites and for the appropriate ritual—that is, for renewing and sustaining the land. Ownership of land was nontransferable: its members held land in trust collectively by means of an unwritten charter deriving from the Dreaming. The second, larger kind of unit was the horde or band, the group concerned with the ordinary practical business of getting a living. This was a land-occupying and land-utilizing group, exploiting all the available natural resources. Fluctuating in size and composition, it comprised members of several families—that is, of more than one local descent group; and the range over which they normally travelled included several estates.



Original distribution of the larger Aboriginal tribes.

In addition to membership in a local descent group and horde, individuals identified themselves by reference to larger social groups: moieties, semimoieties, sections, subsections, and clans. A moiety is one of two basic complementary parts into which a tribe divides itself. Sections and subsections are further subdivisions, splitting the tribe into four and then eight parts. Such social divisions among the Aborigines governed marriage and other contracts. A person in one moiety, for instance, could marry only a person in the other moiety. When there were section or subsection systems, everyone was incorporated in such a system from before birth. The main advantage was that it represented a widely understood grid for categorizing everyone a person was likely to meet. With this went a shorthand statement of basic rules or conventions, which was a simple guide to behaviour, especially useful for classifying distant kin and strangers. Close, genealogically traceable kinship virtually always took precedence over nominal kinship when priorities were at issue.

Aboriginal Australia was not socially stratified in the sense of having more or less fixed classes or strata arranged hierarchically on the basis of descent or acquired status. Status was clearly marked only within the religious sphere; females, for example, were mostly excluded from an executive role in secret-sacred ritual, and areas of privilege were further defined by graded acceptance of youths and adults passing through rites of learning. Essentially, however, Aboriginal society was "open"—there were no social barriers operating against a man becoming a leader in religious matters. Such achievement depended primarily on his own efforts, his kin's, and his observance of ritual. Men acquired prestige through knowledge of religious practices and expertise in directing ritual. Entrepreneurial leaders had considerable status in one or two areas—among the Tiwi and Murngin, for example—but this was derived not so much from commercial activities as from their ability to manipulate certain kin, to arrange polygynous marriages for themselves and remunerative betrothals for their children, and to establish a network of alliances.

Kinship, marriage, and the family. As noted, the Aborigines were deeply conscious of social relations, especially kinship. All Aboriginal kinship systems were basically classificatory—that is, a limited number of terms was extended to cover all known persons: terms for lineal relatives, such as "father," also referred to collateral relatives, such as uncles. This did not mean that one could not recognize his actual father, mother, son, or other relative. The terms simply suggested the appropriate behaviour—filial, brotherly or sisterly, fatherly, or whatever.

Function and significance of kin terms

Kin terms were behavioral signals, indicating, for example, the expectation of sexual access, restraint, avoidance, or responsibilities. Affines (relatives by marriage) were often classified with consanguineous (blood) relatives, although qualifying terms might be used. Certain terms indicated potential spouses or affines. A husband and wife were ideally, before marriage, always related to each other as kin, either actual or classificatory. Some relationships were more prominent than others. The brother–sister relationship, for example, was one of the most emotionally charged and often marked by some form of avoidance. The most outstanding avoidance relationship was between a man and his actual or potential mother-in-law—not just

his wife's mother but all women and girls who on the basis of kinship and the recognition of preferred potential spouses, might become his mother-in-law.

Reciprocity was a fundamental rule in Aboriginal kinship systems and also in marriage. Marriage was not simply a relationship between two persons; it linked two families or groups of kin, which, even before the union was confirmed and most certainly afterward, had mutual obligations and responsibilities. Generally, throughout Aboriginal Australia those who received a wife had to make repayment either at the time of marriage or at some future time. In the simplest form of reciprocity, men exchanged sisters; and women, brothers. Such exchanges took place between different moieties, clans, or local descent groups or between certain types of kin. Most kinship-and-marriage systems provided for the possible replacement of spouses and for parent surrogates. At any one time, a man or woman had available a number of persons, who, because of their kinship, were termed spouses. Access to them depended on several factors, but, as far as premarital and extramarital liaisons were concerned, it was conventionally in this direction that partners should be sought.

Infant betrothal was common. If arranged before the birth of one or both of the prospective spouses, it was a tentative arrangement subject to later ratification, mainly through continued gift-giving to the girl's parents. In some Aboriginal societies, parents of marriageable girls played one man against another, although this was always a potentially dangerous game. Also, there might be a considerable age discrepancy between an affianced pair. Generally, a long-standing betrothal, cemented by gift-giving and the rendering of services, had a good chance of surviving and fostering a genuine attachment between a couple.

For a marriage to be recognized it was usually enough that a couple should live together publicly and assume certain responsibilities in relation to each other and toward their respective families; but it might be considered binding only after a child was born. All persons were expected to marry. A girl's marriage should be settled before she reached puberty, and, ideally, a husband should be older than his wife.

Apart from formal betrothal, there were other ways of contracting marriages, such as elopement, capture during feuding or fighting, and redistribution of widows through the levirate (compulsory marriage of a widow to her deceased husband's brother) or patriate (distribution of a father's widows to his sons, unless they were actual or close mothers). Elopement was often supported by love magic, which emphasized romantic love, as well as by the oblique or direct approval of extramarital relations.

Polygynous marriage

Although most men had only one wife at a time, polygyny was considered both legitimate and "good." The average number of wives in polygynous unions was two or three. The maximum, in the Western Desert, was five or six; among the Tiwi, 29; among the Murngin, 20 to 25, with many men having 10 to 12. In such circumstances, women had a scarcity value. Having more than one wife was usually a matter of personal inclination, but economic considerations were important; so were prestige and political advantage. Some women pressed their husbands to take an additional wife (or wives), since this meant more food coming into the family circle and more baby-sitters. To terminate a marriage, a woman might try elopement. A man could bestow an unsatisfactory wife on someone else or divorce her. A formal declaration or some symbolic gesture on his part might be all that was necessary. In broad terms, a husband had more rights over his wife than she had over him. But, taking into account the overall relations between men and women, their separate and complementary spheres of activity in marriage and in other aspects of social living, the status of women in Aboriginal society was not depressed.

Socialization. A child's spirit was held to come from the Dreaming to animate a fetus. In some cases, this was believed to occur through an action of a mythic being who might or might not be reincarnated in the child. Even when Aborigines acknowledged a physical bond between parents and child, the most important issue for them was the spiritual heritage.

In his early years a child's focus was on his actual parents, and especially on his mother, but there were others close at hand to care for him. Weaning occurred at about two or three years of age but occasionally not until five or six for a youngest child. Through observation of camp life around him and informal instruction, a child built up his sociocultural perspective, learning through participation. At the same time, he became familiar with his natural environment. Small children often went food collecting with their mothers and other women; as girls grew older, they continued to do so, but boys were thrown more on their own resources. Parents were, on the whole, very indulgent. Infanticide, even in arid areas, was much rarer than has been suggested.

Children learned quite early with whom they were allowed to play and whom they must avoid; brothers and sisters, for instance, or "mothers-in-law" and "sons-in-law," would not normally play together. Betrothal put special restraints on a girl, and, at or even before puberty, she normally went to live with her husband, assuming the status of a married woman.

With initiation, a boy's life changed drastically. His formal instruction as a potential adult began, and he was prepared for his entry into religious ritual. His future was henceforth in the hands of older men and ritual leaders who exercised authority in his community. But he was not among strangers: the relatives who played an active role in his initiation would also have significant roles in his adult life. A boy's age at the first rite varied: in the Western Desert it was about 16; in the Kimberleys, about 12; in northeastern Arnhem Land, six to eight; and among the Aranda 10 to 12 or even older. Generally, once he had reached puberty and facial hair had begun to show, he was ready for the initial rituals.

Initiation in Aboriginal Australia was a symbolic reenactment of death in order to achieve new life as an adult. As a novice left his camp, the women would wail and other noises would be made, symbolizing the voice of a mythic being who was said to swallow the novice and later vomit him forth into a new life. The initiation rites themselves were a focal point in discipline and training; they included songs and rituals having an educational purpose. All boys were initiated, and traditionally there were no exceptions.

Initiation rites

Circumcision was one of the most important rites over the greater part of Australia. Subincision (incisure of the urethra) was especially significant in its association with secret-sacred ritual. Other rites, according to the area, included piercing of the nasal septum, tooth pulling (in New South Wales this was central in initiation), and the blood rite, which involved bloodletting from arm vein or a penis incisure—the blood being used for anointing or sipping (red ochre was used as a substitute for blood in some cases). Hair removal, cicatrization (scarring), and playing with fire were also fairly widespread practices. Among the Aranda, fingernails were torn out and heads gashed open and bitten. All such rites were usually substantiated by mythology.

For girls, puberty was marked by either total or partial seclusion and by food taboos (also applied to male novices). Afterward, they were decorated and ritually purified. Ritual defloration and hymen cutting or both were widespread (in the Western Desert, for instance), while on Groote Eylandt the labia majora were removed.

Boys, after circumcision, became increasingly involved in adult activities. Although they were not free to marry immediately, even if they had reached puberty, they might do so after undergoing certain rites, such as subincision. Initiation was a prelude to the religious activity in which all men participated. It meant, also, learning a wide range of things directly concerned with the practical aspects of social living. Adulthood brought increased status but added responsibilities. A vast store of information had to be handed down from one generation to the next. Initiation served as a medium for this, providing a basis of knowledge upon which an adult could build. This was a process that continued through life and that was especially marked in men's religious activity.

For Aborigines birth and death were an open-ended continuum: a spiritual religious power emerged from the

Dreaming, was harnessed and utilized through initiation (as symbolic death–rebirth) and through subsequent religious ritual, and finally, on death, went back into the Dreaming. Life and death were not seen as being diametrically opposed: the Dreaming provided a thread of life, even in physical death.

Social control. Pressures toward conformity began and were formalized during initiation. There were rules and standards, right and wrong ways of doing things, and some allowance for variation. Sanctions upheld the accepted codes and were either positive or negative or a combination of both. Probably the strongest were fear of supernatural punishment and fear of sorcery.

Traditionally, most dissension arose over women, religious matters, and death. Women had trouble with husbands, eloped, and engaged in unsanctioned extramarital liaisons. Such behaviour could mean serious fighting, involving relatives of the parties concerned. Infringement of sacred law was less direct in its social repercussions but was nevertheless regarded as the most serious of all. In many cases, an ordinary or accidental death had wide ramifications, particularly if there were accusations of sorcery. An inquest was held, and, through divination, a supposed "murderer" was found. Punitive measures might or might not be taken against him.

Maintenance of law and order

The maintenance of law and order was quite narrowly localized. Authority was limited and qualified by kinship claims. Precedents were sought in order to guide or influence actions resulting from a breach, and in all tribes there were approved procedures for maintaining the peace. There were no judicial bodies as such, though on the lower Murray River a formal council, or *tendi,* of clan headmen and elders did arbitrate disagreements between adjacent tribes. Generally, simple informal meetings of elders and men of importance dealt with grievances and other matters. There was also settlement by ordeal—the most outstanding example of this sort being the *magarada,* or *maneiag,* of Arnhem Land. During a ritualized meeting, the accused ran the gauntlet of his accusers, who threw spears at him; guilt was proved if the accused was wounded in the thigh.

Although it is inaccurate to speak of a gerontocracy in Aboriginal Australia, men of importance were easily distinguished. They were usually "elders," who had this status not necessarily because of their age or grey hair but because of their religious position and personal energy. This gave them considerable control over the affairs of others in matters not directly related to religion.

Economic organization. Dependence on the environment and on its resources was reflected in seminomadic living. The Aborigines had to be intimately acquainted with all the country within their range of movement. At the same time, their economic system was not conceptually separate from their religious system. Both before and after initiation, graphic stories (or myths) served as guides to specific localities, detailing what people could expect to find there. In other words, information used in obtaining a living was preserved and transmitted in a religious context. Religious ritual, furthermore, was believed to ensure the maintenance of the natural species and the proper fluctuation of the seasons.

Before modification of their economic organization as a result of contact with Europeans, large groups of people did not remain in the same camp over a prolonged period. There were two basic patterns. Generally, in fertile regions of Aboriginal Australia, time-honoured camping areas were recognized, always in close proximity to water and usually with mythological associations; they were places where people always camped at certain times of the year. Camps were bases from which people made forays into the surrounding bush for food, returning in the late afternoon or spending a few days away. The second pattern involved a much larger territory in arid or desert areas across which Aborigines moved from waterhole to waterhole along well-defined tracks in small family groups. The whole camp moved and rarely established bases. Only in good seasons and at sizable permanent waters was it possible for a large number of people to remain for an extended period.

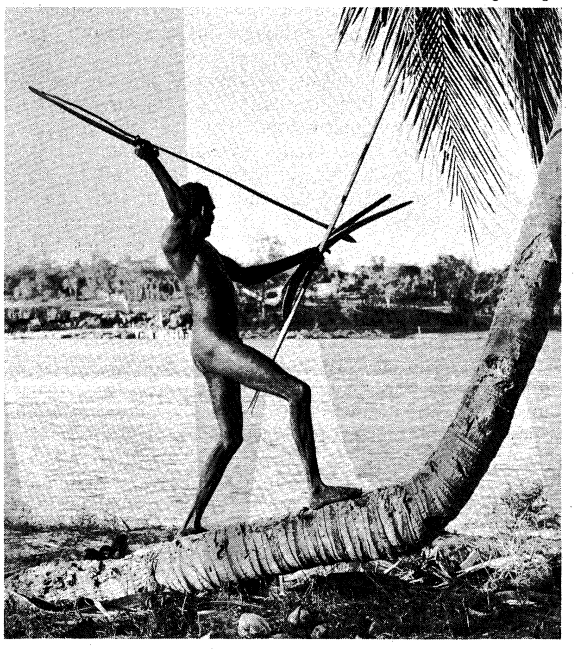
These two patterns were reflected in domestic arrangements. In the north, people made bark shelters, and during the monsoonal rains used caves and stilted huts as protection against flood, mosquitoes, and sandflies. In the desert, windbreaks—bough shelters or saplings covered with brush or bark—were common. During fine weather most Aborigines preferred to sleep in the open, with a windbreak; when it was too cold, dogs helped to provide warmth. Fires were kept alight, and, when moving from one place to another, or even when hunting, people carried live fire sticks.

Aborigines, with rare exceptions (southwest of Darwin and northeastern Arnhem Land, for example), had no form of cultivation and no domestic animals, except the tamed dingo. They preferred to go naked, except for small pubic coverings or decorations. In certain areas, fur cloaks were a protection against cold.

Tools and artifacts

Outside the sphere of religion, material objects were minimal. Grinding stones and platters were left at certain camps, ready for use. Men carried spears and spear throwers and, in some areas, boomerangs. There were bark canoes and rafts and dugout log canoes, some with pandanus-mat sails. Women's digging sticks could double as fighting weapons. Their large deep wooden dishes held seeds, vegetables, or water, even babies. In some areas, painted bark baskets, plaited pandanus bags, and net bags served the same purposes. Rarer objects were the kangaroo-skin waterbags of the arid central areas and the skull drinking vessels of the Coorong in South Australia. Implements included a large selection of stone tools, wedges, bone needles, bobbins, and sharkskin files.

Douglass Baglin



Aboriginal spearfishing near Darwin, Northern Territory.

Generally, men were hunters, women food collectors; and, unless they were on the move, they carried out these tasks separately. A woman's economic pursuits were focussed on her own family and children. A man's obligations were wider. Providing for his own family was often of secondary importance: ideally, his ritual duties and his network of reciprocal obligations took priority. He was expected to provide meat for his family; but women bore the main responsibility, supplying vegetable foods and small game or shellfish. Although this was primarily a subsistence economy, even life in the desert was not reduced to a constant struggle for the bare essentials. It nevertheless called for skill, tenacity, and continuing social involvement. A person could not "opt out" and survive. Reciprocity was basic to economic life, and all adults were caught up in a web of rights and duties.

As far as trading was concerned, goods passed along defined routes from one group to another in an intricate crisscross patterning over the continent. Boomerangs, for

example, went in one direction, red ochre in another; pearl shells from the Kimberleys found their way, gradually, to the Great Australian Bight; and central Australian shields appeared in the Canning Stock Route area.

In traditional Aboriginal society acquisition and gain were important, but the idea of making a profit was almost irrelevant. What really counted was the social relationship inherent in the exchange. The more organized trading partnerships entailed a debtor–creditor association; but prestige still hinged on retaining an equitable balance in the exchanges, on "solvency," and on not dropping out of the network. Private property was recognized, the claims of others upon an item not affecting an individual's right to exclusive use of it until such time as he decided to relinquish it. Traditionally, however, land was inalienable, being held in trust by members of various local descent groups.

BELIEF AND AESTHETIC VALUES

Religion. The Aborigines were concerned that the seasons should come and go in an orderly, predictable way; that human life should continue; and that they themselves should go on living in much the same way as they had always done. Religious beliefs and values gave tradition its power and force in influencing life in the present.

Totemism

Both before and since the French sociologist Émile Durkheim's study of *The Elementary Forms of the Religious Life* (1915), much has been written about Australian totemism. The concept rests on the belief that human beings are an integral part of nature, like all other living things. Totemism has been defined as a representation of the universe, seen as a moral and social order; as a philosophy that regards man and nature as one corporate whole; or as a set of symbols forming a conventional expression of the value system of a society. Such symbols provided intermediate links, both personal and social, between man and the mythic beings. Many of the mythic beings in Australia were "totemic" in the sense of exemplifying in their own persons, in their outward form, the common life-force pervading particular species. Others, originating in human or near-human form, at the end of their wanderings entered some physiographic feature or were metamorphosed as hills or rocks or turned into various creatures or plants. Totemism is another illustration that a sharp dichotomy between sacred and mundane cannot be drawn in Aboriginal Australia, where religion was directly relevant to everyday living.

Among most Aboriginal tribes, the main ritual roles in the big religious sequences were reserved for men. Ordinarily, in such cases, women were forbidden to enter ritual grounds, and what went on there was secret and sacred. But women usually had complementary actions to perform or observed special taboos. In other rituals both men and women participated. Among some tribes, women had their own secret-sacred rites. Generally, any large ritual affair concerned a whole community and involved all its members in some measure.

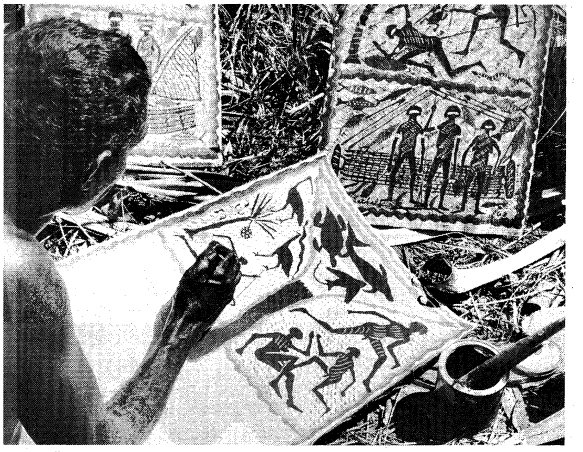
Aesthetics. Sacred ritual provided immense scope for aesthetic expression, especially in dramatic performances with stylized posturing and complicated dance movements. Less intense but sometimes almost as elaborate were the nonsacred ceremonies (or corroborees) designed for entertainment and relaxation. Songs ranged in style from the succinct verses or couplets of central Australia and the Western Desert, which were made up of three, four, or more words repeated in linked sequences, to the more elaborate songs of northeastern Arnhem Land, which were long verses building up complex word pictures through symbolic allusion and imagery. There was no poetry in terms of spoken verse, but there were chants, some of them outstandingly beautiful. The majority of secret-sacred songs comprised mythic cycles, each containing several hundreds of verses. There was also a wide repertoire of songs on everyday events, such as the "gossip" songs of western Arnhem Land, composed by songmen with the aid of spirits. Instrumental music in the north was provided by the didgeridoo (drone-trumpet) and by clapping sticks. In southern and central regions, boomerangs or clubs were rhythmically beaten together or pounded on the ground; in southeastern Australia, women used skin beating pads. Tunes and rhythms varied greatly from area to area.

Songs, chants, and instrumental music

Oral literature was rich. In addition to sacred mythology, there were ordinary stories and tales, either historically true or presumed to be true. Some existed in several versions, depending on the situation in which they were told and on the individual background of the storyteller.

Each cultural area had its own distinctive style of art. *Tjurunga* (sacred object) art, consisting of incised patterns on flat stones or wooden boards, was representative of a large area of Australia, although centralized in Aranda territory. In central Australia, body decoration and elaborate headdresses on ritual occasions, using feather down, blood, and ochres, were especially striking. Everywhere, sacred ritual provided the incentive for making a large variety of objects—mostly impermanent, because the act of making them was itself one of the appropriate rites. In western Arnhem Land, *maraiin* objects—realistic and stylized carved representations of various natural species—were made. The *rangga,* or ceremonial poles, of eastern Arnhem Land, many of durable hardwood, bore ochre designs and long pendants of feathered twine. For mortuary rituals, the Tiwi made large wooden grave posts, and shaped and decorated receptacles for bones were common in eastern Arnhem Land. Also common were carved wooden figures of mythic beings and of contemporary persons, some used in sacred ritual, others as memorial posts for the dead.

Douglass Baglin



Aboriginal artist working on a bark painting, Queensland.

Paintings in ochre on sheets of bark were indigenous to Arnhem Land, although examples could be found in the Kimberleys and in southeastern Australia. They were used mostly on the initiation ground for the instruction of novices. In western Arnhem Land, naturalistic patterns showing figures against an open background were the norm; there was also a unique kind of "X-ray" art that depicted the internal organs of animals and human beings. Also found in various parts of Australia were cave and rock paintings or engravings.

For an outstanding artist or songman, the rewards lay mainly in prestige and in religious advancement. He was not absolved from earning a living in the ordinary way and shouldering routine responsibilities.

ABORIGINES IN AUSTRALIAN SOCIETY

The heritage of early alien contact. Traditionally, the Australian Aborigines had little knowledge of other peoples and were ill-equipped to withstand the rapid inroads of an alien and culturally different people.

Early contact with Indonesians

Two qualifications are necessary here. First, the Aborigines excelled in adjusting themselves to their natural environment. Secondly, in the north they came into contact with Indonesians at an early date (at least two to four centuries ago). The visitors came as sea-going traders to the Arnhem Land coast and made a powerful impact on local art, music, ritual, and material culture. On Cape York Peninsula, the influence of New Guinea and the Torres Strait Islands was evident in the adoption of masked ritual dancing and the use of the drum.

But the coming of Europeans, in 1788, was entirely different. Initially, many Aborigines were willing to welcome them, even regarding them as returning spirits of the dead or as manifestations of the mythic beings. European settlement soon expanded, however, making inroads into tribal territory and interfering with natural resources, excluding the Aborigines from their sacred and other areas and drawing them—often forcibly, always uncomprehendingly—into the life of the developing colony. Communication was minimal. Clashes marked virtually all situations in which Aborigines and Europeans pursued conflicting interests. In the period of "pacification by force," up to the 1880s, a large number of Aborigines were killed. Others were driven into the bush or remained in small pockets subject to the "civilizing" influence of missions, or were left to fend for themselves in the fringe settlements of cities and towns; still others remained in camps or pastoral and cattle stations to become the nucleus of a labour force. Diseases took their toll. The Colonial Office in London had instructed that the Aborigines' rights should be safeguarded and that they were to be accorded the benefits of British subjects. The actual situation was different.

The Aborigines reacted with spasmodic guerrilla warfare and stock killing or with passive resistance. But they soon learned that the only alternative was to adapt, at least to some extent. The result was pauperization. Gradually, missionaries and government welfare agents began to have some effect, and questions of humane treatment came to have a more practical meaning. But in outlying areas, maltreatment and violence lingered on into the early 1940s. Further, wherever European settlement was intensive, miscegenation took place, and part-Aborigines came to replace the full-blooded population. Their traditional life ceased to exist as a living reality over much of the southwestern, southeastern, and middle eastern areas of the continent. In the central and northern regions traditional life remained, even on some pastoral, mission, and government stations, although in a modified form. In more remote areas it was still possible for Aborigines to live approximately in the way they had before but with notable modifications, particularly in the sphere of law and order. It was for some time believed that the Aborigines would eventually die out, and reserves were established in the late 1920s and early 1930s to serve as a buffer between them and Europeans. But many were attracted to the fringe settlements, where they formed tribally and linguistically mixed communities. This meant the emergence of a new form of living, structurally linked to the wider Australian society.

Developments since World War II. From the early 1940s, government policies have been changing. But growing support for the goal of assimilation (in effect, Europeanization) did not include adequate programs for achieving it. Many Aborigines were resistant, skeptical, and disillusioned—the aftermath of neglect and the not-too-distant memory of the traumatic events of early contact. In the early 1950s, some sought withdrawal and a magico-religious escape in the *kurangara* (Kimberleys) or in a revival of quasi-Aboriginal religion (as on the north coast of New South Wales). The Pindan Cooperative at Port Hedland in Western Australia is a further example of a movement concerned with maintaining group cohesion on a quasi-traditional basis.

No Aborigines exist who have not had some contact with modern Australian society. By 1960 only about 7,500 Aborigines still kept even a modified traditional orientation. The great majority of the total population of some 111,000 persons (1981) who identified themselves as Aborigines or Torres Strait Islanders (Aborigines of Papuan ancestry) were living in southern cities and country towns—more European both in physical appearance and in manner of living. Generally, within recent years, the emphasis has been on drawing Aborigines more closely into the wider Australian society and, in the process, erasing unfavourable forms of discrimination. All Aborigines are now Australian citizens, eligible to vote, to receive social service benefits, and to drink liquor. Facilities for primary and secondary education and for technical training have been improved, and more people of Aboriginal descent are taking advantage of this. Nevertheless, many remain poorly trained and educated, caught in the lowest socioeconomic level of Australian society. Prospects are brighter for the younger generation, but social acceptance by many white Australians is still limited.

Population figures

Two major developments are significant. A Commonwealth (Federal) conference in 1965 nominally redefined assimilation policies by giving Aborigines a choice between an Aboriginal and an Australian–European orientation. The trend toward social and cultural alteration has actually proceeded so far that, in most areas, such choice is unreal. But it has left the door open in some areas where Aborigines still have an informed appreciation of the value of their traditions. The other development has been an increasing uniformity between state and Commonwealth policies, partly a result of the Commonwealth Department of Aboriginal Affairs, which serves in an overview capacity and provides economic aid. In 1973 the National Aboriginal Consultative Committee (NACC), composed of Aborigines and Torres Strait Islanders elected throughout Australia, was established to advise the minister of Aboriginal affairs on Aboriginal needs, desires, and policies. The NACC was replaced by the National Aboriginal Conference (NAC) in 1977. The NAC was composed of elected Aborigines and Torres Strait Islanders from each state and territory, and it met annually on a national basis and quarterly on a territorial or state basis.

Articulate part-Aboriginal groups have emerged in the south; they insist on integration rather than assimilation—that is, on retaining Aboriginal identity as a unique status symbol marking them off from other Australians. This movement toward pan-Aboriginality has implications for all people of Aboriginal descent. In the north, the focus has been on questions of land ownership and control, including compensation (and not just royalties) for and a share in the mineral exploitation that is occurring on Aboriginal reserves. The Aboriginal Land Rights Act (1976) gave traditional Aborigines the right to hold reserve land in the Northern Territory. By the late 20th century more than 139,000 square miles of the Northern Territory had reverted to Aboriginal ownership. Minerals on the Aboriginal lands, however, remained the property of the federal government, but royalties from mineral exploitation are paid to the Aboriginal communities holding title to the land. The act also provides the framework for Aborigines to lay claim to other lands if a strong traditional link can be proved. Aborigines are also assisted by the Aboriginal Development Commission in the purchase of non-reserve land for settlement and business. Despite these accomplishments in Aboriginal representation and land ownership, there is still wide dissatisfaction with the progress of socioeconomic development. A likely consequence of this is a reinforcement of the movements mentioned above toward maintaining what remains of the traditional life, reflecting a belief that the last significant aspects of the Aborigines' social and cultural identification should not be dissipated. (R.M.B.)

Australia to 1900

EARLY EXPLORATION AND COLONIZATION

Early contacts and approaches. Prior to documented history, there may have been further Asian contacts. Chinese astronomers are said to have made observations in Australia during the 6th century BC; firmer evidence argues for a Chinese landing near Darwin in 1432. The incursion of Islām into Southeast Asia came within 300 miles (480 kilometres) of Australia, and adventure, wind, or current might have carried some individuals the extra distance. Both Arab and Chinese documents tell of a southern land, but with such inaccuracy that they scarcely clarify the argument. Similarly, the "jave la Grande" shown on some 16th-century European maps may or may not be a genuine representation of Australia. Bugis (Macassar) seamen certainly fished off Arnhem Land, in the Northern Territory, from the late 18th century and may have done so for generations. Perhaps only effective resistance by the Aborigines against the Bugis reserved Australia for white colonization.

The Portuguese. The quest for wealth and knowledge might logically have pulled the Portuguese to Australian shores; and the assumption has some evidential support, including a reference that Melville Island, off the northern coast, supplied slaves. Certainly the Portuguese debated the issue of a *terra australis incognita* ("unknown southern land")—an issue in European thought in ancient times and revived from the 12th century onward. Yet hard, clinching evidence of contact is lacking.

The Spanish. Viceroys of Spain's American empire regularly sought new lands. One such expedition, from Peru in 1567, commanded by Álvaro de Mendaña, discovered the Solomon Islands; excited by finding gold, Mendaña hoped that he had found the great southern land and that Spain would colonize there. In 1595 Mendaña sailed again but failed to rediscover the Solomons. One of his officers—Pedro Fernández de Quirós, a man of the Counter-Reformation, who desired that Catholicism should prevail in the southland, of whose existence he was certain—won the backing of King Philip III for an expedition under his own command. It left Callao, Peru, in December 1605 and reached the New Hebrides. Quirós named the island group Austrialia del Espiritu Santo, and he celebrated with elaborate ritual. He (and some later Catholic historians) saw this as the discovery of the southern land. But Quirós' exultation was brief; troubles forced his return to Hispanic America. The other ship of the expedition, under Luis de Torres, went on to sail through Torres Strait but almost certainly failed to sight Australia; and all Quirós fervour failed to persuade Spanish officialdom to mount another expedition.

Oceanic exploration. The exploration and settlement of Australia began early in the 17th century.

The Dutch. Late in 1605 Willem Jansz. of Amsterdam sailed from Bantam in search of New Guinea. He reached Torres Strait a few weeks before Torres himself and unknowingly saw, and also named, part of the Australian coast—Cape Keer Weer, on the west of Cape York Peninsula. More significantly, from 1611 some Dutch ships sailing from the Cape of Good Hope to Java inevitably carried too far east and touched Australia: the first and most famous was Dirck Hartog's "Eendracht," from which men landed and left a memorial at Shark Bay, Western Australia, October 25–27, 1616. Pieter Nuyts explored almost 1,000 miles of the southern coast in 1626–27, and other Dutchmen added to knowledge of the north and west.

The first landing

Most important of all was the work of Abel Tasman, who won such respect as a seaman in the Dutch East Indies that in 1642 Governor General Anthony van Diemen of the Indies commissioned him to explore southward. In November, having made a great circuit of the seas, Tasman sighted the west coast of what he called Van Diemen's Land (latterly Tasmania). He then explored New Zealand before returning to Batavia. A second expedition of 1644 contributed to knowledge of Australia's northern coast; thenceforth, New Holland was the name for the landmass.

The British. But the Netherlands spent little more effort in exploration, and the other great Protestant power in Europe, England, took over the role. In 1688 the English pirate William Dampier relaxed on New Holland's northeastern coast. On returning to England, he published his *Voyages* and persuaded the Admiralty to back another venture. He traversed the western coast for 1,000 miles (1699–1700) and reported more fully than anyone previously, but in terms so critical of the land and its people that another hiatus resulted.

The middle decades of the 18th century saw much writing about the curiosities and possible commercial value of the southern seas and *terra australis incognita.* This was not restricted to Great Britain, but it had especial vigour there. The British government showed its interest by backing several voyages. Hopes flourished for a mighty empire of commerce in the eastern seas.

This was the background for the three voyages of Capt. James Cook on behalf of the British Admiralty. The first, that of the "Endeavour," left England in August 1768 and had its climax April 20, 1770, when Lieut. Isaac Hicks sighted southeastern Australia. Cook landed several times, most notably at Botany Bay, and at Possession Island in the north, where on August 23 he claimed the land, naming it New South Wales. Cook's later voyages (1772–75, 1776–79) added only a little to Australian exploration but were both symptom and cause of strengthening British interest in the eastern seas.

Captain Cook's voyages

Later explorations. Cook's voyages led to settlement but did not complete exploration of the Australian coasts. Marion Dufresne of France skirted Tasmania in 1772, seeing more, at least of the western coast, than had Tasman. The Comte de La Pérouse, another French explorer, made no actual discoveries in Australia, but his visit to Botany Bay early in 1788 was notable. In 1791 the British navigator George Vancouver traversed and described the southern shores discovered by Pieter Nuyts years before; A.-R.-J. de Bruni, chevalier d'Entrecasteaux, of France also did significant work, especially in southern Tasmania.

Two Britons—George Bass, a naval surgeon, and Matthew Flinders, a naval officer—were the most famous postsettlement explorers. Together they entered some harbours on the coast near Botany Bay in 1795 and 1796; and Bass ventured farther south in 1797–98, pushing around Cape Everard to Westen Port. Flinders, too, was in that region early in 1798, charting the Furneaux Islands. Late that year Flinders and Bass in the "Norfolk" circumnavigated Tasmania, establishing that it was an island and making further discoveries. Several other navigators, including merchantmen, filled out knowledge of the Bass Strait area; most notable was the discovery of Port Phillip in 1802.

Meanwhile, Flinders had returned home and in 1801 was appointed to command an expedition that would virtually complete the charting of Australia. Over the next three years Flinders proved equal to this task. Above all, he left no doubt that the Australian continent was a single landmass. Appropriately, Flinders urged that the name Australia replace New Holland, and this change received official backing from 1817.

France sponsored an expedition, similar in intent to Flinders', at the same time. Under Nicolas Baudin, it gave French names to many features (including "Terre Napoléon" for the southern coast) and gathered much information but did little completely new. It was on the northern coast, from Arnhem Land to Cape York Peninsula, that more work was needed. Two Admiralty expeditions—under P.P. King (1817–22) and J.C. Wickham (1838–39)—filled this gap.

European settlement. The British government determined on settling New South Wales in 1786, and colonization began early in 1788. The motives for this move have become a matter of some controversy. The traditional view is that Britain thereby sought to relieve the pressure upon its prisons, a pressure intensified by the loss of its American colonies, which hitherto had accepted felons. Convicts went to the settlement from the outset, and official statements put this first among the colony's intended purposes. But some historians argue that this glossed a scheme, likely to provoke concern both within Britain and at the diplomatic level, to provide a bastion for British trade in the eastern seas. Supporters of the commercial-strategic view emphasize that Cook had extended hopes that the South Pacific would provide essential naval stores, especially mast timber and flax. (Its lack of documentary support notwithstanding, the argument makes considerable sense.)

British motives for settlement

Whatever the deeper motivation, plans went ahead, with Lord Sydney (Thomas Townshend), secretary of state for home affairs, as the guiding authority. Arthur Phillip served as commander of the expedition; he was to take possession of the whole territory from Cape York to Tasmania, westward as far as 135° and eastward to include adjacent islands. Phillip's power was to be near absolute within his domain. The British government planned to develop the region's economy by employing convict labour on government farms, while former convicts would subsist on their own small plots.

The First Fleet sailed May 13, 1787, with 11 vessels, including six transports, aboard which were about 730 convicts (570 men and 160 women). More than 250 free persons accompanied the convicts, chiefly marines of var-

ious rank. The fleet reached Botany Bay on January 19–20, 1788. Crisis threatened at once. Botany Bay was poor in soil and water and even as a harbour. Phillip therefore sailed northward on January 21 and entered a superb harbour, Port Jackson, which Cook had marked but not explored. He moved the fleet there; the flag was hoisted on January 26 and the formalities of government begun on February 7. Sydney Cove, the focus of settlement, was deep within Port Jackson, on the southern side; around it was to grow the city of Sydney.

The first settlement

Phillip at once established an outstation at Norfolk Island. Its history was to be checkered—settlement was abandoned in 1813 and revived in 1825 only to provide a jail for convicts who further misbehaved in Australia. (The island served a new purpose from 1856 as a home for the descendants of the "Bounty" mutineers, who by then were too numerous for Pitcairn Island.)

Phillip remained as governor until December 1792, seeing New South Wales through its darkest days. The land was indifferent, disease and pests abounded, few convicts proved able labourers, and the Aborigines often were hostile. The nadir came in autumn 1790 as supplies shrank; the arrival of a second fleet brought hundreds of sickly convicts but also the means of survival.

An authoritarian society. While much change proceeded throughout this period, authoritarian and hierarchical elements remained strong. The reception of convicts continued and was a major fact in social and economic life. Entrepreneurs strove hard but did not yet develop a staple industry. Farmers and graziers began to fill out an arc 150–200 miles (240–320 kilometres) around Sydney: this area was designated as the Nineteen Counties in 1829, and settlement beyond that limit was discouraged. Following the discovery of Bass Strait, and in concern to secure southern waterways, new settlements were made in the south.

From Britain, David Collins sailed in 1803 to settle Port Phillip; his sojourn there was unhappy, and in mid-1804 he moved to the River Derwent in southern Tasmania, already settled (September 1803) by a group from Sydney under John Bowen. Collins resettled the amalgamated parties at Hobart. In November 1804 William Paterson founded a settlement in northern Tasmania, the precursor of Launceston. These settlements united in 1812; they were still under supervision from Sydney, although only nominally from 1825. Among penal outstations settled from Sydney were those at Newcastle (1804) and Moreton Bay (1824), the forerunner of Brisbane. Britain extended its possession over the whole of the continent in the mid-1820s, again suspecting French (or even American) intervention. The western boundary of the governor's commission shifted to 129° in 1825 to include Bathurst and Melville islands in the far north, and there was a small settlement in this region (1824–29). At Western Port, east of Port Phillip, another settlement was made (1826–27), while in January 1827 Edmund Lockyer began permanent settlement at Albany, Western Australia. His instructions stated that Britain now claimed all Australia.

British possession of the whole continent

Structure of the government. As remarked above, the constitutional structure was authoritarian. The governors were all military officers. There were no representative institutions; but the Judicature Acts of 1823 and 1828 provided for executive and legislative councils, with the major officers of government serving in both and an equal number of private individuals, chosen by nomination, in the latter. More significant at this stage was articulation of a judicial system, especially the establishment of supreme courts (New South Wales, 1814; Tasmania, 1824); normal trial by jury did not obtain.

Sociopolitical factions. Within this rigid structure, sociopolitical factions developed and affected events. Most important in the early years was the leadership of the New South Wales Corps, stationed at Sydney from 1791. Some officers of the corps sought power and profit with an avidity that led to clash after clash with the early governors; this culminated on January 26, 1808, when John Macarthur, a former officer of the corps, led a rebellion that deposed Gov. William Bligh (served 1806–08), earlier famous for the "Bounty" mutiny. In due course, the imperial government reacted and recalled the corps; but Gov. Lachlan Macquarie (served 1810–21) also clashed with the colony's "exclusives"—former officers and a handful of wealthy free immigrants. Conversely, he associated himself with the "emancipist" faction—a group that argued in favour of erstwhile convicts having a particular claim upon government and the colony's resources.

The "emancipists"

Macquarie's attitude disturbed the imperial government. After an official inquiry (1819–21) by J.T. Bigge, it encouraged the migration of men of some standing and wealth to both New South Wales and Tasmania. Such men received substantial grants of land and appeared the natural leaders of social and economic development. The emancipists continued to be strong, however, especially through the leadership of W.C. Wentworth (himself the son of a convict woman), whose newspaper, the *Australian* (founded 1824), was the spearhead of opposition, especially to Gov. Ralph Darling (served 1825–31). In Tasmania factions never formed so clearly, but there, also, the press led criticism of the government.

The convicts. By 1830 about 58,000 convicts, including almost 50,000 men, had come to Australia (the rate increasing rapidly after 1815). Many were more or less habitual urban thieves. There were a few political prisoners, while a substantial proportion of the Irish convicts (a third of the total) had become offenders through sociopolitical unrest. In Australia the convicts were either employed by government or "assigned" to private employers. Broadly speaking, conditions were not especially harsh or repressive, and "tickets of leave" and pardons provided relatively quick routes to freedom; assignment to the new settlers of the 1820s, however, often had an element of slavery. Most convicts committed some further misdeeds, although only about one-tenth were charged with serious offenses. Those found guilty went to secondary penal stations, the (sometimes exaggerated) horror spots of Australian history—Macquarie Harbour, Newcastle, and Moreton Bay in this period and, later, Norfolk Island and Port Arthur. The convicts gave Australia a large *Lumpenproletariat,* yet success stories were common enough, and many led decent lives. There were only a few militant protests—the most remarkable was an uprising among Irish convicts outside Sydney in March 1804. Altogether, the convict impact was less grim and ugly than might be expected.

Economic activity. The maintenance of convicts was essentially the economic resource of the colony for many years; this function caused very considerable expenditure by the British government. Wealth was won by supplying government stores with food and grain, or by controlling internal trade (especially in rum), which ultimately depended on the government maintaining the settlements, or by both. The officers of the New South Wales Corps were skilled in filling these roles, although civil officers, private settlers, former convicts, and even serving convicts all had their own means of doing business, and the amount of petty commercial activity was large. Farming was pursued on a widely ranging scale. John Macarthur was most notable of those who early believed that wool growing would be a major economic resource; he himself received a substantial land grant in 1805 to pursue this hope, and he persuaded Bigge of its validity. By 1830 these hopes were yet to be decisively confirmed. Sealing and whaling returned a little more, although the richest seal fields (especially in Bass Strait) were soon thinned; and not until the 1820s did colonists have the wealth to engage seriously in whaling, although British and Americans early used Australian ports for this purpose. Maritime adventure led early colonists to make contact with Pacific islands, most importantly Tahiti.

The growth of commerce

Exploration. The period saw some notable exploration by land. From early days in Sydney men sought a way over the mountains, 50–100 miles west. The task was accomplished in 1813; the journalist W.C. Wentworth led the party. A surveyor, George William Evans, followed their route to Bathurst (founded 1815) and reported rich pastoral country. John Oxley further mapped the inland plains and rivers, especially the Lachlan and Macquarie, and also explored the southern coasts of future Queensland (1823), while Allan Cunningham was the great pioneer of that state's hinterland (1827). Meanwhile, in 1824–

25, Hamilton Hume and William Hovell went overland southward to the western shore of Port Phillip. Charles Sturt, in 1828–30, won still greater fame by tracing the Murray–Murrumbidgee–Darling river system down to the Murray's mouth.

Culture. The writings of explorers and pioneers were Australia's first contributions to literary culture. While catering to the European appetite for natural history, they sometimes achieved literary grace. Pictorial illustrations of the new land, some by convicts, also dated from the earliest years. David Collins' *An Account of the English Colony in New South Wales* (1798) and W.C. Wentworth's *Description of New South Wales* (1817) were literate, informed, and impressive. Wentworth showed skill as a versifier, too, especially in his *Australasia* (1823). Newspapers were founded as early as 1803, and they contributed to cultural as well as political history. Outstanding was the architecture of Francis Greenway, a former convict, who, under Macquarie's patronage, designed churches and public buildings that remain among the most beautiful in Australia.

THE GREAT SHIFT: 1830–60

The three decades between 1830 and 1860 saw the most rapid change in Australia's history. The impact was most evident in politics, but the economy and culture were no less affected. Patterns then established persisted.

Settlement. Four of Australia's six states were formed between 1829 and 1859. A British naval captain, James Stirling, examined the Swan River in 1827 and interested English capitalist-adventurers in colonization. Two years later he returned to the Swan as governor of the new colony of Western Australia. The Colonial Office discouraged schemes for massive proprietorial grants; still the idea persisted, with Thomas Peel—kinsman of the future prime minister Sir Robert Peel—investing heavily. But colonization was grim work in a hot, dry land, with the government reluctant to spend a penny. Western Australia's story for decades was survival, not success.

Yet the same enthusiasm quickly generated around proposals to establish a colony in South Australia, inspired by a British social reformer, Edward Gibbon Wakefield, who argued that if land were sold at a "sufficient" price, its owners would be forced to maximize its value by cultivation, while labourers would have to lend their energies to that task before being able to become landowners themselves. Highly doctrinaire, Wakefieldianism appealed to the liberal intelligentsia and to dissenting groups in England who also backed nascent South Australia. The first colonists arrived in July 1837, and Adelaide was settled soon afterward. The colony experienced many hardships, but lasting significance resulted from its founders' stress on family migration, equality of creeds, and free market forces in land and labour.

The northern and southern portions of New South Wales formed separate colonies. Settlement into the Port Phillip district in the South proceeded very quickly from the mid-1830s, colonists coming from both north of the Murray and from Tasmania; Melbourne's history began in 1835 and boomed immediately. Throughout the 1840s there were calls for constitutional independence, granted in 1851, when the Port Phillip District took the name Victoria. Northward, the Moreton Bay District was never quite so buoyant; and the creation of Queensland had to wait until 1859. Short-lived settlements included Port Essington (1838–49) and Gladstone (1847).

Development of responsible self-government

Politics. All colonies except Western Australia gained responsible self-government. New South Wales led the way when an imperial act of 1842 created a two-thirds elective legislature. The Australian Colonies Government Act (1850) extended this situation to Victoria, South Australia, and Tasmania. The act made allowance for further revision of the colonial constitutions, and in 1855–56 this took effect in the four colonies, Tasmania then abandoning the name Van Diemen's Land. Queensland followed, at its separation from New South Wales. All had bicameral legislatures, with ministers responsible to the lower houses, which by 1860, except in Tasmania, were elected on a near-democratic adult-male franchise.

While the imperial power often responded to colonial cries for self-rule, there were some tense moments. Virtually all colonists abhorred paying taxes for imperial purposes, including the costs of maintaining convicts locally; a good many disliked convictism altogether; most disputed the imperial right to dictate land policy; and many, especially in South Australia, disapproved of the imperial government directing that aid be given to religious denominations. These were the main issues, coalescing with that of increased self-rule, which stoked sometimes quite violent debate.

From the outset of the period, the imperial government fostered a freer market in land and labour throughout the colonies, not merely in South Australia. Thus grants of land ceased in 1831, replaced by sale. Attempts to create a satisfactory system caused much friction, with colonists generally hostile to any demand for payment. In New South Wales in 1844, new regulations even prompted talk of rebellion.

Friction over convicts

With regard to labour, colonists agreed with imperial encouragement of free migration, but friction arose over the convicts. British opinion in the 1830s became increasingly critical of the assignment of convicts to private employers as smacking of slavery; it was abolished in 1840, and with it transportation of convicts to the mainland virtually ceased, although more than before were sent to Tasmania. The end of assignment removed the chief virtue of transportation from the colonists' viewpoint and so contributed to a very vigorous movement against its continuation. The British government terminated it for eastern Australia in 1852; in Western Australia transportation commenced in 1850, at the colonists' behest, and continued until 1868. Altogether some 151,000 convicts were sent to eastern Australia and nearly 10,000 to Western Australia.

In the early 1850s the most dramatic political problem arose from the gold rushes. Diggers (miners) resented tax imposition and the absence of fully representative institutions. Discontent reached a peak at Ballarat, Victoria, and in December 1854, at the Eureka Stockade, troops and diggers clashed, with loss of life. The episode is the most famous of the few occasions in Australia's history involving violence among Europeans.

Common suspicion of the imperial authority modified but did not obliterate internal tension among the colonists. Divisions of ideology and interest were quite strong, especially in Sydney, where a populist radicalism criticized men of wealth, notably the big landholders. The coming of self-government marked a leftward, although far from revolutionary, shift in the internal power balance.

The economy. This period saw the first and greatest booms of the two bonanzas of Australian economic growth—wool and minerals.

Wool. Only now did men, money, markets, and land availability interact to confirm that Australia was remarkably suited for growing fine wool. Occupation of Port Phillip was the most vital part of a surge that carried sheep raising 200 miles and deeper in an arc from beyond Adelaide in the south, north, and east to beyond Brisbane. The "squatter" pastoralist became an archetype of Australian history. Although pastoralism contributed to depression in the early 1840s, the industry kept growing, and the whole eastern mainland benefitted consequently.

Pastoral expansion had one grim corollary—disaster for the Aborigines. Relations between the two races were always tense. The early governors all ordered and even practiced benevolence, but to little effect. When the European spread inland, the Aborigine fought hard, but that only strengthened the European's arm. The Tasmanians moved rapidly toward extinction, but the Australians, with further room for withdrawal, evaded that fate. Philanthropists, missionaries, and administrators stood aghast but found no remedy.

Minerals. The first significant mineral discovery was that of copper in South Australia (1842 and 1845). The discovery had the effect, to be repeated time and again, of suddenly redeeming an Australian region from stagnation. Much more remarkable, however, were a publicized series of gold discoveries made from 1851 on, first in east central New South Wales; this ignited the gold years and

spread into and throughout Victoria. As a result, Australia changed from a land of exile to one of golden attraction. The Victorian economy benefitted from the flood of men and money, although the smaller colonies suffered. Despite the Eureka Stockade incident, the diggers proved quite moderate and responsible.

Culture. Governments and citizens paid considerable heed to improvement of soul and mind. From the mid-1830s, generous aid helped all Christian churches to expand. The Church of England had the highest nominal allegiance, but in the eastern mainland colonies Roman Catholicism was notably strong; Methodism had vigorous advocates throughout; Congregationalism and other forms of dissent dominated in South Australia; and Presbyterianism had its chief strength in Victoria. Most churches attended to education, especially the provision of superior schools, while the state struggled to provide a primary system. The universities of Sydney and Melbourne were founded in 1850 and 1853, respectively. Mechanics' institutes, museums, and botanical gardens were also built.

Architects created much beauty in early Australia. Artists and writers were active; drama and music developed in all towns. The first Australian novel, *Quintus Servinton* (1830–31), was by a convict, Henry Savery; Henry Kingsley's *Geoffrey Hamlyn* (1859) is often judged the first major Australian novel. John West's *History of Tasmania* (1852) was a work of remarkable scope and insight.

Further exploration

Various forms of science had their investigators, but land exploration remained the richest field of discovery. Sir Thomas Livingstone Mitchell confirmed Sturt's work on the river systems and first opened the way from New South Wales to the rich lands of western Victoria (1836). The West Australian coastal regions were mapped by George Grey (1837–40) and by Edward John Eyre, who went overland from Adelaide to Albany (1840). Eyre and Sturt both vainly attempted to reach midcontinent from Adelaide; this was at last achieved in April 1860 by John McDouall Stuart, who in 1862 went still farther, to Darwin. Meanwhile, the central north and the northeast had been penetrated from Sydney; the most famous explorer was Ludwig Leichhardt, who led two successful expeditions (1844, 1846–47) before disappearing in an attempted traverse from the Darling Downs to Perth. An equal and more celebrated tragedy ended the expedition of Robert O'Hara Burke and William John Wills, who crossed from Melbourne to the Gulf of Carpentaria in 1860–61 but starved to death on the return. Later explorations of Western Australia in the 1870s added the names of John Forrest and Ernest Giles to the pantheon of explorer-heroes.

SEVERAL SMALL DEMOCRACIES: 1860–1900

During this period the colonies had little formal relation with each other; instead they concentrated their attention inward on their capitals. New Zealand had more in common with the eastern colonies than did Western Australia. Federation came about in 1901, but the more striking fact was its tardiness. Nevertheless, the colonies did follow similar, if independent paths.

Politics. Democracy was largely fulfilled, save that the upper houses remained elitist in franchise and membership. Governments changed rapidly over long periods, but the constitutions survived. Political groupings were extremely intricate, often personal or power seeking in origin, but allowing some expression for liberal or conservative ideology.

Social reform

The liberals made the colonies quite advanced in matters of social reform, if not the average man's paradise that some glib publicists pictured. Breaking up the large "squatter" estates and replacing them with yeoman farming was a constant concern, meeting many difficulties yet achieving some effect where market and environment allowed. Reformers put much faith in education and strove toward providing adequate primary schooling for all. "Free, secular, and compulsory" was a slogan and roughly the final result; this entailed savage controversy with the Roman Catholic Church, which scorned the godless schools and made enormous efforts to provide its own. Other forms of state aid to religion tapered away. Factory legislation and rudimentary social services presaged the welfare state; restriction of nonwhite, especially Chinese, immigration belonged to this context, for Europeans feared these labourers would reduce living standards, but the restriction was also a matter of sheer racism.

The economy. Overall the economy prospered. Wool and metals continued as the great export income earners. Pastoralism flourished, especially up to the mid-1870s; despite land legislation, this was the heyday of the squatter "aristocracy." Expansion of sheep and cattle growing into the more distant hinterland continued the heroic-pioneer theme of earlier years. Railway construction aided rural industry and proceeded remarkably fast, notably in the 1880s: between 1875 and 1891 the mileage rose from 1,600 to above 10,000 and reached as far as 500 miles inland. Most of the required capital was raised overseas on behalf of governments, contributing to the extremely important role played by the public sector in economic growth.

Gold and copper discoveries

Mining. Victoria's gold and South Australia's copper maintained their significance as new techniques allowed more sophisticated exploitation. Gold was found in southern Queensland in the later 1860s, then in the Northern Territory, and in tropical Queensland: the Palmer River goldfield pulled men to the far north in the mid-1870s. By then Cobar, in central New South Wales, had proved the most important of many new copper fields. Tin also became significant, Mt. Bischoff in Tasmania being the world's largest lode at its discovery in 1871. The 1880s was predominantly the decade of silver; western New South Wales proved richest, and in 1883 Charles Rasp, a German migrant, first glimpsed the varied riches of Broken Hill, which were to make that city almost fabulous and to prompt the establishment of Broken Hill Proprietary Company Ltd.—in time, Australia's largest private enterprise. Also from 1883 dated another big and ramifying discovery, the gold of Mount Morgan, Queensland. Gold also became Western Australia's great bonanza in the early 1890s, the Kalgoorlie and Coolgardie fields winning world attention; the copper of Mt. Lyell, Tasmania, was another highlight of that decade. These discoveries were both product and instigator of much wider activity, creating speculation, mobility, boom, and slump of extraordinary impact.

Industry. Urban expansion and the growth of secondary industry, while less distinctively Australian and contributing little to export income, were remarkable. By the criteria of investment, employment, and relative acceleration, the growth of secondary industry outstripped that of primary industry. Secondary industry multiplied its growth some ten times over during the period, so that manufacturing and construction accounted for 25 percent of the national product in the 1880s. The population ratio shifted decisively from country to town, establishing an extreme capital-city concentration and eventually putting Melbourne and Sydney among the world's large cities. Urban building and services attracted much capital; and most manufacturing was directed to providing food, furniture, and clothing for the relatively affluent townsman. City speculation contributed more than its share to a general overcapitalization, which, in company with worldwide forces, produced a depression in the early 1890s, the main impact of which was in the urban-secondary sector.

The colonies. The history of the respective colonies sharpens some points in this general background. For further treatment of the history of an individual state, see below under state name. In the later 19th century, regional characteristics consolidated, and they changed little at least until the 1960s.

Victoria. Victoria retained the impetus of the 1850s for a full generation. This was most evident in its capital, Melbourne, which had a vigorous cultural and social life. Ardent and ideological liberalism was evident in the colony's education controversy and, with greater novelty, in its adoption of protection as a means of developing its industries and living standards. Disputes between the upper (conservative) and lower (radical) houses of the parliament were frequent and sharpened political feeling. A famous clash occurred in 1865–66 over a protective tariff; another, in 1877–78, over payment of members—the liberal cause triumphed on both occasions.

Adoption of protection

New South Wales. With its longer background, New South Wales changed less in this period. Its master politician, Henry Parkes, first came into prominence in the 1840s. Parkes was involved in sectarian disputes, which were especially vigorous in the colony. Another major theme of political debate was protection versus free trade—the latter retaining greater favour, in contrast to Victoria. Sydney had its share of scandals and scalawags, especially late in the period, contributing to a rumbustious image.

Queensland. Physical expansion westward and northward dominated the history of Queensland. Cattle and sugar became industries of substantial importance. A class of small farmers was established that aspired to settle the tropics, which were usually considered unsuitable for small-scale farming by Europeans. Conversely, the established "kings" of the region used Kanakas (labourers from the Pacific islands). Thus the continued immigration of such labour provoked hot debate, which was not resolved until after federation, when the young commonwealth imposed an absolute prohibition. The Kanaka question contributed to regional feeling antagonistic to the capital.

South Australia. South Australia enjoyed less prosperity than its eastern neighbours. Agriculture remained significant in its economy but saw much disappointment; in the decade around 1870 farmers pushed out into semiarid country, hopeful that rain would follow the plough, only to learn with cruel certainty that it did not. Settlement drew back toward Adelaide, the sole sizable town. Landholding did prompt South Australia's most famous contribution to reform: that land transfer proceed simply by registration, rather than through cumbrous title deeds.

Woman suffrage

Another notable contribution was the institution of woman suffrage (1894), which effectively brought nationwide application of the principle at federation; appropriately, South Australia was the home of Catherine Helen Spence, the most remarkable Australian woman of the time, who published a significant novel, *Clara Morison* (1854), and became active in many social and political movements.

In 1863 South Australia took over the administration of the area thenceforth known as the Northern Territory, which earlier had remained technically part of New South Wales; the change entailed adjustment of boundaries. (The Northern Territory became the concern of the federal government in 1911.) South Australia set up its administrative centre at Darwin. In 1872 the construction of an overland telegraph line linked Australia with the outside world via Darwin.

Tasmania. The 1860s imprinted a "sleepy hollow" image on Tasmania, which persisted. The mineral discoveries at Mt. Bischoff and elsewhere were correspondingly important in reviving the economy. Nevertheless, living standards remained lower than in Victoria, and in Tasmania there were still property qualifications for voting in 1900. The colony contributed to democratic practice, however, by experimenting with proportional representation.

Western Australia. Western Australia ceased to receive convicts in 1868; it gained a partly elected legislature in 1870 and responsible government in 1890. Premier throughout the 1890s was John Forrest, as adept at politics as at exploration. Until the gold rushes, economic growth was slow and primitive; in the 1890s the colony was fastest in relative growth, and little short of that in absolute terms. Farming (in the southwest), town and railway building, and social legislation all followed.

Social movements. Working class and radical movements stretched back to the 1830s, although substantial trade union organization came only after midcentury.

Labour. The unions won some job benefits, including widespread adoption of the eight-hour day. The 1870s and 1880s saw extensive mass unionism, notably among miners and sheep shearers. Trades halls arose in the cities, and organization, extending beyond colonial boundaries, began to knit together. The unions early considered using political pressure and gaining political representation. This inclination strengthened in the early 1890s, helped by tougher times and by employers' stiffening resistance to union demands. Thus arose the labour parties, which gained quick success, especially in New South Wales and Queensland. At first the labourites' aim was simply to influence ministries, but for a few days in December 1899 Anderson Dawson was Labor premier in Queensland.

Rise of the labour parties

Other radicals reacted differently to the pressures of the 1890s. A few hundred of them set off for Paraguay in 1895 to establish there a Utopian "New Australia"; they failed. Republicanism was probably stronger in the 1890s than at any other time, sometimes accompanying a Marxist-like militance.

Movement toward federation. Federation was another ideal of the times. Most important politicians supported the cause, with more or less altruism. They could operate on more positive factors than common background and apparent common sense. Since the Crimean War (1853–56) Australians had feared incursion from the north by Europeans or Asians or both; the most emphatic result came early in 1883, when the government of Queensland, fearful of Germany, took possession of Papua, forcing Britain's reluctant connivance. Better defense was one motive for association, and so was the prospect of more effective Asian immigration restriction; intercolonial free trade was another desideratum. The Australian Natives Association (Australian-born comprised 64.5 percent of the population in 1901) rallied to the cause.

Yet the mills ground slowly. A federal council existed from 1885, but this was only a standing conference, without executive power. New South Wales never joined the council; the senior colony was jealous of a movement that would reduce its autonomy, the strength of which was in Victoria. Conventions met in 1891 and 1897–98 to prepare draft constitutions. These then went to referenda, which gained "yes" majorities. The Commonwealth of Australia came into existence on January 1, 1901.

The Commonwealth of Australia

The constitution was federal, with the states (as the colonies now became) forsaking only limited and specified power to the commonwealth government—these included defense, immigration, customs, marriage, and external affairs. While the lower house, the House of Representatives, was elected by single-member constituencies of roughly equal size, each state had an equal number of representatives in the upper house, the Senate. Ministers were to be members of Parliament. A high court would interpret the constitution.

Culture. Men of learning had contributed to the nationalist surge. Especially in the 1890s and through the Sydney *Bulletin,* verse and prose portrayed the "Outback" as the home of the true Australian—the bush worker: tough, laconic, and self-reliant, but ever ready to help his mate. The *Bulletin* was nationalist, even republican, and much more radical than the federalist politicians. Henry Lawson and Joseph Furphy were the supreme writers of the nationalist school. Painters and poets also extolled the nationalist ideal.

Not all cultural achievement belonged to the nationalist context, however. Henry Kendall was a lyricist of nature, and A.L. Gordon wrote of horses and countryside with a skill that won him a memorial in Westminster Abbey. "Rolf Boldrewood" (Thomas Alexander Browne) wrote tales of outback adventure, while the great 19th-century Australian novel was Marcus Clarke's *For the Term of His Natural Life* (1874), based upon convict records and legends. The older universities remained small, but with some outstanding men on their faculties; the universities of Adelaide (1874) and Tasmania (1890) were new foundations. Ferdinand von Mueller was an outstanding botanist who worked primarily at the Botanic Gardens, Melbourne. That city was the home of the great coloratura soprano Nellie Melba (Helen Porter Mitchell, born 1861), not quite the first but certainly the most famous Australian singer to achieve international renown.

Popular culture followed the British model, with music halls, novelettes, and especially sport to the fore. Australian rules football developed first in Melbourne and became strong throughout southern Australia; in cricket, a victory over the mother country in 1882 established one area of colonial equality. Admiration combined with fear to create a sporadic cult of the bushranger (rural desperado-thief): its most famous expression came with the capture of Ned Kelly's gang and Kelly's execution in 1880. Urban youths joined in gangs, or "pushes," and won the epithet larrikin.

Popular culture

Australia since 1900

NATIONHOOD AND WAR: 1901–45

The world's passions and conflict of the early 20th century were to shape the new nation's history, despite its physical distance from their epicentres. By many standards, this was the least attractive of the five major periods of Australian history. Nationalism strengthened, but it killed and sterilized rather more than it inspired; egalitarianism tended to foster mediocrity; dependence on external power and models prevailed. Yet creativity and progress survived.

The economy. Drabness was most evident in economic affairs. At the broadest level of generality, the period did little more than continue the themes of the 1860–90 generation. The most important such themes were the improvement of communications (railways reached their peak of 27,000 miles in 1941, and meanwhile came the motor boom) and increasing industrialization. In the primary field, there was significant expansion of exports, with wheat, fruits, meat, and sugar becoming much more important than hitherto. But just as manufactures received increasingly high tariff protection, so the marketing of these goods often depended on subsidy; and so the sheep's back continued to be the nation's great support in world finance. Metals, gold especially, were important in the early years; but thereafter this resource conspicuously failed to provide the vitality of earlier and later times. The worldwide economic depression of the 1930s affected Australia, especially its primary industries; otherwise the overall rate of growth, and probably of living standards, too, scrambled upward—a little more quickly than average in the mid-1920s and perceptibly so in the 1940s.

Politics and government. In national politics, men fought for office with increasing vigour and resource, while their administrative performances generally began well but then ebbed. Retrospect can espy a strangely regular pattern wherein four crisis episodes—the establishment of national policies, World War I, depression, World War II—alternate with three periods dominated by a political party enjoying its climacteric. The Labor Party filled this role in 1904–15; the Country Party, 1919–29; and the United Australia Party, 1931–41. At the state level politics were more confused, if not impenetrable.

Strengthening of the central government

A constant theme was the strengthening of the central government as against the states. This complemented the high degree of homogeneity, especially in personal and social matters, that extended through Australia's great physical spread; it was expressed primarily through the commonwealth's financial powers—at first especially relating to customs but later by direct taxation. From World War I both levels of government imposed such taxes, but in 1942 the federal government virtually annexed the field, with the high court's approval. The establishment of a national capital at Canberra, where Parliament first sat in 1927 after meeting in Melbourne since federation, symbolized this situation.

Culture. The period produced not only Furphy's *Such Is Life* but also the work of Henry Handel Richardson (H.H.R.; pseudonym of Ethel F.L. Richardson, later Robertson), another contender for "the great Australian novelist." In *The Fortunes of Richard Mahony* (three volumes, 1917, 1925, 1929), H.H.R. told the anguish of the central character, modelled on her father, as he sought to come to terms with Australian life. The tension of dual loyalties to Britain and Australia was a major concern also of Martin Boyd, whose long career as a novelist began in the 1920s. A more exclusively nationalist tone pervaded many tales of outback life and historical novel sagas. The first notable novel of urban life was Louis Stone's *Jonah* (1911); a later contributor to this genre was Vance Palmer (especially *The Swayne Family,* 1934), who, with his wife Nettie, won fame as a literary critic and selfless patron of the aspiring young.

The most significant contribution in poetry came from a group in Sydney influenced by the German philosopher Nietzsche and the late-19th-century French innovators. Outstanding was Christopher John Brennan, a major theorist of symbolism. While calling on the Australian background, these men gave a sophistication to their poetic world that lifted it far from outback balladry. Associated with this group was Norman Lindsay, an artist, novelist, sculptor, and seer.

In art, the rural landscape held domination. Revolutionary changes in European art were markedly slow in affecting Australia, but a few artists did produce some notable work of imaginative technique. Musical composition was hackneyed and mediocre, although in Percy Grainger Australia produced (but did not retain) a musician of remarkable originality and ability. Architecture promised an interesting chapter with the selection of the American Walter B. Griffin's design for the city of Canberra; in practice this was much mutilated, but Griffin did do some interesting work in both Melbourne and Sydney.

Education

One outstanding area to which the universities contributed was anthropology; a chief protagonist was A.R. Radcliffe-Brown (professor of anthropology at Sydney, 1925–31). Australians increasingly filled faculty posts, although most who did so had Oxbridge degrees, while some of the most able native intellects worked overseas. The University of Western Australia, founded in 1911, drew on one of the most substantial philanthropic bequests in Australian history (from the newspaperman Sir Winthrop Hackett) and initially charged no fees. Other university foundations were Queensland (1909) and colleges at Canberra and Armidale. The states developed their own secondary schools throughout the period, although the achievement was scarcely comparable to the development of primary education in the early period.

Australia was in the forefront of filmmaking early in the century, but this early promise soon faded. A.B. Paterson's "Waltzing Matilda" became Australia's best known song—part folk hymn and part national anthem. Radio had an impact in Australia equal to that elsewhere; radio stations became a mark of medium-town status, and the Australian Broadcasting Commission became a major force in culture and journalism. Radio helped make the 1930s probably the most sports-conscious decade in Australia's history. Cricket, tennis, swimming, boxing, and horse racing were areas of athletic excellence. Aviation moved from sport to enterprise to business; Charles Kingsford-Smith was the most famous hero and Qantas the most successful airline.

Growth of the commonwealth. The first two prime ministers were Edmund Barton (served 1901–03) and Alfred Deakin (served 1903–04), who had led the federation movement in New South Wales and Victoria, respectively. They were liberal protectionists. Their ministries established the "White Australia" (*i.e.,* exclusion of Asians) immigration policy, a tariff, an administrative structure, the high court, and a court of conciliation and arbitration that carried probably to the highest point anywhere in the world the principles of industrial arbitration and judicial imposition of welfare and justice through wage and working-condition awards.

Early ministries

In 1904 J.C. Watson led the first, brief Labor cabinet, followed by G.H. Reid's conservative free-trade ministry. Deakin led again (1905–08); then Andrew Fisher was Labor's second prime minister (1908–09), his ministry defeated when liberals and conservatives "fused" in Deakin's third term (1909–10). Then Labor won its first clear majority at election, which it barely lost in 1913 and regained, still under Fisher, in 1914. This kaleidoscope did not hinder—perhaps it even prompted—ambitious governmental policies. Social services were extended with old-age pensions (1908) and maternity grants (1912); protection rose markedly in a 1908 tariff; the Commonwealth Bank was established; and an army and navy developed.

The new nation was psychologically as well as physically prepared for war. Fear of attack became increasingly directed against Japan, and it prompted pressure on Great Britain for a firmer policy in the New Hebrides (since 1886 supervised jointly by Britain and France)—achieved in 1906–07. Although many Australians criticized Britain when the latter appeared negligent of local interests, the dominant note was overweening loyalty to the empire. Colonial troops had fought in both the Sudan and Boer wars. In 1914, when World War I began, politicians of all hues rallied to the imperial cause.

World War I. Some 330,000 Australians served in

World War I; 60,000 died, 165,000 suffered wounds—few nations made such relatively heavy sacrifice. The most famous engagement of the Australia and New Zealand Army Corps (ANZAC) was in the Dardanelles campaign (1915); the day of the landing at Gallipoli—April 25—became a day of national reverence, honoured far beyond any other. Even before Gallipoli, Australian troops had occupied German New Guinea, and the Australian vessel "Sydney" sank the German cruiser "Emden" near the Cocos Islands (November 9, 1914). After the Dardanelles, Australians fought primarily in France—Ypres, Amiens, and Villers Bretonneux were among the battles, all redolent with slaughter. In Palestine, the Australian light horse and cavalry corps contributed to Turkey's defeat.

Effects of war

The war profoundly affected domestic affairs. In economic development, it acted as a supertariff, benefitting especially textiles, glassmaking, vehicles, and the iron and steel industry. Such products as wool, wheat, beef, and mutton found a readier market in Britain, at inflated prices. But the shock of war affected politics much more, especially by giving full scope to the furious energy of W.M. Hughes, who supplanted Fisher as Labor prime minister in October 1915. Soon afterward he visited Britain. There his ferocity as a war leader won acclaim, and he became convinced that Australia must contribute still more. He advocated military conscription for overseas service; but a referendum in October 1916 declared negatively for this proposal, and immediately afterward the Labor parliamentary caucus moved no confidence in Hughes's leadership. But he continued as prime minister of a "national" government, even after losing a second conscription referendum in December 1917. The referenda in particular and war stress in general made these years uniquely turbulent in Australian history. The Labor Party lost other men of great ability along with Hughes. The split cemented a long-standing trend for Roman Catholics to support the party. Hughes's enemies also included the small but growing number of extremists—most notably the Sydney section of the Industrial Workers of the World (IWW)—who opposed the war on doctrinaire grounds.

The postwar years. The aftermath of war echoed, but finally resolved, this turbulence. Some radicals hoped that returning servicemen would force social change; but instead, the Returned Servicemen's League became a bastion of conservative order, its supporters ready to use physical force against local people they considered "Bolsheviks." The Labor Party faltered, its members adopting a more radical socialist type of platform in 1921, but with far from uniform conviction. When the challenge came to Hughes's leadership early in 1923, it arose partly from the conservative-business wing of Hughes's own Nationalist Party (its representative S.M. Bruce becoming prime minister) and partly from the Country Party, which from late 1922 held a crucial number of parliamentary seats. Although led by wealthy landowners, the Country Party won support from many small farmers: it benefitted too from its former-soldier image and from widespread country-versus-city feeling. Its leader, E.C.G. Page, had considerable, if erratic, force.

"Men, Money, Markets"

Bruce continued as prime minister until 1929, with Page his deputy in Nationalist–Country coalitions. Bruce declared his policy to be the discovery of "Men, Money, Markets" and worked hard toward this end. The cost was high, however: tariffs, bounties, prices, and public indebtedness all rose. There was considerable administration innovation—*e.g.,* the Loan Council regulated all governments' overseas borrowing—and the successful Council for Scientific and Industrial Research (later, the Commonwealth Scientific and Industrial Research Organisation [CSIRO]) was established in 1926 to apply scientific expertise to developmental problems. The worldwide development of consumer industry had its impact: the revolution in transportation provided by the automobile is the best example, although full-scale car production was still in the future.

With much economic activity subsidized—the exception being one primary product (wool)—Australia was particularly vulnerable to the Great Depression of the 1930s. It struck hard: unemployment exceeded 25 percent of the work force and imposed a degree of social misery unknown in Australian history. The rate of recovery was uneven, manufactures doing better than primary industry. Population growth slowed; at the nadir, emigration exceeded immigration.

Politics reflected the impact. J.H. Scullin succeeded Bruce as prime minister in 1929, and his Labor ministry suffered the real squeeze of events; within the Labor Party there was considerable division as to how government should react to the Depression. Some favoured a generally inflationist policy, with banks facilitating credit issue and governments extending public works. Right-wing Labor men distrusted such a policy; radicals would have gone further, by renouncing interest payment on overseas loans. Orthodox opinion argued for deflationary policies—curtailed government expenditure, lower wages, balancing the budget, and the honouring of interest commitments. In June 1931 the commonwealth and the state governments agreed on a plan—the Premiers' Plan. Albeit having some inflationary features, this foreshadowed a one-fifth reduction in government spending, including wages and pensions—a considerable affront to Labor's traditional attitudes.

Against this background, the government disintegrated. Before the Premiers' Plan, some right-wingers, led by J.A. Lyons, had crossed to the opposition. In November some leftist dissidents voted against Scullin, forcing his resignation. In the elections that followed, Labor suffered a heavy defeat. The new prime minister was J.A. Lyons, whose followers had coalesced with the erstwhile Nationalists to form the United Australia Party (UAP). Lyons led a wholly UAP government until 1934 and UAP–Country coalitions until his death in 1939.

The governments of J.A. Lyons

The Lyons governments provided stability and not much more. Recovery was uneven and sporadic, quicker in manufacturing than in primary industry, aided more by market forces than by governmental planning. Two policies failed badly to fulfill expectations—the Imperial Economic Conference, held at Ottawa, Ontario, in 1932, improved trade slightly, but the integrated economic community for which some had hoped never developed; and Australia's "trade diversion policy" of 1936, which tried to redress the imbalance of imports from Japan and the United States, offended those countries and actually reduced exports further. A plan for national insurance, the Lyons governments' most ambitious social legislation, also aborted. These mishaps did not much bother the electorate; improvement, even if meagre, was enough to retain favour.

Internal division was the greater threat to the government. This became manifest after Lyons' death. The UAP elected Robert Menzies as its new leader (and therefore prime minister); but the decision was hard fought, and it was criticized publicly and vehemently by Page, still leader of the Country Party. Nevertheless, Menzies retained office; but internal division persisted, the coalition's parliamentary majority was tiny, and Menzies resigned in August 1941. A.W. Fadden, the new leader of the Country Party, then took office, but in October he gave way to John Curtin and a Labor ministry.

Communist and Fascist movements

While the electorate generally voted conservative, Australia shared the common Western experience of the interwar years in the rise of a small, vigorous Communist movement. Founded in 1922, the Australian Communist Party made most headway in the big industrial unions and in Sydney; it also had some influence and supporters among the intelligentsia, especially in the 1930s. The party suffered a share of internal factionalism but for the most part was able to present a united face to the public.

Fascism achieved no formal political recognition in Australia, but there were hints of sympathy toward Fascist attitudes—D.H. Lawrence wrote of such in his novel *Kangaroo,* based on a brief visit in 1922; and an "Australia First" movement began in literary nationalism but drifted into race mystique and perhaps even treason. An intellectual movement of more lasting force developed among a group of young Roman Catholic intellectuals in Melbourne in the mid-1930s. They developed a commitment to social justice and against Communism somewhat in the manner of G.K. Chesterton. This was the "Catholic Social Movement," which had considerable influence on Australian Catholics.

Whereas Australia had been virtually spoiling for war before 1914, passivity became the international keynote after 1920. At the Paris Peace Conference that formally concluded World War I, Hughes was his fire-eating self, especially in defense of Australia's interests in the Pacific; thus he won a mandate for erstwhile German New Guinea and Nauru (an atoll in the central Pacific) and effectually opposed a Japanese motion proclaiming racial equality, which he thought might presage an attack on Australia's immigration laws. In the League of Nations, Australia was an independent member from the outset. Bruce's succession as prime minister marked a new emphasis. Very much an Anglo-Australian, Bruce led the nation into a period when "the empire" became the object of yet more weighty rhetoric and more desperate hope than earlier. Australia did not ratify the Statute of Westminster (1931, embodying the 1926 Balfour Declaration as to the constitutional equality of the Dominions) until 1942. The UAP governments followed Britain closely concerning the totalitarian expansion of the 1930s; if Australian influence counted for anything, it was to strengthen appeasement of Germany and Japan. Although fear of Japan continued, that country's accession to the Fascist camp did not provoke a tougher governmental line. The Labor Party meanwhile was even more incoherent and variable in matters of foreign policy than were its social-democratic counterparts elsewhere in the Western world: isolationism and anti-Fascism were equal and opposing forces.

World War II. When war came again, however, the nation's response was firm—some 30,000 Australians died in World War II and 65,000 were injured. From early in the war, the Royal Australian Air Force was active in the defense of Britain. The Australian Navy operated in the Mediterranean (1940–41), helping to win the Battle of Cape Matapan (March 1941). Australian troops fought in the seesaw battles of North Africa. In mid-1941 Australians suffered heavy losses both in the Allied defeats in Greece and Crete and in the victories in the Levant. Meanwhile, the German general Erwin Rommel was scoring his greatest triumphs in North Africa; out of these emerged the successful Allied defense of Tobruk, substantially by Australians (April–December 1941).

After the Japanese attacked the United States naval base at Pearl Harbor, Hawaii (December 7, 1941), however, the focus shifted homeward. The Japanese victories of the following months more than fulfilled the fantasies that fear and hate long had prompted in Australia. On February 15, 1942, 15,000 Australians became prisoners of war with Singapore's fall, and four days later war came to the nation's shores, when Darwin was bombed. Then came a Japanese swing southward, by August threatening Port Moresby, New Guinea.

Relations with the United States

The United States became Australia's major ally. In a famous statement (December 1941), Prime Minister Curtin declared: "I make it quite clear that Australia looks to America, free from any pangs about our traditional links of friendship to Britain." A sharper note of independence from Britain came when Curtin insisted (February 1942) that Australian troops recalled from the Middle East should return to Australia itself and not help in the defense of Burma, as British prime minister Winston Churchill wished. Conversely, American needs prompted total response to Curtin's call. U.S. general Douglas MacArthur established his headquarters first in Melbourne and then in Brisbane; the Australian Navy assisted in the U.S. victory in the Battle of the Coral Sea in May, which retrospectively appears a turning point in the war; and the two nations' troops thereafter fought in many joint land battles. The American soldier became a common figure in the Australian capitals, forging the biggest single link in the social relations between the two countries (although not always a harmonious one, as competition for girls and grog sparked jealousy).

On land, the fortunes of war turned against the Japanese in August–September 1942, beginning with an Allied victory at Milne Bay, New Guinea. More prolonged—and of more heroic dimension in Australian eyes—was the forcing back of the Japanese from southern New Guinea, over the Kokoda Trail. Then followed a long attrition of Japanese forces elsewhere in New Guinea and the islands, with Australia playing a role secondary to American forces but nevertheless significant—initially in New Guinea, and at war's end especially in Borneo. Australian volunteers and conscripts fought in these campaigns, the government and people having accepted the legitimacy of sending conscripts as far as the Equator and between the 110th and 159th meridians.

The war brought some passion into domestic affairs, albeit less than that of 1914–18. Curtin's government exercised considerable control over the civilian population, "industrial conscription" being scarcely an exaggerated description. Overall this was accepted—partly because of the sheer crisis, partly because the government showed purposefulness and capacity. Curtin easily won the 1943 elections; thereafter his ministry and the bureaucracy gave considerable thought to postwar reconstruction, hoping to use war-developed techniques to achieve greater social justice in peace.

Wartime growth

The war carried industrialization to a new level. The production of ammunition and other matériel (including airplanes), machine tools, and chemicals all boomed. Meanwhile, primary production lost prestige, aids, and skills, so that the 1944 output was but two-thirds that of 1939–40. Urban employment was bountiful, and concentration in capitals became more marked than ever; many families had two or more income earners. Thus affluence quickened; federal child endowment from 1940 and rationing of scarce products helped distribute this wealth. The gross national product, estimated at $1,089,000,000 in 1918–19 and $1,860,000,000 in 1938–39, rose to $2,936,000,000 in 1942–43.

The states. As noted above, the period saw a steady loss of state power to the federal centre and a general lessening of local consciousness. Nevertheless, repeated referenda showed the people reluctant to expand commonwealth powers formally. At many levels state loyalties and state governments still mattered most.

New South Wales. New South Wales resumed its primacy as the most populous, wealthy, and industrialized area. Most of the generalizations made in any account of Australia derive from the particular experience of New South Wales; the federal Labor and Country parties both drew their major strength from it. Its internal politics illustrated the common theme that Labor did much better in state than in commonwealth elections. The dominant Labor figure, J.T. Lang, as premier of New South Wales in 1925–27, emphasized child endowment and other welfare services. Again in office in 1930–32, Lang refused to endorse the play-safe policies with which the federal and other state governments responded to the Depression; feeling became high and a "New Guard" organization, led by military officers from World War I, pledged itself to save the state from Lang's radicalism. Finally, the state governor dismissed Lang, who broke also with the federal Labor Party and fought it for years afterward.

Federal parties' major strength

Victoria. Victoria offered much less drama; the dynamism of the colony's first half-century never returned. Deakin, Bruce, and Menzies were all Melbourne men, and Lyons had his power base there—indicative that this city remained the bastion of financial interests. Yet in state politics the Labor and Country parties generally supported one another and the Nationalists did poorly. The generation of electricity from brown coal deposits in eastern Victoria was especially important in maintaining economic development.

The smaller states. The smaller states probably lost economically by federation. Interstate free trade meant that there could be no local tariffs to offset Sydney's and Melbourne's advantages in industrial production. As fiscal matters generally became more complex, the states were left farther behind the commonwealth and more at its mercy. Tension developed, most famously in Western Australia, where in a referendum the people voted for secession in 1933. The federal government then established the Commonwealth Grants Commission, which thereafter administered subsidies intended to ensure approximate equality in living standards.

Each of the smaller states strove vigorously, if not very

effectually, to close the gap between itself and New South Wales–Victoria. Chief reliance fell on sugar and cattle in Queensland; timber and gold in Western Australia; apples, other fruit, and hydroelectric power in Tasmania. South Australia sank into deepening gloom until the mid-1930s; revival then came through the success of a campaign to attract industry.

AUSTRALIA SINCE 1945

Social and economic history. The rate of change in Australia after 1939 was greater than that during the preceding 80 years (in both centuries the '50s were especially dynamic). Sheer affluence was an important factor. Australians had long been among the world's wealthy, and, relatively speaking, this situation remained much as before. In absolute terms, however, the average Australian acquired appreciably more consumer goods and comforts. Large-scale *embourgeoisement* was all the more significant in a society in which working-class attitudes had always set overall norms. By 1980 about one-half of the population owned automobiles, and more than two-thirds of all families owned their domiciles. These basic possessions were the lodestones of Australian life.

In the midst of Australia's prosperity and its relative freedom from tensions and problems, there were those who bewailed the mediocrity of its ruling elites—business, academic, administrative, political. This criticism had considerable force. Wealth seemed as often to dampen as to inspire a drive for excellence. Yet there were many exceptions to mediocrity. Canberra—with its buildings and its institutions, as developed from the mid-1950s—probably had the most excellence to offer: the federal public service was efficient and skilled, while the Australian National University and the National Library of Australia reached international calibre.

The affluence of the 1960s did not reach rural industry. The economic role of both the small farmer (his life often a grim compound of debt, drought, and disaster) and the big farmer (who had been a member of a wealthy caste) seemed certain to diminish. Australian leaders had long encouraged industrialization, and the masses had preferred urban living; still, the "bush" had meant much in national mythology, and the shrinkage of its economic importance was thereby significant for all.

Immigration

Second to affluence in shaping a new Australia was the federal government's assistance of mass European-wide immigration from 1946 until restrictions were imposed in December 1974. Roughly 100,000 immigrants entered Australia each year; fewer than one-third came from Britain, about one-sixth from Italy, and almost one-tenth each from Germany and The Netherlands. Thus not only did migration ally with increasing birthrates to increase the population, but the traditional homogeneity of British background (often exaggerated in vernacular, but sizable enough in fact) ended. The inner suburbs of Melbourne and Sydney became polyglot communities. Far-reaching changes were made in Australia's immigration policies and procedures in 1978, effectively raising immigration quotas and changing the rules for selecting immigrants.

At other points, too, ties with Britain loosened. After 1941–42 the Royal Navy was no longer Australia's shield against the world. The United States and Japan became trading partners and investors of comparable weight—and the West Indies became a more stimulating opponent at cricket. Thus, no longer could an Australian speak of Britain as "home" without appearing ridiculous. The change was not absolute—Australia remained an active member of the British Commonwealth—but it was enough to leave a vacuum in Australian life.

Discoveries of minerals accelerated in the 1960s. Iron in Western Australia was the most considerable; bauxite, especially in north Queensland, and nickel, in Western Australia, ranked next. Uranium was a more exotic but transitory resource, while tin, copper, lead, and zinc all rallied. Petroleum and natural gas were discovered in substantial quantities in southern Queensland and off the coast of southeastern Victoria. Japanese capital was particularly important in the development of these resources.

Stock exchange speculation became commonplace in talk and action. The scale of business multiplied. In terms of economic activity and profitability, Australia placed high among the world's second-rank nations.

Government played a part of some significance. The Labor governments before 1949 believed in fiscal supervision as a matter of ideology, and this tendency survived the change of administration. "Protection all around" remained a basic policy, although it had some opponents, and in 1973 the Labor government reduced import tariffs by 25 percent. Tariff preferences for less developed nations were instituted in 1974. Through membership in the General Agreement on Tariffs and Trade (GATT) and other organizations, the government sought beneficial trade agreements, especially for primary resources. Marketing boards and commissions even came to include wool in 1970–71, markets for wool being scarce. Yet investors and managers, including those from overseas, went comparatively without restraint. Legislation directed against monopolies, for example, was insignificant, and overseas investors were not obliged to reinvest profits. Restrictions on foreign investment were imposed by the Labor administration early in 1973, but they were canceled in August 1974.

Protective tariff policy

Industrial relations became a matter of enormous complexity, involving executive and judicial arms of government as well as employer and employee organizations. They seemed to work, with generally less controversy than in earlier periods and in other places, although a strike in the petroleum industry in 1972 was especially disruptive. In May 1975 Australian workers began a trial period of wage indexation, imposed by the Conciliation and Arbitration Commission, to be based on the consumer price index; but indexation was abolished in 1981 following trade union protests and strikes against the rigidity of the system. The Australian Council of Trade Unions was one of the nation's powers.

Change in attitudes toward Asia

Attitudes toward Asia also changed in ways difficult to generalize. Arrogant antipathy diminished; sympathy, interest, and knowledge increased. Relations with Japan became vital to the economy, many Asian students attended Australia's universities, and revisions of immigration policy caused an increasing number of Asians to be allowed to enter. In the late 1970s Australia became one of the UN's designated countries of resettlement for refugees from Indochina, and by 1980 some 45,000 refugees had settled in Australia. Yet foreboding remained; its bases were Australia's heritage of Western history and Eastern geography and the end of European, especially Anglo-Saxon, world supremacy.

Not everyone shared in the nation's affluence. Some European immigrants worked hard at unpleasant jobs. Overall, about one-tenth of the population might be considered underprivileged, with old-age pensioners suffering most. Social services, though expanded considerably by the Labor government in 1973–75, were not adequate to meet these problems. After 1975 the national health scheme was improved, culminating in the institution of a national medical plan in 1984.

The Aborigines' experience remained the nation's atrocious scandal. A mining boom in northern Australia extended the long history of territorial expropriation, but the Aboriginal Lands Rights Act (1976) in the Northern Territory offered new hope (see above *Aboriginal Australia: Developments since World War II*). A mere handful of Aborigines achieved distinction—the best known were Lionel Rose, who became a world bantam-weight boxing champion, and Evonne Goolagong, women's tennis champion. In politics, the professions, or the academy, they had no counterparts.

State boundaries and loyalties suggested a division in internal affairs. The crises of the nation's first half-century had forced unity, or at least the acceptance of federal power. The relative calm of later years was accompanied with growing antagonism to "dictation" from Canberra. Growing wealth in Western Australia especially reduced the economic dominance of the southeast corner.

Culture. Painting, sculpture, poetry, and the novel all flourished in the postwar period. The writing of Australian history produced some notable prose. Ray Lawler might

have produced "the great Australian play" with *The Summer of the Seventeenth Doll* (1955), its heroes two Queensland sugarcane cutters. Affluence enabled architects to build more private homes of distinction. Construction of an opera house in Sydney was begun in 1957 and finally completed in 1973. Melbourne had its cultural centre that housed an art gallery of splendid quality.

ydney
)pera
louse

The Commonwealth government assisted theatrical work; national opera and ballet companies developed considerable reputations. The Australia Council was established in 1975 as a statutory authority to advise the Commonwealth on the arts and to sustain and promote the arts. In the late 1960s a small group of established writers and composers was invited to compose a new national anthem; "Advance Australia Fair" was chosen in 1974 to replace "God Save the Queen." The soprano Joan Sutherland achieved overseas fame almost comparable to that of Nellie Melba.

Universities grew enormously after World War II with the admission of former servicemen. The effort overextended resources, and not until the late 1950s did federal government expenditure redress the situation. The large cities acquired two or three universities, while in the late 1960s colleges of advanced education began to diversify higher education. Meanwhile, governments had reversed the generations-old policy of refusing aid to nonstate (*i.e.*, denominational and especially Roman Catholic) schools. Considerable sums were spent on education, but the system remained far from perfect.

Medical research advanced notably. The first Australians to win Nobel Prizes, Macfarlane Burnet (1960) and J.C. Eccles (1963), both worked in this area. Astronomy also merited particular emphasis: the Mt. Stromlo observatory was another of Canberra's world-ranking institutions, and Australian scientists were in the forefront of radio and X-ray astronomy. This was one of several fields in which the Commonwealth Scientific and Industrial Research Organisation (see above *Growth of the commonwealth: The postwar years*) contributed much, departing from narrow limits of applied research. The CSIRO and the universities dominated research, but some smaller institutions were important: Burnet directed one such, the Walter and Eliza Hall Institute of Medical Research.

cientific
esearch

Australia was one of the few countries to participate in all the modern Olympic Games; and in 1956, when Melbourne was host city, Australian swimmers dominated the scene. From time to time, golf, track and field sports (especially for women), cycling, and pugilism occupied the spotlight in Australia. European immigrants played soccer, but the old Australians still flocked to see Australian rules football, especially in Melbourne.

Television, introduced in 1956 and so able to capitalize on interest in the Olympic Games, soon became dominant in everyday entertainment. Many programs came from Britain and the United States, but the Australian contribution slowly increased. Radio diminished as television boomed but later found a continuing audience. The cinema was forced to accept a more modest role until the late 1970s, when international interest in Australian films grew. Popular live entertainment benefited much from the boom, with clubs that provided liquor, gambling (poker machines), and variety shows.

Domestic politics. J.B. Chifley succeeded John Curtin as Labor prime minister in mid-1945, Curtin having died in office. A growing efficiency and enthusiasm in pursuit of social justice led to extended social welfare, national development, the establishment of the National University, and the provision of university scholarships. Public servants provided the expertise for more sophisticated supervision of the economy. The party won the elections of 1946 fairly comfortably, although the government's majority fell.

But the Australian Labor Party soon became the victim of the Cold War. During World War II Communists had increased their influence throughout Australian society and especially within the labour movement. Many believed that the Labor Party itself was Socialist. This judgment exaggerated the truth, but Chifley's belief in controlling the economy gave it some force; when he planned to nationalize all banks (1947), the reaction was accordingly intense. Although declared unconstitutional in 1948, this banking nationalization counted heavily against the government in the electorate's eye.

Meanwhile, Communist influence was in large measure responsible for many strikes; the culmination came in the coalfields of New South Wales in 1949. These strikes also aroused public concern and hostility. The Labor government suffered heavy defeat in the election of December 1949. The defeat was intensified by a restructured electoral system: thenceforth the House of Representatives had about 120 members, representing single-member electorates of approximately equal size; the Senate had 60 members, each state electing five members every three years by proportional representation, the members serving for six-year terms.

Liberal and Country party coalition governments ruled continuously after the 1949 election until 1972. The Liberal Party, formed in the mid-1940s, was largely based on the erstwhile United Australia Party. Its founding genius was Robert Menzies, and as its head he was prime minister from 1949 until 1966.

Menzies and the Liberal Party

Reverberations of the Cold War helped Menzies stay in office. In 1954 the defection from the Soviet Embassy in Canberra of the Soviet agent Vladimir Petrov led to a royal commission into alleged espionage within Australia. This not only revitalized the Communist threat, which Menzies often invoked, but prompted the Labor opposition to defend some of those implicated in the commission's inquiries. This in turn antagonized the right-wing and Roman Catholic elements in the Labor Party—the groups that had been the most vigorous elements within the party in opposing Communists in the trade unions. In 1955 this group established the Democratic Labor Party (DLP); although it won few parliamentary seats, the DLP took enough votes from the Labor Party to lessen the latter's chance of federal office.

The DLP contributed what little ideology there was in Australian politics after 1950. The Labor Party increasingly disowned any Socialist belief but found no alternative doctrine; two successive party leaders, Herbert Vere Evatt, from 1951 to 1960, and Arthur Aloysius Calwell, from 1960 to 1967, were frustrated leaders of the opposition. The Communist Party split between small pro-Soviet and pro-Chinese factions.

In the late 1960s politics took a different turn. Harold Holt, Menzies' successor as prime minister, had scarcely established himself when he died by accidental drowning (December 1967). There followed a Byzantine contest for his office, won by John Gorton. But Gorton failed to provide solid, consistent leadership, and early in 1971 the Liberal Party replaced him with William McMahon. In December 1972 a new Labor government under Gough Whitlam took power and entered upon a program of social, economic, and political reform.

Labor Party in power again

The Labor program included a large increase in appropriations for social welfare, including a national health service, urban development, a doubling of aid to education, and the removal of investment incentives for the mining industry. Opposition to Whitlam's program was effective in the Senate as early as the following May in its defeat of his electoral reform bill; in December 1973 the nation defeated a referendum that would have permitted federal price and wage controls. Increased spending meant increased inflation, and this became a campaign issue in April 1974 when Whitlam dissolved Parliament in the hope of winning a majority in the Senate. He succeeded in eliminating the DLP from the Senate and increasing his numbers so that he could pass his program through a joint sitting of Parliament. A sudden increase in unemployment in July 1974, a 16 percent loss in state elections in Queensland in December, and a financial scandal in the Cabinet in early 1975 plagued the Labor government.

New leadership came to the opposition in March 1975 with the election of Malcolm Fraser to head the Liberal Party. By October 1975 Labor strength in the Senate had been reduced, two states having filled vacancies with members from the opposition parties (an unconventional practice). Fraser took advantage of the situation by refusing to pass necessary appropriations bills, thereby forcing

another election. Whitlam refused to submit to this pressure, so the governor general, Sir John Kerr, originally nominated to the post by Whitlam, dismissed the prime minister—the first time a representative of the crown had dismissed a prime minister in 200 years. With assurances from Fraser that the appropriations bills would be passed, he was made caretaker prime minister, and a new election for both houses was set for December 13. A bitter campaign ensued, with inflation and unemployment as the primary issues. The results were sweeping majorities for the Liberal–Country party coalition. With the loss of more than 30 seats, Labor was reduced to a weak opposition, and the equilibrium of the Menzies era of the 1950s and early 1960s was returned.

Liberal–Country party coalition

Fraser was successful in reducing inflation and foreign indebtedness, but the unemployment rate remained high. Despite internal dissension, which had resulted in the formation of a splinter party—the Australian Democrats led by Donald Chipp—the coalition dominated the 1977 elections. Problems continued, however, as rifts grew deeper, and controversies arose involving individual party members. In 1983, leadership changes strengthened the Labor opposition so that it gained the majority in elections and returned to power with Robert J. Hawke as prime minister. Hawke led the Labor Party to victory in the general elections of 1984, 1987, and 1990.

Foreign affairs. The Labor governments of the 1940s, and especially that of Labor minister H.V. Evatt, hoped that Australia would play an independent, substantial part in creating a stable international situation. Thus it was active in support of the United Nations, Evatt himself serving as president of the General Assembly in 1948–49. Support of Indonesian independence was another important policy, and there was movement toward building a civil service expert in international affairs.

These positive trends continued with the Liberal victory of 1949. The next year, for example, Australia signed the Colombo Plan, under which substantial aid, especially through education, was sent to Asian countries. Australia remained an active supporter of the United Nations. Menzies, especially, was a great admirer of the United Kingdom and its empire/commonwealth.

Meanwhile, the movement of world politics soured the optimism of the 1940s. Turbulence in Asia, especially a Communist victory in China, gave a new edge to Australia's long-standing fear of attack from the north. In response the country moved increasingly close to the United States, which alone offered the possibility of replacing the United Kingdom as a protector. Thus Australians fought along with UN troops in Korea; in 1951 the country accepted the U.S. view of Japan as a bulwark against Communism, while the ANZUS (Australia, New Zealand, United States) Pact promised U.S. help in case of attack; and in 1954 Australia joined the Southeast Asia Treaty Organization (SEATO), which extended U.S. responsibility throughout the area. The United States built missile bases throughout Australia, and Australia's increasingly large defense expenditure bought much U.S. equipment. Australia also supported the United States in Vietnam; it sent troops (including conscripts) to fight there, but these troops were withdrawn between August 1971 and February 1972.

With the return of Labor to power in December 1972, a greater emphasis was put on relations with Asia. The People's Republic of China was recognized immediately, and North Vietnam was recognized in February 1973. Whitlam visited Indonesia and India and stated that the Indian Ocean "should be free from international tensions, Great Power rivalry, and military escalation." A visit to China followed in November. Closer relations with the new Labor government of New Zealand began at the outset, and ties to the United Kingdom were lessened, including reduced military support in Singapore and Malaysia. In 1974 Whitlam toured six nations in Southeast Asia, reemphasizing his government's interest in the Eastern Hemisphere. At the end of the year, Whitlam visited members of the European Economic Community, declaring that Australian prime ministers had neglected this important area, second only to Japan as Australia's trading partner. Hawke continued Fraser's style of personal diplomacy, improving relations with Indonesia and Vietnam.

Asian foreign policy

Papua New Guinea, which was composed of the Australian external territory of Papua and the Australian-administered UN Trust Territory of New Guinea, became a self-governing state in 1973. After there had been several postponements, it achieved complete independence in 1975. After independence Papua New Guinea continued to receive substantial amounts of Australian aid.

Foreign policy after 1975 shifted its emphasis from defense and security to economic matters, particularly the relationship between trade and foreign policy. (MI.R./Ed.)

For later developments in the history of Australia, see the *Britannica Book of the Year* section in the BRITANNICA WORLD DATA ANNUAL.

For coverage of related topics in the *Macropædia* and *Micropædia,* see the *Propædia,* sections 967 and 977, and the *Index.*

AUSTRALIAN CAPITAL TERRITORY

A separate political entity of the Commonwealth of Australia, the Australian Capital Territory is situated within the confines of the state of New South Wales and comprises the national capital, Canberra, and some surrounding land. The area of the Capital Territory was ceded to the commonwealth in 1911 by the state of New South Wales. Four years later the additional area of Jervis Bay was also ceded to the Capital Territory.

Like certain other federal cities, Canberra was deliberately chosen and planned as a capital, on a site formerly occupied by a sparse population of sheep raisers, along the banks of the Molonglo River. The area was first entered by Europeans in 1820. A small permanent settlement was established, probably in 1824, when stockmen squatted at Canberry (or Canbury), as it was called.

Physical and human geography

THE LAND

The 925 square miles (2,395 square kilometres) of the Capital Territory are situated on the open and gently undulating Limestone Plains, explored and named by Charles Throsby in 1820. The Murrumbidgee River passes through the territory from the southeast to the northwest along with its tributaries the Cotter and the Naas. Another tributary of the Murrumbidgee, the Molonglo, runs through Canberra. Beyond the Murrumbidgee, to the west and south, the terrain becomes much more rugged, rising to the Brindabella Range, the northernmost extension of the Snowy Mountains. Along this range and within the Capital Territory are some of the highest peaks in Australia, such as Mounts Bimberi (6,274 feet [1,912 metres]), Morgan (6,144 feet [1,873 metres]), Murray (6,040 feet [1,841 metres]), Tidbinbilla (5,114 feet [1,559 metres]), Franklin (5,400 feet [1,646 metres]), McKeahnie (4,915 feet [1,498 metres]), and Coree (4,657 feet [1,419 metres]). During the winter months, the higher parts of the mountains are covered with snow. The mountains are densely forested with eucalyptus and pine trees. The mean height above sea level in the region is 1,900 feet (580 metres), while the lowest point is about 1,500 feet (460 metres).

The climate of the Capital Territory produces temperatures colder than those of most of the state capitals of Australia. The annual mean temperature is 55° F (13° C). Monthly figures for average temperatures provide a better picture of the climate, ranging from 82.4° F (28° C) in January to 42.8° F (6° C) in July. Minimum average temperatures for the same months are about 55° F (13° C) for January and 32° F (0° C) for July. Frost is common on most winter nights, but the days are usually sunny and often warm. During the summer, temperatures are high in the daytime but not extreme, while the nights are

The climate

invariably cool. In the autumn, conditions are pleasant with long periods of sunshine and moderate temperatures. The average annual rainfall for the Capital Territory is about 25 inches (628 millimetres). In the mountains of the territory the average rainfall rises to about 60 inches (1,514 millimetres) each year.

THE PEOPLE

The early population

In 1841 the county of Murray, which includes present-day Canberra, had a population of 2,111 (1,562 males and 549 females). Almost one-third of the population were convicts, of whom 666 were male and 24 female. When the commonwealth assumed control of the area Canberra had a population of 1,777, many of whom were temporary residents involved in the construction of the Royal Military College. By 1926, when Canberra became the official federal capital, the number of persons had grown to 6,550. The vast majority of the Capital Territory's present population resides in Canberra proper. Since World War II the population has increased because of expanded government activities. Government employees account for more than half of the territory's work force. Other major population groups of the work force are construction workers, teachers, and persons involved in retail and wholesale trade.

THE ECONOMY

Much of the Capital Territory's economy revolves around government employment and services. A large portion of this government employment involves the Department of Housing and Construction. This governmental agency provides construction funding for homes, education facilities, roads, public transport, health and welfare facilities, and more, creating many jobs in the construction industry.

The forest authority manages some 150,000 acres of woodlands, including about 32,000 acres of land at the Capital Territory's two coniferous plantations. There is very little commercial production of hardwood from the woodlands of the territory. Several hundred thousands of dollars worth of softwood, Monterrey, Slash, and yellow pine, are produced each year. Timber is processed at the Capital Territory plantations, producing structural plywood; sawn, dressed, and kiln-dried lumber; and wood chips.

The agricultural and pastoral industries adopted by the region's early settlers are still a prominent part of the economy. Wheat is grown in the territory. Livestock, primarily sheep, cattle, and some pigs, is raised in the region, providing wool, milk, and meat for secondary industries.

Tourism has become a major contributor to the Capital Territory's economic welfare. The most popular attractions are the Australian War Memorial, the Black Mountain Telecommunications Tower, the High Court of Australia, Parliament House, the National Library, the Royal Australian Mint, and the mountain lookouts and reserves.

ADMINISTRATIVE AND SOCIAL CONDITIONS

The Department of the Capital Territory is primarily responsible for administering the Capital Territory. The functions administered by the department include housing, public transport, and municipal services. Education, public health, and justice are under the jurisdiction of the department of education, the department of health, and the attorney-general, respectively. The House of Assembly (formerly the Legislative Assembly) is the legislative body of the Capital Territory. The Assembly is composed of 18 members (nine from each of the electoral districts of Canberra and Fraser) serving on a part-time basis. The procedures of the Assembly are similar to rules and regulations used by the federal House of Representatives. Its role is to advise the minister of the Capital Territory on matters affecting the region and to act on legislation and ordinances concerning the territory.

The National Capital Development Commission

The National Capital Development Commission Act of 1957 created a commission by the same name to undertake and carry out the planning, development, and construction of the city of Canberra. Construction work has been carried on in conjunction with the Department of Housing and Construction and private consultants.

As Commonwealth or Crown territory, the Capital Territory lands cannot be sold for residential or commercial enterprises. Lands are leased for private development. In the late 20th century, one-twentieth of the Crown land was reserved for public parks and recreational reserves while about one-third was being leased for private purposes (excluding mining and forestry).

The city of Canberra is known as a garden city because of its many parks and recreational facilities. They are administered by the City Parks Administration of the Department of the Capital Territory, which also administers the National Botanic Gardens and the Horticultural Research Centre, both located in Weston.

Other municipal services, such as water supply, sewage collection and disposal, and storm drainage, are operated by the Department of Housing and Construction for the entire Capital Territory.

The Education Ordinance of 1937 calls for the compulsory school attendance of children between the ages of six and 15 years. The Schools Authority of the Capital Territory has the responsibility of administering all government schools. The Authority represents teachers, parents, and the community. Within the structure of the Authority, the Schools Accrediting Agency gives accreditation to courses and administers procedures regarding student assessment. There is a preschool system available to children less than six years of age. There are several specialized schools and learning centres that are available to children who have physical disabilities and learning handicaps. These specialized centres prepare such children in order that they may be integrated into neighbourhood schools.

Higher education

The Office of Further Education, which is a division of the Department of Education, is responsible for the Technical and Further Education Colleges. Institutions of higher education include Wooden College (1980), Canberra School of Art, and Canberra School of Music. In addition, Canberra College of Advanced Education and the Australian National University offer programs in undergraduate and postgraduate education.

Apprenticeship programs are conducted in many trades in the Capital Territory. Apprenticeship training is supplemented by classroom training in the trades at many colleges.

The Office of Further Education is also responsible for adult immigrant education in the territory. Free classes in English are available on a full-time or part-time basis for those who have become permanent Australian residents. Those attending classes on a full-time basis are eligible to receive a living allowance.

Transportation

The railway system of New South Wales connects Canberra to Queanbeyan. The rail line was opened for freight traffic in May 1914 and for passenger travel in October 1923. Service is provided to Sydney and Melbourne by direct and linking lines. Daily domestic airline service is available to Sydney, Melbourne, and other provincial centres. Bus services connect the territory to cities and towns in South Australia, Victoria, and New South Wales. The Capital Territory is served by an inner-urban and outer-urban system of highways and roads, connecting Canberra with suburban developments and the recreation and tourist attractions within the territory. The Internal Omnibus Network is a public bus service throughout the Capital Territory.

CULTURAL LIFE

As the federal capital, Canberra and the Capital Territory offer many cultural amenities not found elsewhere in the country. Both the Australian National Library and the National University are located in Canberra. Old Canberra retains much of the Australian town charm associated with less urban settings, while new Canberra, with its multistoried buildings and avenues, adds a touch of dignity and sophistication. The architecture ranges from the Mediterranean-style of old Canberra to the Neoclassical style of the National Library. Belts of parkland separate Canberra from its suburban satellites. Commuting distances, in and out of the capital city, are short. The country and mountains, covered with snow in the winter, are close to the city, and the seaside is about two hours away by automobile.

History

Charles Throsby, a naval surgeon who arrived in Australia in 1802, was probably the first white man to explore the lands of the present-day Capital Territory. In 1820 Throsby and his overseer Joseph Wild explored the region in search of the Murrumbidgee River. During the search, Throsby and Wild camped on the banks of the Molonglo, near the site of the Royal Military College at Duntroon. Wild decided to abandon the search for the Murrumbidgee, but Throsby continued on and discovered the Murrumbidgee in March 1821. Wild returned to the region in May 1823, discovering the Tuggeranong and Monaro plains and camping near Tharwa.

The early settlers of the Capital Territory were pastoralists, raising sheep and cattle on the plains. Jousha John Moore established the first settlement along the bank of the Molonglo at a site called Canberra or Canbury (a derivation of an Aboriginal term meaning "meeting place"). The homestead established by Moore survived until 1941. Other early settlers were Robert Campbell, who established settlement at Duntroon, and John MacPherson, who settled in the area near Black Mountain.

During the late 19th and early 20th centuries, life on the Limestone Plains was rough, crude, and often lawless. Outlaws and bands of convicts plagued the settlers of the region. By 1841 signs of permanent residence—a post office, a resident doctor, and a police magistrate—were in evidence, and the site of the township was gazetted.

The establishment of Canberra as the federal capital began several years before the government acquisition of land in 1911. As the demand for federation grew, the need for a federal capital became clear. The commonwealth of Australia was inaugurated in 1901 and the Commonwealth Constitution Act of 1900 provided for the establishment of a capital in the state of New South Wales, but not within 100 miles of Sydney. After two royal commissions and much public debate, the commonwealth parliament in 1908 selected the Yass-Canberra district and in the following year the site for the new city was determined. On January 1, 1911, an area of approximately 900 square miles was transferred from New South Wales to the Commonwealth thereby creating the Australian Capital Territory. A worldwide competition for a design for the capital was launched in the same year. The first prize was awarded to Walter B. Griffin (1876–1937) of Chicago. Construction was begun in 1913 but was delayed by World War I; from 1915 to 1920 Griffin took charge of the city's construction.

Early construction of the capital

After the war, work was accelerated by the Federal Capital Commission, which assumed control over all construction and administrative activities in the territory on January 1, 1925. On May 9, 1927, the Duke of York (later King George VI) opened the new parliament house and the transfer began of the central staffs of the administrative departments from Melbourne (hitherto the temporary seat of government of the commonwealth). Attractive houses, schools, and other buildings for this garden city were built in the suburbs. In 1930 the Canberra University college was opened and in 1936 a wing of the National Library.

The early administration of the Capital Territory was under the control of the minister of interior. The early law of the territory was the law of New South Wales, according to the Seat of Government (Administration) Act of 1910. Most of the state laws were soon replaced by territorial ordinances. The Capital Territory was administered by numerous different departments from 1910 to 1925. The Federal Capital Commission was the administering body from 1925 to 1930.

From its inception until 1930, the Capital Territory had no form of local or municipal government. One commissioner of the Federal Capital Commission was elected by the people in 1929, but this came to an end the following year when the commission was abolished.

In May 1930 the Australian Capital Territory Advisory Council was created to act as adviser to the ministers of the federal government on matters concerning the territory. The council was composed of three members elected by the people of the territory for three-year terms and four members selected by the federal government, one representing the departments of health and works, and two representing the Department of Interior. In 1952 the membership of the council was altered so that five members were elected by the people of the territory and four members selected by the federal government. The advisory council had no executive or legislative functions.

Post-World War II construction

The Depression and World War II curtailed construction, but after World War II a great surge in the development of Canberra took place. More than 8,000 houses and flats were built, as well as administrative offices, extensive buildings for the Australian National University (founded in 1946), shops, schools, and community facilities.

In 1958 the National Capital Development Commission (NCDC) was created. The NCDC is considered the creator of modern Canberra just as Griffin is credited as the creator of old Canberra. Since the early 1960s, the NCDC has been forced to deal with the problems of rapid urban and suburban growth of the Capital Territory.

In 1974 the advisory council was replaced by the Australian Capital Territory Legislative Assembly. The assembly consisted of 18 members (nine from each of the electoral districts in Canberra and Fraser) who served on a part-time basis. The first assembly was elected on September 28, 1974, and the first meeting of the assembly was held on October 28, 1975. In 1979 the name of the Legislative Assembly was changed to the House of Assembly. (Ed.)

NEW SOUTH WALES

The first British colony in Australia, New South Wales is the oldest, richest, and most industrialized state of the Australian Commonwealth. Originally, the name New South Wales was applied to the entire east coast of Australia when the British explorer Capt. James Cook claimed the territory for the British crown in 1770. The separate colonies of Tasmania, South Australia, Victoria, and Queensland were proclaimed in the 19th century, and in 1911 and 1915 the Australian Capital Territory around Canberra and Jervis Bay was ceded to the commonwealth. New South Wales was thus gradually reduced to its present area of 309,433 square miles (801,428 square kilometres). It is bounded by the Pacific Ocean to the east, the states of Victoria to the south, South Australia to the west, and Queensland to the north. By no means the largest Australian state, it is the most populous. Its capital, Sydney, is the nation's largest city. Lord Howe Island, a dependency of New South Wales, is administered as part of the state.

New South Wales reflects the dynamism and the growth problems of a fast-developing country. It is Australia's focal point of commercial farming, industry, and cultural development. It is also plagued by an imbalance between its urban and rural populations and often chafes under the financial restrictions of the commonwealth government.

Physical and human geography

THE LAND

The Great Dividing Range

Relief. The dominant geographical feature is the Great Dividing Range, which separates the narrow coastal strip from the great plains to the west. The coastal strip varies from 10 to 50 miles in width and is bounded along its entire length by a natural barrier of steep mountains. Beyond this barrier, however, the Great Dividing Range consists not of true mountains but of a series of high plateaus, or Tablelands, which slope gently to the west until they merge imperceptibly into the plains beyond. The average height of these Tablelands is about 2,500 feet. In some areas they rise to 5,000 feet, and they attain a height of 7,370 feet (2,228 metres) at Mt. Kosciusko—the highest mountain in Australia—in the Australian Alps in the south. The gentle western side of the Tablelands

is known as the Western Slopes. The plains cover nearly two-thirds of the state. Lying below 1,000 feet, they are interrupted only by the elevated country between Orange and Cobar in the east and by the Main Barrier and Grey ranges in the west.

Drainage. The Great Dividing Range is the state's main watershed. Numerous rivers flow eastward from the range to the Pacific Ocean. Though often beautiful, they are too short and rapid to be of much economic value. The major rivers that flow west—the Namoi, the Macquarie, the Lachlan, and the Murrumbidgee—cross some 500 miles of sunburned plains before joining the Murray and Darling rivers. These brown, muddy, inland rivers meander across the plains and lose a great deal of their water by evaporation. Over 1,600 miles of the Darling River, which rises in Queensland, flow to the southwest across the plains to join the Murray. Rising in Victoria, the Murray runs for more than 1,200 miles along the southern border of New South Wales before crossing South Australia to reach the Pacific Ocean.

Soils. A great deal of New South Wales is naturally fertile, and the red and black soil plains are extremely rich. The coastal strip, however, consists mostly of poor and sandy soils. Agricultural potential is severely limited by inadequate and uncertain rainfall and intense evaporation.

Climate. About 19 percent of the state receives less than 10 inches of rain a year, and approximately 23 percent receives only 10 to 15 inches. The coastal districts have the most annual rainfall, varying from about 30 inches in the south to about 75 inches in the north. Precipitation diminishes progressively away from the coast. The average annual rainfall in the northwest is only eight inches, and some of the land beyond the Darling River merges into desert.

ainfall

If the dry climate and brilliant sunshine present severe agricultural problems, they are yet attractive to the inhabitants of the coastal cities. Newcastle, Sydney, and Wollongong have a delightful climate. It is rarely too hot in summer or too cold in winter, and Sydney is without sunshine for an average of only 23 days a year. Inland it is both hotter in summer and colder in winter. Average temperatures range from 74° to 83° F (24° to 29° C) in the summer months and from 49° to 54° F (7° to 13° C) in winter. Temperatures of over 100° F (38° C) are frequent in summer, and frost at night is common in winter. Heavy snow falls on the southern mountains and, more rarely, on the northern and central Tablelands. On Mt. Kosciusko snowdrifts linger throughout the summer.

The seasons are fairly well defined; autumn begins in March, winter in June, spring in September, and summer in December.

Plant and animal life. The natural vegetation ranges from the dense semitropical forest of the northern coast to the sparse plant life of the western plains. Nearly one-tenth of the state is forested, and, except for the plains, a much larger area is covered with bush and scrub. The forests, concentrated mainly on the coast and Tablelands, give way to shrub eucalyptus on the Western Slopes and to salt bush and spinifex (a spiny grass) in the far west. The predominant tree is the eucalyptus, which is the state's chief source of hardwood. Smaller quantities of softwoods, such as the red cedar, the hoop pine, and the rosewood, are found on the northern coast, while the cypress pine grows on the Western Slopes. Native grasses are found everywhere but in the extreme west, and there are many wildflowers.

The rich birdlife includes many species of parrot and cockatoo, the flightless emu, the mound-building scrub-bird, and the mallee bird. Most of the species of marsupials, mammals that do not develop a placenta and that carry their young in an external pouch, are represented. These include the wombat, the koala, the common and ring-tailed opossum, the common and long-nosed bandicoot, and a variety of kangaroos and wallabies. Many of the smaller marsupials, their populations diminished by agriculture and forestry, are retreating into uninhabited regions. The platypus and spiny anteater are common.

Marsupials

Several species of poisonous snakes, including black, brown, and tiger snakes, are found throughout the state. The best known fish is the Murray cod, which is found in the western rivers and is an excellent food source. Trout have been introduced into the streams of the Great Dividing Range.

Settlement patterns. There are four regions of traditional activity. The coastal strip is mostly used for dairy farming, the Tablelands for sheep and cattle raising, and the Western Slopes for wheat cultivation. The plains are the site of the great sheep stations, for more than 100 years the basis of the state's economy. They still provide almost 40 percent of Australia's wool.

There are also four distinct political regions. The northern Tableland is known as New England. Its population, almost entirely rural, tends to resent the government's preoccupation with Sydney. There has long been an unsuccessful movement to form a new state. Similar feelings can be found in the Riverina district, located between the Murray and Murrumbidgee rivers. There, too, people tend to favour separatism; in practical matters, they look toward Melbourne, the capital of Victoria, rather than to Sydney. The Monaro sheep country comprises the windy, upland district of the southern Tableland, including what is now the Australian Capital Territory. The sheep graziers of the Western plains, living a remote and lonely life, feel a common loyalty born of common interests.

Rural settlement. Climatic conditions have also dictated the size of landholdings. The smallest are in the coastal strip, where dairy and fruit farming are the chief occupations. Holdings are considerably larger—between 500 and 5,000 acres—on the Tablelands and Western Slopes, where mixed farming is the normal practice. Further west they grow larger still; in the dry lands of the Western Division, many sheep stations cover over 100,000 acres. The exception to this general rule is in irrigated areas along the Murrumbidgee and other inland rivers, where diversified cultivation makes small holdings possible.

Land-holdings

The tenure of landholdings in New South Wales is mostly either freehold or leasehold from the crown. Tenancy, as understood in Europe, is uncommon, and, except in the Western Division, most land is occupied by the owners.

Because of the large size of the average field, or paddock, and the relatively uncultivated appearance of the land, a typical sheep station presents a characteristic appearance. The heart of each property is the homestead, with its cluster of low buildings and well-watered trees and gardens, surrounded by bare, brown paddocks. There are no villages, and the nearest country town may be 30 or 40 miles away. In spite of the severe problems faced by sheep graziers because of the unstable price of wool and the constant threat of drought, life on the larger properties is still considered a pleasant and privileged one.

Urban settlement. The largest industrial urban area is the coastal complex of Newcastle, Sydney, and Wollongong. There are several towns of 30,000 or fewer inhabitants in the interior, such as Albury and Wagga Wagga in the southeast and Orange and Tamworth in the east central region. These are essentially country towns serving the surrounding rural population. The only interior industrial town is Broken Hill, in the far west, which depends upon the rich mineral deposits in the Barrier Range.

THE PEOPLE

Population groups. *Ethnic origins.* The people of New South Wales are in no way different from those of Australia as a whole. More than three-quarters are of British origin, and more than one-fifth are of Irish descent. The small remainder is composed mainly of continental Europeans. There are also some 30,000 native tribesmen, or Aborigines, and a few thousand Chinese.

Religious groups. Almost the entire population professing a religion is Christian. The largest denomination is the Anglican Church of Australia, followed by the Roman Catholic Church. Other groups are Presbyterians, Methodists, Greek Orthodox, and Baptists. Lutherans and Congregationalists each comprise less than 1 percent of the population.

Demography. The birth rate and death rate and other vital statistics do not vary substantially from those of the rest of Australia. The most striking feature of the popu-

lation is the disparity between those living in the rural areas and those in the cities. Three-fourths of the state's people are crowded into 2 percent of its area in Newcastle, Sydney, and Wollongong. The state government has long been concerned by this trend, and various schemes have been proposed to correct it.

THE ECONOMY

Economically the most important state in Australia, New South Wales contains about one-third of the country's sheep, one-fifth of its cattle, and one-third of its pigs. It produces a large share of the nation's grain, dairy products, and wool and mines about half of its black coal and most of its silver, lead, and zinc. It is also the country's most industrialized area and produces over two-fifths of the nation's manufactured goods.

Resources. *Biological resources.* There are several million sheep, a few hundred thousand horses, and a few million cattle grazing on the state's vast grasslands. There are also a few hundred thousand pigs. The state forests provide an important natural resource of hardwood timber, although the percentage of forestland is low by international standards. The Pacific Ocean provides valuable fish.

Mineral resources. The most important mineral resource is the vast black-coal deposits of the central coastal region. The main silver, lead, and zinc deposits are located at Broken Hill. Large copper deposits have been discovered at Cobar. Other mineral resources include tin from the Ardlethan and Tallebong fields and rutile (a red or black mineral cut into gems), found in the coastal sands.

Power resources. Coal is the most developed power resource, though the Snowy River offers important hydroelectric potential.

Wheat and wool

Agriculture, forestry, and fishing. Agriculture is spread throughout the state. About three-fifths of the acreage under crops is devoted to wheat, which is grown for both domestic consumption and export. Other grains grown include oats, maize (corn), and rice; fodder, potatoes, grapes, sugarcane, and fruit are also raised. Excellent wine is produced in the Hunter Valley and cotton is grown on the Namoi River. Sheep are raised mainly for their wool, which is also exported. Both slaughter and dairy cattle are important.

New South Wales is the most important timber-producing state, accounting for about one-half of Australia's production. Hardwoods and softwoods are exploited, and there is a regular pine reforestation program.

Tuna fishing is the most important marine industry, and mullet, shark, and Australian salmon are also caught in significant numbers. The state provides about one-third of the national fish catch, as well as all of the oysters consumed domestically.

Mining and quarrying. The most important coalfields are in Hunter Valley above Newcastle, around Wollongong, and at Lithgow. Silver, lead, and zinc, as noted above, are mined at Broken Hill. Copper is mined chiefly near Cobar, and tin is mined in the central and northern Tablelands.

Industry. About two-thirds of the manufacturing industries are located in Sydney. Other important industrial centres are Newcastle, Wollongong, Lithgow, and Broken Hill. The fastest growing industry is the manufacture of iron and steel and metal goods. Other important products are textiles, food, beverages, tobacco, chemicals, paints, paper, and printed material. Factories are mainly small, and only a few employ over 1,000 persons. A very few large concerns, such as the Broken Hill Proprietary Company Ltd. (iron and steel) and the Australian Consolidated Industries Ltd. (glass), employ most of the industrial workers.

Power. Electricity is generated and distributed by the Electricity Commission of New South Wales. Power is sold to local authorities, the state railways and tramways, and large industrial users. Almost all the state's electricity is generated by thermal-power stations. The Snowy Mountains hydroelectric scheme, completed in 1974, has diverted the waters of the Snowy and other rivers westward into the Murrumbidgee.

Finance. Since 1942 the commonwealth has collected all income taxes and reimbursed the states on an agreed formula. As a result, New South Wales is almost entirely dependent upon commonwealth grants for its revenue. In 1971 the states were given the right to levy a payroll tax.

Administration of the economy. Like the rest of Australia, the state has essentially a capitalist economy, with the great majority of industry owned by public companies. The state government, however, owns and manages the railways, some coal mines, and the production of electricity. There is a vigorous trade-union movement. The Chamber of Manufacturers and other employers' organizations represent the interests of private enterprise.

Transportation. New South Wales has excellent internal air services. The state railways provide an adequate link between Sydney and larger population centres. They have suffered, however, from competition with air and road transport. Sydney has Australia's only underground rail network, and electric rail services connect the city with many of its suburbs.

Metro services in Sydney

Roads. There are hundreds of thousands of miles of public roads, including tens of thousands of miles of state highways and main roads. Most of these are paved, but many, including the main highways to Brisbane and Melbourne, are too narrow for the traffic they now bear. Most country roads are surprisingly good, though not all are yet paved.

Inland waterways. There are no commercial waterways, although before the railways were built the Murray and Darling rivers and their tributaries were used extensively for carrying freight.

Ports. The four major ports are Sydney, Newcastle, Port Kembla, and Botany Bay. Together they handle several million tons of cargo each year. All ports are administered by the Maritime Services Board of New South Wales.

Railways. The railways cover several thousand miles and centre upon Sydney. They also provide adequate facilities to the other major urban centres including the major mining centre at Broken Hill, 700 miles to the northwest. The lines run from north to south along the coast and roughly from east to west in the Tablelands and plains.

Air services. Air services, largely for passenger traffic, are provided by a number of lines. Air links are operated to all the major towns, and the airport at Sydney can accommodate jet aircraft.

ADMINISTRATIVE AND SOCIAL CONDITIONS

Government. *Constitutional framework.* The state government administers internal matters, while the national government (commonwealth) is responsible for defense, foreign policy, immigration, trade, customs and excise, post and telegraph services, and air and sea transport. Within those limitations, the state is sovereign and has powers to make laws for the welfare and good government of New South Wales. It has no armed forces other than the police.

Houses of Parliament

The Parliament consists of two houses—the lower house, or Legislative Assembly, directly elected by the people, and the upper house, or Legislative Council, one-fourth of whose members are elected by both houses every three years. The Cabinet is chosen from the party that commands a majority in the Legislative Assembly; it is headed by a premier. Parliament meets for three years but can be dissolved earlier.

The state also has its own governor, who is appointed by the Queen on the recommendation of the government. The titular head of the government, he is now always an Australian, and his duties are almost entirely formal.

Local government. New South Wales is divided into more than 200 local-government areas, which are controlled by councils. These councils are elected every three years by adult residents of their areas.

Elections. All elections are conducted on the basis of universal adult suffrage, and the voting age is 18 years. Voting for the Legislative Assembly is compulsory, but voting for local councils is not.

Political parties. Political parties are usually state branches of the federal political parties and tend to have the same policies and interests. The three principal parties are the Liberal Party, the Country Party, and the Austra-

lian Labor Party. There is also a branch of the Australian Democratic Labor Party, without a seat in either house, and a small Communist Party, which still retains some influence in a few trade unions. Elections are fought on state issues, but the fortunes of the parties are often affected by the fortunes of the same parties in the commonwealth.

Justice. State law and its administration are largely based upon the British system. Legal procedure includes trial by jury in criminal cases, the right of appeal, and an independent judiciary. The highest state court is the Supreme Court, against which appeals can be made to the High Court of Australia and, in certain cases, to the Privy Council in London. Minor offenses are dealt with by magistrates in the Courts of Petty Sessions, while more serious cases are brought before a judge and jury in the Courts of Quarter Sessions.

Education. Education is compulsory for all children between the ages of six and 15. Most children are educated in free, nondenominational primary and secondary schools. A minority go to private schools, most of them administered by the Roman Catholic Church. There are also a few denominational secondary schools for the children of the wealthy. There are five universities in the state; the cost of administration is shared with the commonwealth.

Health and welfare. The state government is responsible for the supervision of public health, hospitals, and medical services. There are more than 280 public and a large number of private hospitals. The national health-insurance scheme is financed and administered by the commonwealth, as are the social services such as family allowances, child endowment, unemployment benefits, and pensions.

tandard
f living

Social conditions. New South Wales shares in the high standard of living and generous social services enjoyed by all Australians. The state has its own Industrial Commission to settle industrial disputes, though here, too, the Commonwealth Arbitration system has become predominant. The basic wage—adopted from time to time by the Commonwealth Conciliation and Arbitration Commission—is adopted for state awards and agreements. The 40-hour workweek is now standard.

CULTURAL LIFE

The state cannot claim a unique culture that distinguishes it from the rest of Australia. It is perhaps the most representative and typical of all the states, and its greater population and wealth give it a leading position over all but Victoria.

Sydney has the oldest and reputedly the best of the symphony orchestras. There are many small theatres, but no professional drama company of first-class quality. Sydney's Opera House, which opened in 1973, is a major arts centre with a concert hall, a large theatre for opera and ballet, and a small theatre.

Many of Australia's leading poets and novelists were either born in the state or live there. Sydney is the centre of a school of contemporary painting that looks toward the United States and Europe for its inspiration. The Art Gallery of New South Wales has the best collection of Australian painting in Australia. In general, the cultural climate is marked more by a healthy and democratic vitality rather than by any special distinction. This is equally true of the popular arts, where there are large audiences for and keen interest in both Australian and foreign folk singers and jazz and rock groups. (J.D.Pr.)

History

In 1768 James Cook began a voyage that took him and his crew around Cape Horn, to Tahiti and New Zealand, reaching the east coast of Australia (present-day New South Wales) in 1770. For the first time European navigators had seen a part of Australia that appeared suitable for settlement. Reaching the northern point of Queensland, after a hazardous journey, Cook took formal possession of eastern Australia on August 23, 1770. Cooks's second voyage in 1772 and his third voyage in 1776 added a great deal to European knowledge of the Pacific and showed that successful voyages could be made from Europe.

SETTLEMENT

The first suggestion for the establishment of a British colony in Australia was made before a committee of the House of Commons in 1779 by Joseph Banks, who with Cook had seen Australia for himself. Transportation overseas of persons convicted of offenses against the law in England had been practised during the 17th and 18th centuries, but the Revolutionary War then being fought in the American colonies prevented further transportation there of convicted persons. The eventual victory of the American colonies, the continued practice of judges of sentencing convicted persons to transportation, and the marked increase in offenses caused by the social upheavals of enclosure of land, obsolescence in several lines of manufacture and trade, and the sudden growth of towns in England made it certain that a new outlet for offenders had to be found. Banks believed that Australia was ideally suited for this purpose. It was so distant from any other civilized place that escape would be difficult; there was little chance of opposition from the natives; the climate was mild; the soil was capable of sustaining a large number of people; there were no beasts of prey; and there was an abundance of timber and fuel.

In addition to wanting a place to which convicts could be sent, the government thought it would have to provide a home for American loyalists in difficulties resulting from their enemy's victory. But when the Pitt government came to office (1784) it was concerned only with the practical problem of getting rid of convicts. To solve the problem an order in council dated December 6, 1786, appointed "the eastern coast of New South Wales or some of the islands adjacent" as the place to which offenders might be transported. Capt. Arthur Phillip of the Royal Navy was commissioned as the first governor. At this date Great Britain, under the belief that New South Wales might be separated from what the Dutch called New Holland, did not lay claim to all Australia, but just the eastern coast and all islands adjacent in the Pacific. The governor was to take the usual oaths of office; he was given power to pardon and reprieve offenders, to levy forces for the defense of the colony, to withstand pirates and rebels, to enforce discipline, to make land grants, and to exercise general jurisdiction.

The penal colony

The powers and duties of the governor made it clear that the British government intended the colony to be controlled as a jail and meant the governor, through his staff and "police," to employ directly most of the convicts and to keep their produce as a government stock. The possibilities of private land ownership and employment of convicts and others were always visible, but it was clear that the British government considered that for many years the government in the colony would perform most of the administrative and directive work. It was clear also that the British government intended to devote the very minimum in money and materials to the settlement. Between his appointment and the departure of the first fleet for Botany Bay nine months later on May 13, 1787, Phillip had persisted in efforts to obtain more stores and equipment for his expedition, so much so that it has been concluded that by the time it sailed the idea of founding a commercial settlement was very definitely afoot. It was, however, never much more than "in the air." The first fleet was very poorly equipped, and the convicts and their guards arrived with very little with which to contend with the forests of New South Wales. Whatever had been the intention of the government in 1787, the outbreak of war with France in 1793 made it certain that little time and money would be devoted to a convict settlement on the other side of the world.

THE EARLY YEARS

Soon after arrival Phillip decided that Botany Bay was unsuitable and moved his 11 vessels to Port Jackson. This harbour, which he described as the "finest in the world," became the scene of his settlement and of Australia's greatest city. There reached Sydney 736 convicts, of whom 188 were women. Guarded by 191 marines, under 19 officers, they reached a country accurately described as "forest-clad, unkempt, uncanny and unknown," and one

Early settlement of Sydney

on which Phillip reported: "No country offers less assistance to first settlers." Hardly had the fleet arrived when the marine guards, claiming that their duties did not extend beyond the voyage out, refused to act any longer. A special unit, the New South Wales Corps, was enlisted in England to act as guards, or a police force, in the colony.

In the meantime Phillip had to do the best he could but found his community of convicts was as well behaved as most of the towns in England from which they had come. Expecting supplies and assistance, he was soon dismayed to see the second fleet—three ships—loaded not with provisions but with more convicts. It left England with 1,017 aboard; 267 died on the voyage and 486 of those who did arrive were sick. Of these, 50 died in the first month, and those "not classed as sick were hardly able to attend themselves." It was described as a "ghastly company of sick and dying. . . . Great numbers were slung over the ship's side in the same manner as they would sling a cask." However, Phillip showed an energetic devotion to the growth of the settlement, which was in part shared by many of these poor, cruelly treated people, and he felt very soon able to report a belief that "This country will prove the most valuable acquisition Great Britain ever made." But other officers of the colony were of a different opinion, seeing at first no future in the colony for themselves and then no future in it for anyone but themselves. Some obstructed the governor in his work, as well as encouraging the British government to do nothing to change a policy of neglect.

During Phillip's time there was therefore revealed a political separation of interests—on the one side capable, energetic free men with privilege and money anxious to obtain control of convict labour on slave conditions and of all commercial activities; and on the other a government with a considerable measure of responsibility to see to the employment and rehabilitation of the convicts.

Houses and public buildings were erected of timber and, when bricklayers were discovered among the convicts, of bricks. Phillip led exploration of the colony, and settlements were established at Parramatta, on the Hawkesbury River, and at Norfolk Island. Almost immediately grants of land were made to men who had arrived free, and convict labour was provided for its working. By 1791 more land was in use under government direction than private, but by that year 150 persons were in possession of farms. Cultivation had been attempted on all farms, but crops were poor, for hardly any rain fell in the year before Phillip's departure from the colony. By the middle of 1790 only 38 convicts had been assigned to private employment, and care was being taken to see they were properly employed and treated. By 1792 the settlement was producing its requirements in several types of manufactured goods.

Land grants

Failing health compelled Phillip to relinquish his command, but by his departure in 1792 the settlement had progressed sufficiently far to allow it to be said that by that date the foundation of Australia had been achieved. Phillip was able to report: "The colony is approaching to that State in which I have so long and anxiously wished to see it." Further, he felt confident that "Time will remove all difficulties." While the achievement had been remarkable, there were great difficulties ahead, but never was there any doubt of importance that the settlement would survive.

EXPLORATION, 1796–1845

Australia was explored both by land and by sea. George Bass and Matthew Flinders in 1796 explored much of the coasts of New South Wales in small whaleboats, and in 1798 Bass made a long and successful voyage in his small whaleboat around the east and south coasts of the colony, proving the existence of a strait between the mainland and Tasmania.

Until 1813 nothing was known of New South Wales beyond a coastal strip about 150 miles long and no more than 50 miles wide. The colony was shut in much more by the Blue Mountains on the west than by the sea on the east. In 1813 Gregory Blaxland, William Lawson, and W.C. Wentworth penetrated through this canyoned plateau, almost as deep as it is high, on a journey that marked the beginning of the problems of its more extensive exploration. Soon it would be no longer any easier to confine the graziers to the boundaries of settlement than "to confine the Arabs of the desert within a circle drawn on the sands." From the bases established beyond the mountains a few small expeditions went farther and discovered rivers flowing west and southwest. The nature of the country farther to the west was not, however, to be revealed for a good many years.

In 1824 Hamilton Hume and William Hovell began what was to prove the most important journey since the Blue Mountains were first crossed. They travelled southwest to reach first the Murrumbidgee River, then Australia's greatest river—the Murray, as it was later named—and then to cross the full width of what was to become the state of Victoria. This expedition revealed not only great possibilities for the extension of settlements in New South Wales but even greater possibilities farther to the south.

Exploration of the Murray Basin

In 1831 Charles Sturt travelled down the Murrumbidgee River and into the Murray. This journey of nearly 2,000 miles proved the Murray to be the centre of a northern network of rivers, but still there were rivers farther to the north that flowed out to the west; perhaps they flowed toward some inland sea, perhaps they came to an end in dry, shingly beds.

THE NEW SOUTH WALES CORPS

When Governor Phillip departed in 1792, the senior officers of the New South Wales Corps took over. Probably no less than 16,000 acres of land were granted between Phillip's departure and the arrival of Gov. John Hunter in 1795. The small group of corps officers and their friends had a great advantage in trade—few others received income in currency acceptable by those who shipped goods to the colony. Also, they had powers to apply or not to apply the law, even if not the power to say what the law was. Acquisition of land, of advantages in trade, and of favour at the hands of "the law" ensured that they would become a wealthy and powerful group.

Hunter's main problems continued to be those of making the colony survive—the employment of convicts and emancipists (convicts who had served out their sentences) and the production of food and other necessities; but it was soon apparent that the activities of the New South Wales Corps, particularly the officers, represented a serious difficulty. The governor endeavoured to apply the policy laid down by the British government, with the aid of men whose interests were that it should not prevail. The substance of this political conflict was economic interest. The officers and their associates desired to have convicts assigned to them for employment at their own terms and to dispose of their products at the best prices obtainable. They desired to buy, or otherwise acquire, goods at prices most favourable to themselves and enter directly into the disposal of them. They desired nothing less than a monopoly of the import trade, particularly that in rum. The early governors, on the other hand, proceeded to employ many convicts on government land, to take their produce into government stores, and to dispose of it at prices determined by considerations other than making the most profit. Goods were also imported by the government and sold from its own stores. This practice had arisen because at first there had been no other way and then because of the existence of men able to exercise monopoly powers.

The acquisition and disposal of rum, in particular, became a problem. It became the customary currency, the normal means of paying wages, and its possession depended either upon having silver and gold money which shippers would accept or upon distilling it unlawfully. This "exchange economy based on tyranny and rum," which on balance was most harmful to the small community and its future, produced also the kind of initiative and ambition that gave Australia, in its very early years, the sheep industry upon which all its future prosperity was to depend. John Macarthur, the pioneer of the wool industry, was not only the personification of all that had produced so much evil and so much good but he was the man whose egotism and drive was directly responsible for so much of it.

The exchange economy

While the colony in its first decade was growing in population (only a trickle of it was free), in land under cultivation, in output and range of goods, in the number of buildings erected and the extension of settlement, difficulties with the group led by the corps officers grew also. Hunter returned to England in 1800 exhausted by his labours. Uprisings by Irish and other convicts early in the 19th century were perhaps more dramatic but far less significant than the acquisition of power by the corps. P.G. King, Hunter's successor as governor, a man of strong character and fiery temperament, believed he could do better. However, despite his orders and his indignation, the balance of political, social, and economic power continued to move in the direction of the corps officers. Conditions of life and employment of convicts and emancipists—well over 90 percent of the community—did not improve during King's term of office, but the colony made progress. The conflict of governor and corps officers came to a head in King's time with a duel between Macarthur and his own senior officer, William Paterson, who was carrying out the governor's orders. The opposition led by Macarthur was not completely successful, and Macarthur himself was sent to England to face court-martial. King, however, was left in a weakened position and was soon recalled. Macarthur, in England, took the opportunity to resign from the corps and to develop his plans for land and sheep farming in Australia.

SHARPENING OF THE CONFLICT

Gov. William Bligh

The appointment in 1805 of William Bligh (famous for the "Bounty" mutiny) as governor of the colony made a sharpening of the conflict inevitable. Bligh knew what his instructions meant and he determined to carry them out. The general lines of British government policy remained the same throughout the years 1788–1821, and Bligh proceeded to give effect to it. His defeat at the hands of the New South Wales Corps merely meant that the British government realized that its governors could not work without power independent of the dominant colonial pressure group; and, under Macquarie, the policy the corps sought to defeat triumphed. Issues that led to an insurrection by the New South Wales Corps centred in matters of trade, particularly illicit dealings in rum, and in the acquisition of land rather than in anything relating to wool growing or the affairs of the pastoralists. Bligh's policy was recognized by the "little men" to be in their interests, but the power of the corps officers in effecting Macarthur's aims was too great.

Following Bligh's arrest in Government House and his detention, the main activity of the military government was that of granting land, strengthening its own position, and clipping the wings of its competitors. The employment of convicts became little better than slavery, and the retention of a society permanently divided into free and unfree seemed certain.

MACQUARIE

The success of the New South Wales Corps and Macarthur was their undoing. Orders were given for the corps's removal to England, and Macarthur, too, left the scene of his conquests for some time. Lachlan Macquarie was appointed governor with the power to bring his own regiment to the colony. With this power to enforce his decisions, Macquarie arrived to find the select group, "pure merinos" as they were later to be called, in a dominant economic and social position. Macquarie has often been described as a despot, and no doubt a despot he was to those who desired to see a colony of great inequality between free and unfree, rich and poor. His policy increased the prospects of emancipation for convicts and their chances of gaining status and becoming small farmers in their own right. He caused public buildings and roads to be constructed, established a bank, encouraged exploration and the pastoral industry, and introduced currency to take the place of rum. He redressed the balance of social and economic power over a period of 10 years and, as it became certain that no section of the community would ever succeed in completely dominating all others, production and settlement increased and extended. There would have been more economic progress in certain fields if sectional initiative, if not ruthlessness, had been given its head; by 1817 there was room for a change in policy in certain directions. Complaints about Macquarie's administration led to the appointment that year of Commissioner J.T. Bigge to investigate. The British objected to the amount of Macquarie's public works expenditure, and some influential persons in England and Australia desired the emancipists to be given less and the big landowners more scope. This conjunction of interests resulted in the termination of Macquarie's period of office in 1821.

It could now be truly said that Australia was established on a firm economic foundation and in a manner that would not allow any great disparities to development between groups of its people. Both Macarthur and Macquarie deserve credit for this situation, the first for vigour and personal ambition and an interest that served the colony well, and the second for holding a proper balance within the colony itself and for determining to give it the public facilities essential if it was to function.

PASTORAL EXPANSION AND SELF-GOVERNMENT, 1822–55

In these years the foundation of modern New South Wales were securely laid. Self-government was attained; radicalism emerged; voluntarism was established as the basis of colonial religion; transportation of convicts ceased; trial by jury and a free press were allowed; and the rapid expansion of the pastoral industry made the colony prosperous. New South Wales changed from a prison farm into a self-governing colony with a free and expanding economy. The basis of the expansion was wool. Exploration in the 1820s and 1830s opened up the whole colony, and the demand for wool in Yorkshire enabled the squatter pastoralists to avail themselves of the new land, which provided excellent pastures for the fine-wooled Merino. Wool was a source of sterling funds and an inducement for capital imports and immigration. The consequent wealth was a solvent of autocracy. The growing body of free settlers was not prepared to suffer autocratic government and transportation, which lowered the moral tone of the colony. But the impetus for colonial reform came also from Britain.

Increasing importance of wool

Thus, acts of 1823 and 1828 gave New South Wales a nominated Legislative Council; an act of 1842 made this partly representative; and the Australian Constitutions Act, 1850, made it completely representative with the powers of writing a constitution. Nevertheless, until 1855 executive control rested entirely with the governor and his Executive Council, consisting of officials. The Legislative Council was dominated by the pastoralists, who tried to frame a constitution to thwart the democratic aspirations of the city and labour interests; but the bicameral legislature that was finally adopted avoided the autocratic ambitions of the pastoralists and was the future source of much liberal legislation.

The most important social conflict between 1822 and 1855 concerned the status of the former convicts, but, after transportation was discontinued in 1840, wealthy emancipists and exclusives were brought into an alliance to fight for self-government and against the colony's growing radical movement. In religion the privileged status of the Church of England was modified, and the various Christian denominations (including Catholicism) demanded and received legal equality. The attempt by Gov. Sir Richard Bourke to introduce public secular education foundered on the opposition of the churches. But the most important political conflict concerned the alienation of land. The squatters had settled on land beyond "the limits of location" (a defined small area, to prevent undue dispersion and make government easier) and until 1847 had the most temporary titles to their estates. Gov. Sir George Gipps tried to prevent the complete alienation of the state's pasturelands but was unable to prevent in 1847 orders-in-council which gave the squatters favourable treatment. The quarrel over land, however, continued throughout the 19th century.

COLONIAL LIBERALISM, 1856–85

In February 1851 gold had been discovered near Bathurst, but the main tide of the gold rushes soon swung to the

richer fields of Victoria. New South Wales nevertheless gained in wealth and population from gold. In 1856 the new constitution was implemented with a bicameral legislature and responsible Cabinet government. One of its early acts in 1858 introduced the ballot and universal adult male suffrage, and the new democracy soon moved against the squatters. "Radical" legislation in the next 30 years included the abolition of state aid to religion (1862), the Triennial Parliaments Act (1874), the Public Instruction (public schools) Act (1880), and public health legislation (1881). Sir John Robertson's land Act of 1861 aimed at facilitating closer settlement by allowing selection before survey; anyone might select from 40 to 320 acres (16 to 129 hectares) within the "settled" or "intermediate" districts on payment of a quarter of the price, the balance being due, with the title, after three years' residence. Some genuine closer settlement resulted, especially on the coast, but the squatters' devices of "peacocking" and dummying; *i.e.,* picking the best land and using dummy selectors, left most of the western lands in their hands. The failure to settle much of the increasing population on the land meant greater concentration in the towns, besides the rapid growth of trade unionism and radicalism among the growing body of workers. Radicalism before 1870 was reflected in the land legislation and opposition to the British government. Between 1870 and 1885 it encouraged state intervention in social and economic life and increased the political aspirations of the trade unionists. Sir Henry Parkes was the most influential personality in politics, and between 1872 and 1891 his ministries introduced free trade, established nonsectarian public schools, and sponsored railway development. He finally lost office because he favoured federation before the idea was popular in New South Wales.

The ballot and universal male suffrage

This was a period of great economic development. By 1890 New South Wales had become self-supporting in foodstuffs; valuable minerals had been discovered and mined—gold at Captain's Flat in 1861, copper at Cobar in 1869, tin at Inverell in 1871, and silver–lead–zinc at Broken Hill in 1883; 1,215 miles (1,955 kilometres) of railway were completed by 1884; unassisted immigration between the years 1873 and 1893 exceeded 230,000. These developments were greatly promoted by the capital imports that occurred during the 1870s and 1880s, many of them by the government for public works.

LABOUR, NATIONALISM, AND FEDERATION, 1886–1914

This period was dominated by growing nationalism, the rise of the political labour movement, and the federation of the Australian colonies into the Commonwealth of Australia. The period began with the collapse of export prices and the great reduction of capital imports, both of which were consequent on British financial difficulties. This led to wage reductions, industrial unrest, the great strikes of 1890–91 (in which the trade unions were defeated), and the financial crisis of 1892. Although the 1890s were a time of misfortune, culminating in the big drought of 1902–03, these years were the most significant in the history of New South Wales. The failure of direct action in the 1890 strike forced the unions into politics with immediate influence on legislation. G.H. Reid, with Labor Party support, introduced financial reforms (including income tax), removed the public service from political control, reformed the land law to allow the breakup of large estates, and passed the Factories and Shops Act (1896). In a similar manner, after 1900, liberal legislation supported by Labor introduced old-age pensions (1900), compulsory industrial arbitration (1901), women's franchise (1902), and free public education (1906). After federation Labor gradually increased its power until in 1910 J.S.T. McGowen was able to form the first Labor ministry. Without a doubt, the protection that came with federation favoured the development of New South Wales as the centre of Australian heavy industry and thus of the industrial proletariat and the Labor Party. But a similar development also occurred in agriculture, where, between 1900 and 1914, the area of cultivation doubled, which resulted in a rapid increase in the production of wheat, fruit, and dairy products. In 1912 the Riverina was opened up for closer settlement.

Labour strikes

BOOM AND DEPRESSION, THE 1920S AND 1930S

At the beginning of World War I, New South Wales had a Labor ministry under W.A. Holman, but his support of conscription led to his expulsion from the Labor Party and his formation of a Nationalist ministry in 1916. Labor was returned to office in 1920, and Labor and Nationalist ministries alternated until the Depression discredited Labor and put the Nationalists in office for a decade. The failure of Labor after its promising beginning was due mainly to its loss of both social purpose and emphasis on social experimentation, and to internal dissension. The 1920s were a period of boom, with considerable immigration and capital imports, and the expansion of public works (*e.g.,* railway building and the Sydney harbour bridge) and private industry. In the world Depression after 1929, however, New South Wales suffered badly. Recovery was slow and not complete by 1939. John T. Lang, Labor prime minister in 1925–27 and 1930–32, is one of the most controversial figures in the history of the Australian Labor movement. He introduced some important social legislation—*e.g.,* widows' pensions—but was dismissed from office by the governor in 1932 for repudiating overseas debt payments after his government had legally committed itself, by agreement with the commonwealth, to paying them. In a landslide victory the Nationalists under B.S.B. Stevens were returned in 1932 and retained office until 1941. By then, however, Lang had been expelled from the party, and Labor had regained its unity of purpose.

Nationalist election victory

WORLD WAR II AND POSTWAR INDUSTRIAL EXPANSION

Labor governments were in office continuously from 1941 to 1965, with W.J. McKell, J. McGirr, J.J. Cahill, R.J. Heffron, and J.B. Renshaw as prime ministers. The only notable legislation of this period was that for the 40-hour week (1947), compulsory voting for local government (1947), compulsory trade unionism (1953), new liquor trading hours (10 PM closing; 1955), equal pay for women (1958), and reform of secondary education (1961). In May 1965 Labor was defeated at the polls, and a Liberal and Country Party government under R.W. Askin was formed. An extensive plan to modernize the state's port facilities was announced in 1966. In 1967 a referendum in northern New South Wales for a proposed new state of New England was defeated. The history of New South Wales since 1940 has been dominated by the greatest industrial expansion in its history, much encouraged by substantial immigration and capital imports.

Industrial unrest continued to be part of life in New South Wales during this period of economic expansion. Grievances of workers in the coal industry culminated in a prolonged labour strike in 1949, which dislocated the transport and power supply segments of the economy. Additional periodic labour unrest has continued into the late 20th century primarily in public transportation, power supply services, and in shipping.

Some major capital programs began during World War II. In 1940, the production of petroleum from oil-shale commenced at Glenn Davis but was later abandoned because it was determined to be uneconomical. Dock facilities at Port Jackson were expanded in 1944 to accommodate the British Fleet during wartime. The Captain Cook Graving Dock was constructed during the port expansion, enabling this facility to accommodate the largest ships afloat at the time. Expansion of the iron and steel industry also took place during this World War II and postwar industrial activity.

One of the more monumental projects undertaken since World War II is the Snowy Mountains hydroelectric scheme. In 1949, the Snowy Mountains Hydro-electric Power Act created an authority and empowered it to generate electricity hydroelectrically in the Snowy Mountains. All electricity produced from the Snowy Mountains would be supplied to the commonwealth government for defense and other purposes, and for consumption in the Australian Capital Territory. Surpluses of electric energy would be supplied to the states of Victoria and New South Wales. The act was supported by a detailed agreement between the states of Victoria, New South Wales, and the commonwealth regarding responsibilities of each party in

Snowy Mountains hydroelectric scheme

the construction and operation of the hydroelectric system. The agreement established the Snowy Mountains Council, which consists of representatives of each party. The council directs and controls the operation and maintenance of the system. The objective of the Snowy Mountains scheme is to transfer waters, which would normally flow unharnessed to the sea, from the Snowy River and its tributaries to inland rivers and streams for irrigation and hydroelectric power.

Additional public works since World War II include massive improvements of roads, highways, and bridges; the modernization and electrification of railways; school construction; and the building of the Sydney Opera House.

An increasing amount of attention has been paid by government and citizens to the arts, conservation, and environmental protection. Irrigation areas were established as areas of Crown lands subdivided into farms to which intensive irrigation is used for livestock, agriculture, and domestic purposes. Irrigation districts were also created for privately owned lands as well as flood-control districts. Anti-pollution measures have also been instituted in New South Wales.

Commonwealth's dominant state

New South Wales remains the dominant state within the Commonwealth of Australia in the late 20th century. The state derived its prosperity from a sound agricultural base, diversified industry and manufacturing, and strong commercial and financial institutions.

(J.F.Ca./R.M.Hl./Ed.)

NORTHERN TERRITORY

The central section of northern Australia, the Northern Territory is bounded on the north by the Timor and Arafura seas, and by Western Australia to the west, Queensland and the Gulf of Carpentaria to the east, and South Australia to the south. It is approximately 1,000 miles (about 1,600 kilometres) from north to south and 600 miles (about 1,000 kilometres) from east to west. Its area is 519,800 square miles (1,346,200 square kilometres)—17.5 percent of the Australian Commonwealth. It is largely tropical.

Constitutionally, the territory was, until 1978, inferior in status to the states, and it had limited legislative powers until self-government was granted in that year. Its development since 1911, when it was transferred to the commonwealth from South Australia, has been a major item of expenditure in terms of works, services, and inducements to producers to accept the risks of an uncertain physical and economic environment. The nature of the climate, the poor soils, distance from assured markets, and problems of recruiting labour have been considerable handicaps.

Physical and human geography

THE LAND

Relief and drainage. The unspectacular coastline is flat with low headlands and is mostly fringed with mangrove swamps. There are many offshore islands, of which Melville and Bathurst islands and Groote Eylandt are the largest. Inland from the coastal belt there is a gradual rise to the town of Tennant Creek (1,229 feet, or 375 metres) on the vast Precambrian plateau (1,000–2,000 feet) that extends south and west into the neighbouring states. Farther south, Alice Springs (1,790 feet, or 545 metres) is situated on an alluvial plain in the Macdonnell Ranges, of which Mt. Zeil (4,955 feet, or 1,510 metres) is the highest point in the territory. There are some remarkable tors 200 miles southwest of Alice Springs: Mt. Olga (3,507 feet, or 1,069 metres) with 30 separate domes, and Ayers Rock (2,845 feet, or 867 metres), a red, ovoid monolith rising 1,100 feet.

A number of rivers, of which the Finke and the Todd are the largest, flow from the central ranges after rainstorms. Areas north of the plateau are drained by some substantial rivers: the Victoria, 350 miles long, and the Daly, 225 miles long, flow to the Timor Sea; the Katherine flows southwest from Arnhem Land to join the Daly; the Adelaide, Mary, and South and East Alligator rivers enter Van Diemen Gulf; and the Roper and the McArthur flow east and northeast into the Gulf of Carpentaria.

Plant and animal life. Some forests with Indo-Malaysian vegetation elements exist in Arnhem Land, but otherwise the northern vegetation is open woodland with low eucalypts and tall grasses of low nutritive value. In the main cattle areas of the Victoria River Downs and the Barkly Tableland, an open-tussock grassland on heavy, gray-brown cracking soils is dominated by Mitchell grass (a perennial Astrebla species) with subdominant Flinders grass and herbs. Between the Barkly Tableland and the ranges a wide belt of pervious sand supports only spinifex. In the Alice Springs district the rocky hills carry spinifex, but on the light-textured soils of the lower slopes and valleys an association of mulga (an acacia tree) with short grasses is a valuable fodder resource. Mixed mulga–spinifex scrub with seasonal herbs and ephemerals occupies the nearby red plains of desert loam, and farther to the west is a desert of hummock grassland composed of widely spaced clumps of spinifex and *Plechtrachne*.

Higher plants exceeding 5,000 species are represented. Species endemic to the Macdonnell Ranges include a cycad, *Macrozamia macdonnellii;* a palm, *Livistona mariae*—a surviving representative of the vegetation that once covered central Australia; and several acacias and eremophilas. A broad-leaved species of mulga, the witchety bush, harbours a grub much sought by Aborigines as food. The baobab tree, with a girth of 30 feet, occurs near the Victoria River.

Exotic animal life

Principal birds are the princess parrot of the central, rocky spinifex country, the flock pigeon of the grasslands of the Barkly Tableland, and lorikeets, parrots, and rock pigeons in rugged areas of the north. The black-banded pigeon is known only in the western escarpment of Arnhem Land. Mammals include one of the egg-laying species—the echidna—and a variety of marsupials. The kangaroo is widely distributed, but some species have restricted habitats: the rat kangaroo is adapted to the arid regions; the rock wallaby and antilopine wallaroo inhabit the rocky ridges of the northwest; and the black wallaroo is restricted to the granitic ranges of Arnhem Land. The unique rock-haunting ring-tailed opossum is an interesting evolutionary development. Formerly domestic animals now existing as large wild populations include camels, buffalos, cattle, pigs, goats, horses, and donkeys. With the exception of the buffalos, which are hunted for hides and meat, these are serious pests. The cattle tick also causes great economic problems.

Climate. There are two seasons: a wet summer, from November to April, and a dry winter. Rainfall is extremely variable and of marked summer incidence. The change from the north, where the annual rainfall is 60 inches (1,525 millimetres), to the southeast, where it is only five inches, reflects the diminishing influence of the Australian-Asian monsoon and an increasing dependence on tropical thunderstorms. Only 30 percent of the territory has an annual rainfall of between 15 and 40 inches—the effective limits of agricultural development. The climate is hot and, in the north, uncomfortably humid for eight months of the year. Mean temperatures at Darwin are 84° F (29° C) in January and 76° F (24° C) in July, and at Alice Springs 81° F (27° C) in January and 53° F (12° C) in July. The north is free from frosts.

Settlement patterns. The pastoral potential of the territory was first recognized during construction of the Overland Telegraph Line, and on its completion, in 1872, areas near Alice Springs and Darwin were stocked with cattle. The establishment of cattle stations on the Barkly Tableland during the 1880s completed the foundations of the pastoral industry. The land is rented from the government; tenure of pastoral leases is normally 50 years, and the covenants stipulate conditions of maintenance and improvement. These have been minimal, and there

are inadequate safeguards against ill-advised practices of absentee leaseholders, which have caused damage to vegetation and extensive soil erosion.

The two traditional regions

The territory is traditionally viewed as two regions, roughly delineated by latitude 20° S and commonly referred to locally as the North and the Centre. The two areas differ in topography and climate and represent the spheres of interest of Darwin and Alice Springs. As natural outlets for cattle and mineral concentrates from their respective regions, these centres have developed rapidly. Little urban growth has occurred elsewhere: Tennant Creek and Katherine have urban populations, but there are few other groupings of more than 20 dwellings.

THE PEOPLE

Demography. During the gold boom at Pine Creek in 1872, labour problems led to the engagement of Chinese from Singapore, and by 1881 the population comprised 670 Europeans and 2,781 Chinese. By 1888, when immigration restrictions were imposed, Chinese numbered about 6,000. Subsequent immigration was mostly European, and at the 1966 census the population was almost two-thirds European and one-third Aboriginal. The vast majority of the whites are of Australian birth. Age structure and other population characteristics are typical of Australia. Of the tens of thousands of persons claiming to be Christian, one-third are Catholic and two-thirds Protestant. In the middle and late 20th century more than one-third of the population was under 21 and more than three-quarters under 40 years of age. There was more than one male to every female. Whereas most of the white population is concentrated in the towns, about eight out of 10 Aborigines live in rural areas or in the Aboriginal reserves, which total more than 94,000 square miles.

In the Northern Territory the birth rate was much higher than the Australian average, though the death rate was on a par with the rest of the country. Infant mortality was more than double the national average, with greatest incidence among Aborigines.

The Aborigines. Few Aboriginal tribes are wholly nomad. Many have retained their tribal structure and their customs and religious rituals that govern their social relationships. Their religious and magical rites draw on a rich repertoire of songs and dances in which sacred objects and personal adornment play an important part. These are accompanied by a variety of musical instruments. Tribes and dialects are numerous. Languages are of the agglutinating type; *i.e.*, they combine into single words two or more elements of distinct and separate meaning. Though many of the languages have characteristics in common, the prefixing languages of the north are basically different in structure and vocabulary from the Aranda languages of the south. Many tribes have permitted in the main their languages and religious rites to be placed on permanent record. Government policy on Aboriginal welfare is one of assimilation, promoting their participation in the privileges and responsibilities of citizenship so that they may become an integral part of the Australian way of life.

Aboriginal languages

Aborigines, as Australian citizens, are entitled to equality before the law. The Department of Aboriginal Affairs administers various programs designed to benefit Aboriginal citizens. In order to better administer these programs the department, in conjunction with other commonwealth agencies and departments, established a legal definition of Aborigine as a person of Aboriginal or Torres Strait Islander descent, who identifies himself as such and as accepted by the community in which he or she lives as an Aborigine or Torres Strait Islander.

The Northern Territory is committed to governmental policies of Aboriginal self-management and self-sufficiency. The government attempts to secure equal access to services for Aborigines. Additional programs are oriented toward the social and economic betterment of the Aboriginal citizen, providing special benefits not available to other citizens. Bilingual education programs were initiated in 1973 in many Aboriginal communities. Legal services are provided by the Department of Aboriginal Affairs to ensure that all Aborigines are represented in the courts in a competent manner.

The Aboriginal Land Rights Act of 1976–78 was established for the Northern Territory, giving traditional Aborigines inalienable freehold title to former reserve land while providing a procedure for Aborigines to claim areas of unalienated Crown land. A judge of the Northern Territory Supreme Court acts as the Aboriginal Land Commissioner, ajudicating claims. Aboriginal communities are also assisted in the purchasing of land off the reserves. Money from the Aboriginal Land Fund is used for this purpose.

THE ECONOMY

Agriculture and mining. The traditional dependence of the economy on cattle raising has been substantially changed since 1966 by developments in mining and other primary industries. Agricultural production was small and related to local needs, until 1969, when the exportation of grain to Japan began. Seven companies commenced prawn fishing and processing in the same year. Pastoral production is confined to beef cattle, grazed largely under open range conditions.

Mining of iron ore at Mt. Bundey and Frances Creek (near Darwin), and of manganese ores at Groote Eylandt (an island off the east coast), commenced in 1965–66. Much of the iron ore was exported to Japan while much of the manganese ore went to Europe, Japan, and the United States. Copper and gold ores are mined chiefly at Tennant Creek. Since 1971 the Northern Territory has been among the country's most important copper- and gold-producing regions.

Manufacturing and services. Secondary industries are small service industries meeting local needs. Tourism is increasing, and many visitors are attracted by the climate of Alice Springs and by the stark, colourful scenery of the Macdonnell Ranges and the Mt. Olga-Ayers Rock National Park.

Transport. The Northern Territory has one standard gauge railway: from Alice Springs to Port Augusta in South Australia. The Stuart Highway (954 miles, or 1,538 kilometres), from Alice Springs via Tennant Creek and Katherine to Darwin, and the Barkly Highway (403 miles, or 649 kilometres), connecting Tennant Creek with Mount Isa, Queensland, are the main road links. Air services are well developed: many overseas airlines serve the international airport at Darwin, and there are services between state capitals and Darwin and intermediate towns, a substantial internal network, and charter services at Darwin and Alice Springs. There are many small airports, and most homesteads have airstrips. Shipping from Darwin has increased and necessitated a large extension of port facilities.

Development of air services

ADMINISTRATIVE AND SOCIAL CONDITIONS

Government. The Northern Territory Self-Government Act of 1978 established the Northern Territory as a self-governing entity. Under this act the commonwealth transferred most of its powers of government to the territory. The government of the Northern Territory is similar to that of the states in fields of transferred authority. There are, however, differences in office titles. For example, the Northern Territory has an administrator instead of a governor and a chief minister in lieu of a premier.

The administrator is appointed by the governor general and has the responsibility for administering the Northern Territory's government. The Executive Council, composed of all ministers in the Northern Territory, acts as adviser to the administrator. The administrator acts on all matters within the realm of the territorial government. Matters not within the realm of the government are handled in conjunction with commonwealth advice.

The Legislative Assembly serves as the Northern Territory's parliament. The Assembly is composed of members elected for four-year terms. A speaker is elected by the members of the Assembly as well as six ministers to serve on the Executive Council. A number of territorial departments and authorities operate within the realm of transferred powers. The only territory-related administrative responsibilities still retained by the commonwealth deal with uranium and other prescribed-substance mining and Aboriginal land rights.

Local government was established in the city of Darwin in 1957. Since that date local government has been extended to other towns and cities in the territory. Local municipal councils are elected by universal adult franchise. Elections are held at intervals of not more than three years.

The Northern Territory is represented in the federal government in Parliament. The territory elects one member to the House of Representatives and two members to the Senate. The legal system and the jurisdiction of the courts are similar to the other states. Darwin, as a defensive port, has extensive army, air force, and naval installations.

Health. The Department of Health provides medical and dental services and maintains hospitals at Darwin, Alice Springs, Katherine, and Tennant Creek, and a leprosarium near Darwin. Its aerial services to remote areas are based at Darwin; it also provides the medical personnel for the Royal Flying Doctor Service of Australia, which operates within 500 miles of Alice Springs. Two-way radio consultation usually suffices for minor ailments. Health inspectors visit all settlements, and medical and dental officers make periodical examinations of all schoolchildren.

The Flying Doctor Service

Education. Education is the responsibility of the minister of education, under the Northern Territory Education Act of 1979. Advisory councils were established by the act to assist the minister in special areas of education. The territory is served by several preschools, primary schools, and high schools. Educational services are also available through correspondence and private institutions. Area schools offer courses in secondary education, and three residential colleges are available for Aboriginal students. Government schools offering post-primary courses are located in many Aboriginal communities. Mission schools in some Aboriginal communities offer similar educational opportunities. Special schools for the handicapped are located in Darwin and Alice Springs.

Aboriginal education is conducted, for the most part, away from the population centres in government and mission schools close to the Aboriginal communities. Government schools at the preschool and primary levels provide tuition assistance for Aboriginal children. Bilingual education programs are widespread throughout the Aboriginal population. Courses in Aboriginal culture are also prevalent in many government and mission schools. Children wishing to continue their education beyond the primary level may attend one of the regional colleges for Aborigines. Instruction in a wide range of secondary courses is obtained at the college or in a regional urban high school.

Darwin Community College (1973) is a multipurpose institution serving the Northern Territory. It provides chiefly technical and professional courses at the university level.

The Northern Territory Industries Training Commission administers vocation and apprenticeship training programs. Members of the commission represent government, education, employers, and employees. The commission is involved in research, accreditation, promotion, financial assistance, and advice on matters pertaining to industrial training within the Northern Territory.

Technical and adult education is also offered in the territory. On-site Aboriginal adult education courses are provided by local educators who teach a variety of vocational skills. Other colleges that provide vocational training include Batchelor College, Community College of Central Australia, and Katherine Rural Education College.

CULTURAL LIFE

Cultural institutions are naturally limited in a region of small population of very low density. Because of its isolation—Darwin, for example, is nearer to Hong Kong than to Sydney—the territory is largely deprived of opportunities afforded by visiting artists. Partial destruction of Darwin during World War II led to replanning on modern lines, but even though the decade following its proclamation as a city in 1959 saw the population increase from 13,000 to 32,000, Darwin lacked museums and art galleries and a seat of higher learning. There are, however, active art, musical, and dramatic societies, and the performing and fine arts are occasionally brought to the larger towns through the sponsorship of the Arts Council of Australia.

Aboriginal arts

Indigenous arts are those of the Aborigines, which symbolize the ritual of their religion. Throughout the territory there are examples of sculpture, rock carvings, bark and rock paintings, and "X-ray" drawings (which show not only the outlines but the bone structure and internal organs of creatures of esoteric significance). The best of the decorative arts are produced in Arnhem Land and range from gravestones and ceremonial posts to personal ornaments in attractive, polychromatic designs. Aboriginal art has received wide recognition by artists and designers, and Aboriginal motifs figure prominently in the decor of many public buildings in the state capitals.

Regional national broadcasting stations are situated at Darwin, Alice Springs, Tennant Creek, and Katherine, and there is one commercial station at Darwin. The first television station came into operation at Darwin in 1971. Library services are provided from the five principal centres. Newspapers are published weekly at Alice Springs and Tennant Creek, and the *Northern Territory News* is published daily at Darwin.

(A.E.Sc./Ed.)

History

EARLY EXPLORATION AND SETTLEMENT

For some centuries before the beginnings of European exploration, the coast of the Northern Territory was visited by traders from Macassar for the purpose of obtaining sandalwood and bêche-de-mer. The earliest European explorations were those of the Dutch from Java, between 1606 and 1756. In 1623, the Dutch ship "Arnhem" explored the coastline of the Northern Territory, naming it Arnhem Land. Later, in 1644, Abel Tasman sailed along the coast, landing near the mouth of the Victoria River. Dutch explorers were followed by a period of British exploration, notably by Capt. Matthew Flinders (1803), who charted part of the coastline, and Lieut. Philip King (1817), who discovered that Melville Island was not part of the Australian mainland. The Frenchman Nicolas Baudin explored the area of Melville Island in 1803, applying French names to several physiographic features. No immediate results were derived from these explorations, and the territory remained without European settlement until 1824, when, in order to forestall possible French annexation, a settlement was founded by Capt. Gordon Bremer at Port Essington; it was removed to Melville Island, where Fort Dundas was constructed 10 years before the settlement of Melbourne, Perth, or Adelaide. A further settlement was established by Capt. James Stirling at Raffles Bay in 1827, but both this and Melville Island settlement were abandoned in 1829. A settlement was established on a new site at Port Essington in 1839 by Bremer. The settlement grew and soon had a hospital and more than 20 cottages. In July 1839 the HMS "Beagle" entered Port Essington, and three months later its first officer, Lieut. John Lort Stokes, explored Port Darwin. Prior to the arrival of the "Beagle," Bremer had been called upon to help suppress an uprising in Canton. The settlement had continued under the leadership of Capt. John McArthur. Conditions began to deteriorate at Port Essington, and many settlers died of fever. Finally, in December 1849 the settlement was abandoned.

Melville Island settlement

PERMANENT SETTLEMENT

No further settlements were attempted during the next 15 years, but the interior of the territory was extensively explored, the most outstanding expeditions being the crossing of the continent by John McDouall Stuart (in 1860 and 1862), and by Robert Burke, William Wills, and John King (in 1861). In 1863, through an issue of royal letters patent by the imperial government, the Northern Territory came under South Australian administration. In the same year a land settlement scheme was sponsored by the South Australian government, and a permanent centre of administration was established at Palmerston (now Darwin) in 1869.

The township of Palmerston was surveyed (1869–70), and Capt. Bloomfield Douglas was appointed the government resident. Between 1870 and 1872 the overland telegraph line was constructed linking Adelaide and Darwin. The

first newspaper was published in the territory in 1874, and in the same year Port Darwin grew to nearly 800 persons (three-quarters European and the remainder Chinese and Malay). The discovery of gold at Pine Creek boosted the population to more than 1,700 persons (50 females) by the end of the year. The first railroad began operation in 1889 between the goldfields in Pine Creek and Port Darwin.

19th-century cattle raising

In the late 1860s cattle were introduced from the south, and by 1888 there were 218,000 cattle, 107,000 sheep, and 5,600 horses in the territory. In 1872 gold was discovered south of Grove Hill. Further discoveries were made at Pine Creek, Union Town, Brock's Creek, and Burrundie, and within a year the population of the territory increased from 1,000 to 7,500. The introduction of Chinese labour in 1875 gave some stability to the mining industry, and by 1888 most of the gold produced was being won by Chinese. Copper was found in 1882 at Daly River and in the McArthur River area in 1891, and by 1911 several other minerals had been found in the territory.

In 1890 the Northern Territory was given parliamentary representation at Adelaide, allowing the election of two members to the House of Assembly. In 1901 the residents of the territory were given representation in the federal House of Representatives and were allowed to vote for South Australia candidates for the Senate.

COMMONWEALTH ADMINISTRATION

The territory, despite the efforts of the South Australian government, remained underdeveloped, having a population of only 3,310 (2,734 males and 576 females) in 1911. On January 1, 1911, following nearly 10 years of negotiations, it was transferred to the commonwealth under the Northern Territory Acceptance Act 1910. The Northern Territory (Administration) Act 1910 provided for the territory to be administered, subject to the instructions of the minister for external affairs, by an administrator appointed by the governor general. In the first years of commonwealth administration a number of investigations into the resources of the territory were carried out, several agricultural experimental farms were established, the mining industry was subsidized, the South Australian land laws were amended, and stock routes were improved with wells and bores. In 1917 the railway was extended from Pine Creek to Katherine. Representation in the Federal House of Representatives by a nonvoting member was granted in 1922, following local unrest.

Northern Australia Act of 1926

From 1927 to 1931 under the Northern Australia Act of 1926 the territory was administered in two parts separated by the 20th parallel of south latitude—Central Australia and North Australia—each with a government resident assisted by an advisory council. The act also provided for the establishment of the North Australia Commission with certain powers in relation to the development of North Australia, in particular in connection with transport and communications and water boring and conservation. During this period the North Australia railway was extended from Katherine to Birdum, and in the same year (1929) the extension from Oodnadatta to Alice Springs was completed. The onset of economic depression, however, brought about a serious curtailment of developmental work, and in 1931 the North Australia Act was repealed, the commission was abolished, and the administration of the territory as a whole was once again placed under an administrator at Darwin.

Electricity and water reticulations were introduced in Darwin in 1934 and 1940, respectively, and in Alice Springs in 1937 and 1942; and in 1940 a beginning was made on the Barkly and Stuart highways, which were completed in 1943.

WORLD WAR II AND LATER

Women and children were evacuated from Darwin shortly after the bombing of Pearl Harbor in December 1941. United States troops and naval units were stationed in Darwin early in 1942. On February 19, 1942, Japanese aircraft attacked Darwin, sinking and damaging several naval vessels and killing and wounding several hundred persons.

Bombing of Darwin

Following the Japanese bombing of Darwin, the northern part of the territory was placed under complete military control, and the civil administration, responsible for the remainder of the territory, was transferred to Alice Springs, where it remained until the end of the war.

Although Darwin was severely damaged by air raids and the civilian population had to be evacuated, the territory emerged from the war with greatly improved facilities for air and road communication provided by defense construction and with its main industry, the pastoral industry, in a healthy state. After the resumption of civil control the former residents returned and there was an influx of new arrivals.

Elaborate plans to redevelop a large area around Darwin were drawn following the war. These plans were abandoned in 1950 and a more feasible approach of rebuilding Darwin undertaken. A number of surveys and investigations into the territory's resources were then carried out, especially by the Commonwealth Scientific and Industrial Research Organization and the Bureau of Mineral Resources. The output of the mining industry increased steadily, due mainly to the larger copper and gold production from Tennant Creek, and to the uranium deposits discovered at Rum Jungle in 1949. Other deposits included iron ore at Mount Bundey and Frances Creek, manganese on Groote Eylandt, and bauxite at Gove Peninsula.

Legislative Council

In 1947 the Northern Territory Administration Act was amended, creating a Legislative Council composed of both appointed and elected members, with an administrator, appointed by the governor general, as president of the council. The council was charged with the responsibility of making laws for the peace, order, and good of the Northern Territory government. Laws passed by the council were forwarded to the administrator for approval. Certain laws required the approval of the governor general. The governor general could approve the laws, withhold approval or approve only a part of the law, or return the law to the council with recommended amendments. In 1959 the act was amended once again.

The Legislative Council was converted into a wholly elected body in 1974, when the Northern Territory Administration Act was again amended. The amendment called for an elected council of 19 members with a speaker elected from the membership.

The Commonwealth began a program of transferring administrative powers to the Northern Territory government in 1977. Administrative powers were transferred to the newly created Northern Territory Public Service. The positions of executive member were created, with each executive member acting as a minister in charge of certain services such as fire protection, local government, and correctional services. The Administrator's Council was replaced by the executive members, known as the Executive Council. On July 1, 1978, the Northern Territory Self-Government Act went into effect. (Ed.)

QUEENSLAND

Comprising the entire northeastern portion of Australia, the State of Queensland reaches well north of the Tropic of Capricorn, and it was the first successful settlement of European peoples within a tropical climate. The 1,000 miles (1,600 kilometres) of its eastern coastal region are separated by the northern reaches of Australia's Great Dividing Range from the vast inland plains. Extending for 1,250 miles off this coast is the Great Barrier Reef, one of the most remarkable coral formations in the world.

Queensland's 667,000 square miles (1,730,000 square kilometres) make up nearly one-quarter of all Australia and nearly one-third of the occupied part of the continent. It is bounded on the north and east by the Pacific Ocean, on the south by New South Wales, on the southwest by South Australia, and on the west by the Northern Territory. Although Brisbane, the capital, contains slightly less

than one-half of the residents of the state, smaller but sizable cities, especially along the coast, have dispersed the population across a larger area than in any other state of the commonwealth. Settled originally on the basis of the grazing potentiality of its great grasslands, Queensland has succeeded in increasing the diversification of its economy, a significant part of which derives from the allure of its tropical resorts.

Physical and human geography

THE LAND

Relief and drainage. The inland two-thirds of Queensland comprises a vast plain, broken in a few places by low ranges and hills, whereas the eastern (coastal) third is hilly, occasionally mountainous, with wide valleys and plains between the ranges. The Great Dividing Range runs the entire length of the state, separating east-flowing from west-flowing rivers. In South Queensland it lies close to the coast, and peaks reach to 4,500 feet; but it becomes a low range 250 miles inland at the Tropic of Capricorn, and again nears the coast in North Queensland. Between the Dividing Range and the coast are many other ranges, more conspicuous than the Divide in Central and North Queensland, where the state's highest peak is Bartle Frere, at 5,287 feet (1,611 metres).

Mountains, rivers, and soils

The rivers of Inland Queensland are intermittent; that is, they flow only after heavy rains, when they flood widely across the flat plains. In the arid southwest, the rivers branch into thousands of distributaries, and big floods are followed by lush growth of clover and other herbage in what is known as Channel Country. Most coastal rivers are also intermittent in flow and shallow, navigable only in the tidal reaches.

Most coastal soils have low fertility, though occasional alluvial and other richer soils occur. Large areas of rich black, gray, and gray-brown soils are found inland, forming the basis for one of the greatest natural grasslands of the world. The Great Artesian Basin underlying the plains provides water for livestock.

Climate. The rainfall of Queensland ranges from 180 inches (4,500 millimetres) a year on the wettest part of the northeastern coast to five inches (125 millimetres) in the southwest. All Queensland receives more summer rain than winter rain, but the latter is important in the southeast, in which large areas of wheat and other winter crops are grown. Rainfall decreases in quantity as one moves inland, where droughts are frequent.

Queensland's summer temperatures are warm to hot, humid on the coast but usually dry inland. Winters are mild and sunny, with frosts occasional on the southern coast and frequent inland. Average maximum summer temperatures, registered in January for the southern coast, the northern coast, and the central inland area, are about 85° F (29° C), 88° F (31° C), and 100° F (38° C), respectively; comparable winter minimums in July are 49° F (9° C), 63° F (17° C), and 45° F (7° C). Humidities are greatest in the north, lowest inland.

Plant and animal life. The wetter or more fertile parts of the east coast foster several areas of dense rain forest and its lesser variant, the softwood scrub. Otherwise, the subcoastal vegetation is more open forest of eucalyptus. Further inland, the dominant trees are mostly acacia species—brigalow, mulga, and gidyea—interspersed with large grassy plains and with some areas of desert spinifex, a spiny grass growing in clumps.

Tropical rain forests and exotic animals

Queensland has a great variety of animals—50 species of marsupials, 400 of birds, and 1,600 of fish. Both the short-nosed echidna and the platypus—egg-laying mammals—are found there. Marsupials range from the large red and gray kangaroos to koala bears, opossums, cuscusses, and marsupial mice. Birds include the flightless emu of the plains, great flocks of coloured parrots, and songbirds of the forests. Fish include the freshwater Queensland lungfish, eating and gamefish, and beautiful reef fish.

Settlement patterns. The strong regional sentiment within Queensland arises from the size and dispersion of the populated areas, some 1,000 miles from south to north, 800 miles east to west. The main regions are South Queensland, Central Queensland, North Queensland, and Inland Queensland.

South Queensland extends north to Bundaberg and some 200 miles inland, including the Darling Downs farming subregion. About three-quarters of Queensland's inhabitants live in South Queensland. Rockhampton, Central Queensland's largest city, is the outlet and commercial centre for the beef cattle and wool produced in its hinterland. This region's vast coalfields also have been developed.

North Queensland is identified with sugarcane, producing about three-quarters of Australia's crop. Subregions are Mackay and the Atherton Tableland. Townsville serves as the commercial centre for the pastoral hinterland and has a copper refinery nearby. Coalfields lie inland from Mackay and Bowen. Inland Queensland is the great pastoral area of the state. Popularly called the West, it has an identity overriding its 900-mile spread and is sharply different from the coast.

Almost all of Queensland has been occupied by pastoralists for 80 to 100 years, but large areas of the state are little changed from the original landscape. In western Queensland, the flocks of sheep and herds of cattle mainly graze the natural grasslands, the interspersed belts of trees providing only light grazing. In southern Inland Queensland the mulga is a useful drought reserve, since the leaves are edible by livestock. It is cut or rolled down for this purpose, and it regenerates.

Nearer the coast, some large areas of open eucalyptus forest have been "ringbarked," a process that kills the trees and allows grass to grow. A large belt of brigalow timber, an acacia looking much like an eucalyptus, in the 20-to-30-inch-rainfall belt of Central and South Queensland has been partially cleared for farming and grazing, since the presence of brigalow indicates good soils.

In the crop-growing and dairying region of the southeast and on the sugar-growing north coast, timber has been completely cleared from the better soils, mainly river alluviums and basalt areas. Elsewhere, much original timber remains, and there are many national parks and hardwood forests.

The outstanding feature of land settlement in Queensland is the large size of the grazing holdings and farms. Most of the western grazing country is held in leasehold from the crown, and well over a thousand pastoral holdings average more than 100,000 acres in area. On the better grassed country the holdings are 15,000 to 25,000 acres, carrying 5,000 to 10,000 sheep. In the grain-farming areas much of the land is freehold, and most farms are between 600 and 3,000 acres in size. Large holdings, with a high degree of mechanization and a small labour force, suit the rather unreliable rainfall pattern of Inland Queensland. In the coastal sugar areas, farms range from 50 to 100 acres and are highly mechanized.

Farm holdings and township patterns

The large holdings in rural Queensland have produced a sparse pattern of villages, or townships. In the far west and in the rougher rangy areas of the subcoastal belt, the grazing properties, or stations, are very large, employing six to 20 people and carrying more supplies than a village store. Hence, townships often lie as far apart as 30 to 100 miles.

Recent decades have seen a decrease of population in most of western Queensland. In the agricultural areas containing mainly single-family farms, the small towns are closer, about 10 miles apart; but with the improvement of roads and vehicles some small towns are disappearing, and towns of more than 1,000 people and lying 40 to 60 miles apart are serving the rural areas.

Queensland has a more decentralized population than any other state. Apart from Brisbane, there are fewer than a dozen cities with more than 20,000 inhabitants. The development of ports and commercial centres along the Queensland coast and the growth of the sugarcane industry are in large part responsible for this widespread settlement.

THE PEOPLE

Ethnic groups. More than four-fifths of Queensland's population is Australian born, and, as in other states, it is predominantly of English, Scottish, or Irish ancestry. Important minorities came from Germany in the 19th century and from Italy after 1920, the latter frequently

becoming cane farmers in North Queensland. After 1947, a wider range of European immigrants arrived from The Netherlands, West Germany, and southern and eastern Europe, as well as some from the United States.

The assimilation of Europeans has been rapid once some fluency in spoken English is achieved, and religion has not proven to be a barrier. The assimilation of the Australian Aborigines, on the other hand, has been slower.

Queensland's birth and death rates are close to the Australian average. Historically, Queensland's birth rate has been significantly higher than the Australian average, reflecting a more rural population. The margin has narrowed as urbanization has increased.

Scourge of tropical diseases in the 19th century

From 1860 to 1890 the death rate was substantially higher than the Australian average. During this period, Queensland's future was threatened by tropical diseases brought in by troops from India, by Chinese miners, and by South Pacific islanders. Almost every tropical disease was introduced, and Queensland seemed likely to suffer the same fate as previous European attempts to settle in the tropics. Intense and continued effort by the medical profession and the Queensland health services, however, progressively wiped out the diseases. Australia's policy of restricted immigration was strongly influenced by Queensland's experience of unrestricted immigration, and after about 1890 the state was developed almost entirely by European labour.

The number of rural workers has declined with increasing mechanization of farming and the abandonment of some small farms and dairy farms. These trends have far more than offset the increased population in those few rural areas that still are expanding, notably the brigalow lands.

Although Queensland is the most decentralized of Australia's states, the trend toward increased urbanization is strong. The development of Queensland's great mineral resources, contributing to the growth of mining towns and ports, is both a decentralizing and urbanizing force.

THE ECONOMY

Queensland's economic pattern is similar to that of the other states, but its manufacturing is less developed than that of New South Wales and Victoria in such areas as steel, chemicals, and transport equipment. It is more developed in the production of certain minerals, in tropical crops, and in extensive cattle raising.

Manufacturing, commerce and finance, construction, agriculture, and professional and personal services are the areas of greatest employment in Queensland. Seasonal unemployment is more noticeable in Queensland than in other states, mainly because of the seasonal nature of sugar harvesting and cattle slaughtering and the decline in construction work in the wet months from December to April.

Agriculture. The original attraction of Queensland is the great natural grasslands of its interior. The grasslands remain the state's largest single resource, even with the growth of manufacturing, mineral exploitation, and agriculture.

Domination of agriculture by livestock

Wool production is almost all of the fine-textured Merino variety. Beef cattle are tending to replace sheep in some areas and already have replaced many dairy cattle. Some of the best croplands in Queensland are fully developed: for sugar production on the coast and for grains on the subcoastal Darling Downs. Other crops include peanuts, tobacco, cotton, and many tropical and temperate-climate fruits and vegetables. Queensland has a greater area of land suitable for further development than any other state, especially brigalow and similar lands suitable for beef grazing and summer grain crops.

Manufacturing. Queensland's manufacturing and processing industries now exceed the output of the rural industries in their contribution to the state's income. Brisbane, which employs almost two-thirds of the factory workers, has by far the greatest range of manufacturers. Elsewhere in the state, most manufacture comprises the processing of such primary products as sugarcane, meat, timber, and alumina and various base metals, as well as in industry supplying local needs—printing, baking, vehicle repair, and the like.

Mining. The most spectacular development in Queensland in the middle and late 20th century was in minerals, including alumina and coal. The great bauxite field at Weipa on Cape York Peninsula exports bauxite overseas and supplies the big alumina refinery at Gladstone. The Mt. Isa mines in the northwest greatly expanded their production of copper and of silver, lead, and zinc. Several large coal mines for the export market have been opened up in the hinterland of the central coast. Brisbane is supplied natural gas and oil from the Roma fields and oil from those at Moonie.

Tourism. A major Queensland industry, the tourist facilities of the state are serving a rapidly increasing number of Australian and overseas visitors attracted by the mild, sunny winters and moderate summers. Prime attractions include the surfing beaches and the Great Barrier Reef islands, as well as the rain-forested national parks and tours of the eastern coast. The surfing beaches are in South Queensland. Gold Coast, 20 miles of beaches near the southern border, supports a booming tourist and holiday industry.

Administration of the economy. The state government in Queensland regulates those parts of the economy not under federal control and directly provides some services, especially railways, main roads, education, health services, police and law, some forestry and irrigation, and some housing. Operating under state legislation, local authorities provide minor roads, some bus services, water supply, drainage and sanitary services, and town planning. Statutory bodies provide electricity, harbours, slaughterhouses, and marketing of some primary produce.

Federal and local services

State industrial awards (wages and hours fixed by the state-appointed arbitration authority) cover some 65 percent of the employees in Queensland, a larger percentage than other states. Notable among employee unions is the large Australian Workers Union, covering shearing, grazing, canefarm, and mill employees and construction workers.

Transportation. Queensland developed as a series of ports, with railways tapping the inland, but now coastal shipping is employed exclusively for heavy cargoes such as steel and sugar. Brisbane is the largest receiving port, while Gladstone (alumina and coal) and Weipa (bauxite) are the largest shipping ports. Several thousand miles of railway operate throughout the state of Queensland. Closure of some branch lines in agricultural areas has occurred; but two railways have been constructed for coal exports, and other lines have been improved. Air services are important in Queensland, especially in getting about the large area of the inland.

A steadily improving road network links all parts of the state, but maintenance of unpaved roads is costly and difficult with heavy summer rainfall. A paved road reaches Cairns and is partly constructed to the main inland towns. Paved beef roads have been built to get fat cattle to railheads from the southwestern Channel Country and the northwest.

Brisbane is the only city with serious traffic congestion, and a 25-year program of freeway construction is under way. Most provincial cities also have planned similar but less extensive systems.

ADMINISTRATIVE AND SOCIAL CONDITIONS

Government. The state government follows the usual British pattern of a legislature, an executive, and a judiciary. Queensland is the only state with a unicameral parliament, the upper house having voted itself out of existence in 1922. Local authorities are established under the Local Government Acts; they are elected through adult franchise, and voting is compulsory. Unlike the situation in other state capitals, most of the metropolitan area of Brisbane is controlled by a single city council. Public justice is vested in the Supreme Court and district courts and, for civil jurisdiction, in lower courts as well. The Supreme Court holds periodical sittings in centres throughout the state.

The Australian Labour Party controlled the state for two long periods, 1915–29 and 1932–57. Since 1957 a coalition of the Country and Liberal parties has governed.

Social welfare. Social and economic divisions are even less marked in Queensland than in the rest of Australia, which itself makes up the world's largest area of uniform language, social customs, and standards of living.

Education. The vast majority of primary and secondary school children attend state-government schools, and the rest are in schools run by religious and other bodies. Education is compulsory between ages six and 15.

Private schools receive governmental subsidies, as well as grants for libraries and science laboratories. Correspondence schools always have been important to the remote parts of the state, and thousands of children, including apprentices, are educated in this way. Regular school assessments of pupils are made, and university-admission examinations are held.

Universities are located in Brisbane and Townsville and institutes of technology in Brisbane, Toowoomba, and Rockhampton. An agricultural college, a conservatory of music, and many technical colleges complete the state's offerings in higher education.

Health and housing. Queensland is the only state with a public hospital system that provides free wards as well as the usual subsidised private wards that charge fees. Among the medical-research institutes is the Radium Institute, which successfully deals with skin cancers, a high-incidence affliction in Queensland. The Flying Doctor Service, which commenced in Queensland in 1928, now makes hundreds of flights a year to remote farms or settlements, especially in the inland region.

Prevalence of single-family homes

Several thousand dwelling units a year are built in Queensland, of which the vast majority are separate houses. With a warm climate and ample hardwoods, relatively cheap timber houses have been the most common type in Queensland. Older houses often were erected on timber piles to give room below the house for laundry, automobiles, etc., and to keep houses airy and free from termites. The trend in recent years is to build low-set houses, as in other states, and brick or brick veneer is now very common.

CULTURAL LIFE

A feature of recent cultural development in Queensland is the greater opportunity now enjoyed by people in the nonmetropolitan areas. A network of radio stations, both governmental and commercial, extends over the whole state, except in the most remote areas, which must depend on shortwave transmissions of national programs. Television also covers most of the state. Several companies sponsored by the Arts Council make continuous tours of music, ballet, and theatre into the provincial cities and inland country centres and schools, covering a far greater rural area than is done in any other state. Provincial newspapers are strong in Queensland, and the larger daily papers carry international news.

Local participation in press, radio, television, theatre, and vocal, orchestral, and band music is growing in provincial and rural Queensland, as is the time spent on the arts in school courses. At a higher level, the new James Cook University at Townsville offers arts courses at both undergraduate and graduate levels.

Thus, the earlier isolation and lack of cultural opportunities of rural dwellers in Queensland is, to an extent, being overcome. The state government has established a Cultural Activities section in the Department of Education to advise and coordinate cultural work and to help remote areas.

Brisbane, with a population 10 times that of any provincial city, is naturally the strongest cultural centre. It has a history of good performance in orchestral music, and the Queensland Symphony Orchestra, though small, is accorded high rank. There is an excellent junior orchestra as well. Vocal music also has been strong, especially in the city of Ipswich, where many Welsh coal miners settled. In theatre, Brisbane is notable for its number of little theatres, run for many years by competent amateurs and now receiving financial aid; some groups have become professional companies. Several new theatres have been built recently, and playwriting has shown strength in Queensland. Brisbane also has many privately owned art galleries, in addition to the state art gallery and museum, and interest in the visual arts has been increased by leading Australian painters working in Queensland at various times. On the more popular level, annual competitions of fence painting are held in Brisbane and elsewhere, which attract many entries, and there are competitions for works of sculpture. The state Public Library, with its associated Oxley Memorial Library of specifically Queensland items, is in Brisbane. Libraries are maintained by the majority of Queensland's local governments, some having several branches.

Popular culture shows little that is distinctively Queensland in origin. One widely enjoyed form of popular culture, however, is the appreciation of the many miles of unspoiled landscapes and shorelines within the state. This is reflected back in the popularity of separate houses that offer outdoor living amid trees and gardens. (H.W.H.)

History

EARLY EXPLORATION AND SETTLEMENT

During the period of initial European exploration of Australia and the region of present-day Queensland, it has been estimated that more than 200 Aboriginal tribes inhabited the area, totalling about 94,000 persons. The Aboriginal tribes were either fishing people, living along the coast where food was plentiful, or mountain people, occupying the central and western areas where survival was more difficult.

The Talgai skull, discovered in 1884 and estimated to be about 10,000 to 25,000 years old, exists to prove the immemorial stirrings of human life on Queensland soil. But although by 1606 both Dutch explorers and the Spanish explorer Luis de Torres had found Cape York Peninsula, it was 1770 before Capt. James Cook discovered the east coast of Australia and noted signs that a great river might empty into Moreton Bay. Cook named and charted many capes, bays, and islands along the coast, landing on the shore of what is now Queensland nine times. Matthew Flinders explored and charted Moreton Bay, refuting Cook's claim that a great river might enter into the bay. In 1823 John Oxley, aided by castaways, discovered the Brisbane River and chose its vicinity for a new and stricter penal settlement remote from Sydney. The following year, Lieut. Henry Miller, accompanied by Allan Cunningham, set out with 30 convicts and their guards to establish a penal settlement on the site recommended by Oxley. After six months at the site, the settlement was abandoned and reestablished at present-day Brisbane in February 1825. In the next few years exploration of the region continued as Capt. Patrick Logan and Edmund Lockyer explored the hinterland of the penal settlement, discovering coal and limestone deposits in the process. In 1827 Cunningham discovered the Darling Downs.

Cook's explorations

THE PENAL SETTLEMENT

Following the establishment of the penal settlement at Brisbane, more settlements were founded at Ipswich and Stradbroke Island. Accounts of life in the penal settlements report harsh treatment of the convicts, particularly those confined to chain-gang duty. Colonization of Moreton Bay region was strictly forbidden because the more dangerous convicts were housed in this region. The number of convicts varied from the initial 30 to more than 1,100 (including 30 females) in 1833. In 1840 the penal settlement was abolished, at which time the convict population numbered about 100.

FREE SETTLEMENT AND SEPARATION FROM NEW SOUTH WALES

Allan Cunningham's discovery of the Darling Downs pointed the way to more flourishing settlement, which had already begun overland from the south when in 1840 the abolition of the penal colony facilitated healthy development. The early squatters were followed in 1842 by the first free settlers, and sales of land took place. The main hindrances then became the remoteness of Moreton Bay district, the lack of understanding of the region in Sydney, and the consequent small number of settlers—no more

than 2,000 in the mid-1840s. It was 1840 before Patrick Leslie ventured on the downs at Warwick with sheep, and the first crop farmers, apart from the Moravian missionaries at Nundah in the penal days, were brought by J.D. Lang in the "Fortitude" in 1849. It was therefore mid-century before emigrant ships from Britain sailed directly to Brisbane. Nevertheless, political separation from New South Wales was achieved in 1859 and the colony of Queensland proclaimed, when the population numbered only 23,520 and before any industry had established itself, though gold had been discovered in 1858.

Queensland gold rush

Coal was found at Ipswich as early as 1827 and the first mine opened in 1846, but it was the gold rush that really inaugurated Queensland's mining industry. From 1858 to 1873 the east and north of the state were invaded and opened up by diggers at Canoona, Peak Downs, Gympie, Ravenswood, Charters Towers, and Palmer River. Within a few months isolated spots in the bush became townships of 10,000 people. There some of the beginnings of the "White Australia" policy (1877) may be seen in the discrimination by act of Parliament against Chinese and other Asian speculators. By 1867 the population of the state had grown to 100,000. Even where the gold failed, as at Canoona, a town such as Rockhampton could rise from the ruins. In 1882 came the discovery of a mountain of gold and copper, Mt. Morgan, near Rockhampton.

In the 1860s westward migration was very rapid wherever there was water, and the movement was encouraged in 1869 by an act of Parliament which granted 21-year leases to those who had 25 sheep or five cattle to the square mile. Many of the early pastoralists failed. Wool had to be transported in drays by long trains of bullocks and markets were far away and uncertain. But by 1873 Victorian prosperity (in gold) began to be invested in Queensland. The first railways, from Ipswich to Dalby and Warwick, were operating by 1870. The population of Queensland was more than 200,000 in 1880, with the largest cities being Brisbane, Townsville, and Rockhampton. Pastoral industry spread across the Darling Downs in the 1860s and 1870s. By 1880 about 3,000,000 cattle and 7,000,000 sheep were in the colony. Sugar and cotton production, established in the 1860s, increased. In the 1880s attempts were made, not always successfully, to end the bitter struggle between grazier (livestock rancher) and squatter by a succession of acts of Parliament. Artesian boring was introduced in 1881 and was energetically developed after the bad years of drought and depression between 1890 and 1902, which wiped out flocks and herds. In addition,

The cattle tick

after 1894, cattle suffered the onslaught of the cattle tick from northern Queensland, and an industry that had been stimulated by the advent of refrigeration and that had made its mark on Australian life in the tradition of droving over huge distances saw its cattle population halved by the 20th century.

There was similar strife between the early crop farmer and the grazier who had come before him. Hewing out his farm by clearing the bush, attempting crops where none had ever been cultivated, the farmer won the contemptuous name of "cockatoo farmer." But, supporting himself, he advanced from maize (corn) and pumpkins to lucerne and sorghum; his livestock increased; and the plow, reaper, and farmhouse replaced the hoe, scythe and sickle, and bark hut. By 1900 the government had recovered much pastoral land for the arable farmer, and by the mid-20th century land that had originally belonged to graziers had become renowned for wheat, sugarcane, and bacon.

Dairying did not flourish until the 20th century, when there were sufficiently populous towns to give it encouragement. Until 1888 Queensland imported much of its butter from the south, but by 1895 it had become self-sufficient in dairy products.

Between 1860 and 1904 Queensland was involved in the dispute over the employment of South Sea islanders (Kanakas) as cheap labour. The recruiting of Kanakas (often against their will) began with the introduction of cotton cultivation during the American Civil War. After cotton failed at the end of the war, recruiting continued as the sugarcane industry grew. The replacement of the hoe by the plow increased production and decreased the need for Kanaka labour; by a 1904 act the new federal government ended the importing of labour and agreed to support sugar financially, when necessary, in the interests of the "White Australia" policy.

The Kanaka problem influenced the separatist movement in northern Queensland (where both sugarcane and the goldfields were), but whereas the planters wanted separation the miners did not, fearing that the planters would impose a policy of continued importation of labour.

Toward the end of the 19th century the increase in population, the advance of social legislation in Europe, the socialist idealism of William Lane, and the depression of the 1890s encouraged the growth of trade unionism and led to the emergence of a Labor Party with well-defined policies. The most notable event in Queensland at this period was the shearers' strike in 1891. It was broken, but subsequent legislation providing for shorter working hours, improved conditions of work, and the encouragement of the smaller settler showed clearly the trend of the times. In 1899 Queensland had Australia's first Labor government, though a minority one, but Labor's position rapidly became so strong that after World War I it was almost continuously in office.

Emergence of the Labor Party

FEDERATION AND THE STATE OF QUEENSLAND

On January 1, 1901, the Commonwealth of Australia was proclaimed, creating the state of Queensland. The census of the same year recorded 498,129 inhabitants, excluding Aborigines. In 1904 women gained the right to vote, and in 1914 compulsory voting was enacted first in Queensland, followed by the other states.

In 1922, at the instigation of the ruling Labor government, the Legislative Council (upper house of the bicameral parliament) was abolished, leaving the Legislative Assembly. Three years later Greater Brisbane was established as the largest municipal council in Australia. From the beginning, the Brisbane council was elected by a complete adult electorate.

The Labor government was replaced by the Country Party as the ruling government in 1929. The new government came to power amid difficult times of rising employment, falling incomes, and the social distress of the Depression years. During its tenure of government, the Country Party abolished state trading and established the Bureau of Economics. The first woman was elected to the Queensland Parliament during this period of Country rule. The Country Party was unable to stem the tide of the Depression and was replaced by the Labor Party, who attempted to stimulate the economy through large capital improvement projects. Several substantial projects were undertaken, such as the Story Bridge, Stanley River Bridge, and construction of the University of Queensland at St. Lucia. The worst of the Depression was ending by 1934, and many social services, suspended during the bad economic times, were reinstituted. In 1946 free hospital service was introduced and large-scale irrigation projects were begun by the government, which included the Burdekin and Tully hydroelectric scheme. In 1957 the Labor government was ousted by a Country–Liberal (now called National–Liberal) coalition.

The Country–Liberal coalition

In the 20th century improved farming methods, irrigation, insecticides, communications, and new markets at home and in Japan greatly strengthened primary production. In 1923 vast silver–lead–zinc deposits were found at Mt. Isa, and in 1969 further vast deposits were found. Uranium was discovered and mined at Mary Kathleen from 1950 to 1963, and bauxite was found in great quantity at Weipa. In the 1960s the Mooni oil field, the natural-gas field at Roma (piped to Brisbane), and the huge mining port of Gladstone were developed.

In 1901, Queensland was represented by six senators in the national Parliament. This number was increased to 10 in 1948. The state's representation in the House of Representatives remains proportional to Queensland's population.

In the 1980 elections held in Queensland, the National–Liberal coalition won control of the state Parliament, as it has done consistently since 1959.

(R.H.G./Ed.)

SOUTH AUSTRALIA

One of the six federated states of the Commonwealth of Australia, South Australia lies on the south central coast of the nation. Western Australia lies to the west, the Northern Territory to the north, Queensland to the north and east, and New South Wales and Victoria to the east. The great majority of its people cling to the seacoasts, most of them living in Adelaide, the capital. The rainfall of two-thirds of the vast interior is insufficient to support significant human or animal populations. In the remainder there are scattered towns, mostly small, linked with the coasts by rail, road, or air. The aridity of the interior is only slightly compensated for by the presence, in the northeast, of a portion of the Great Artesian Basin, the world's largest region of natural springs. In two reserves in the northwest and southwest live remnants of the Aboriginal peoples of Australia.

Physical and human geography

THE LAND

Most of South Australia is flat, with large tracts of sand and of the stony plains known as gibber desert. Over 80 percent of the land is less than 1,000 feet (300 metres) above sea level.

Relief. The state is divided into five major regions. The western shield, a vast barren region that also takes in much of Western Australia, is underlain primarily by Precambrian rock, over 570,000,000 years old. It swoops down from the north and west to engulf the entire Eyre Peninsula and much of the Yorke Peninsula on the coast. Bordering it in the northeast and reaching nearly to the coast, great layers of sandstone, lying 4,000 to 5,000 feet underground atop the Precambrian Shield, act as the water bed for the Great Artesian Basin. The several ranges of the Precambrian highland chain stretch mainly north from the Adelaide region and east into New South Wales.

The Eucla and Murray basins

In the extreme southwest and southeast, the Eucla and Murray basins reach into Western Australia and Victoria, respectively. Most of the land in these areas is of Mesozoic origin, between 65,000,000 and 225,000,000 years old. Near the town of Mount Gambier, close to the border of Victoria, a few lake-filled craters have become tourist attractions, but the state is geologically inactive in terms of volcanic activity or surface glaciation.

The coastline is indented by two large, navigable gulfs separated by Yorke Peninsula, which is about 30 miles wide. Gulf St. Vincent, the more easterly of the two, is about 100 miles in length; Adelaide lies on its eastern shore. Spencer Gulf, twice as long, has near its northern limits the industrial towns of Port Pirie, Port Augusta, and Whyalla. The only large stream is the Murray River, which enters South Australia after forming the New South Wales–Victoria boundary. The rivers shown on maps as lying in the north rarely contain any water. To the south and east of Lake Eyre is a striking group of "mound springs," mineral deposits looking much like miniature volcanic cones and rising as much as 130 feet in the air. They have been formed by subterranean artesian action, which forces the water to the surface. The highest point in the state, Mt. Woodroffe, rises to 4,723 feet in the far northwest.

Climate. More than 80 percent of the state receives less than 10 inches (250 millimetres) of rain annually, all of it sometimes falling within a few months to be followed by months of almost complete drought. Only 0.3 percent gets 30 inches or more. The driest known area on the continent occurs around Lake Eyre. A few areas in the southeast receive more than 35 inches, but the average is between 20 and 30 inches. During the winter months of June, July, and August, most of the state has mean temperatures between 50° and 60° F (10° and 16° C), whereas summer figures stand at 65° F (18° C) in the south and between 80° and 85° F (27° and 29° C) in the north.

Land usage. Land suitable for farming other than extensive grazing is confined to the south, excluding the far west. For the analysis of rural statistics, the commonwealth statistician divides the state into nine divisions. Seven of these represent about one-eighth of the state's area. About two-thirds of the state was thus unused for either agriculture or pasture. Some of this area is set aside for Aboriginal reserves.

Plant and animal life. No vegetation grows at all in the stretches of sand and gibber desert, and only scrub grasses and saltbush are hardy enough to survive in the dry regions of the north and west. Eucalyptus and native pines are scattered in the Flinders Ranges; the Mt. Lofty Ranges, near Adelaide, are more heavily wooded, mainly with eucalyptus. In the southeast, more than 170,000 acres of man-made forest, 90 percent of it in softwood, have been created since the late 19th century to assure forest preserves. These have softened considerably the landscape of the entire region.

Man-made forests

Most Australian species are represented, and some species are unique to the state. However, in comparison with the eastern states, South Australia has a sparse population of native fauna, partly because of the climate and soil but also because of the use of better land for farming. Some species have become extinct. Kangaroos and wallabies are common, as are wombats and bush rats. There are many kinds of birds, lizards, and snakes. The coastal sea has a wide variety of fish, but few fish are to be found in inland waters except in the Murray River.

Drainage. Maintaining and transporting supplies of fresh water for human, animal, and agricultural use is a constant concern in nearly all of Australia, the world's most arid continent. A few small streams rising in the Mt. Lofty Ranges near Adelaide have been dammed, but supplementary supplies must be pumped over these mountains from the Murray River. Another pipeline to the industrial town of Whyalla stretches for 223 miles from the Murray, whose water is also used for irrigation. The need to ensure a regular flow of high-quality water in the Murray has led to disputes between South Australia and its partners in the River Murray Agreement—the commonwealth, New South Wales, and Victoria—on the location of additional dams. In the interior, supplies are even scarcer, and private water catchment devices are still widely used.

Settlement patterns. The first surveyor general of the South Australia colony, Col. William Light, delimited the colony into nine counties in 1842. In 1846 the lands of counties were divided into hundreds, based on the system used in many English counties. The boundaries for counties and hundreds were based on natural features, but as more counties and hundreds were surveyed natural features became more difficult to use. In most cases, boundaries were based on parallels of latitude and longitude. The last of the 49 counties of South Australia was delimited in 1933. The counties of South Australia average 293 square miles in size.

Adelaide was surveyed and lands were allotted or sold in March 1837. The colony had grown to about 18,000 by 1844, with about 10,000 living in or near Adelaide. By that time the colony was producing an excess of wheat, and copper had been discovered at Kapunda and Burra. The colony began to prosper and heavy immigration followed the copper discoveries, continuing until the discovery of gold in Victoria in 1851. The concentration of people in Adelaide created a demand for community services. Water was laid in 1861 and gas in 1863. Adelaide became more attractive to settlers because of the added amenities and greater employment opportunities.

Much of the colony had not been exploited. Very few settlers lived in the Murray Valley or in the northern district. After 1961 the pattern of settlement in South Australia began to change. The opening of the northern wheat lands, the drainage of the southeastern swamps, the increase of farmers on the two peninsulas and in the Murray Mallee, and irrigation of the Murray Valley, altered the distribution of settlement. Port Lincoln and Mount Gambier be-

came more important distribution centres, and improved roads and railways made the movement of persons and goods easier. The settlement continued to concentrate in Adelaide and the surrounding cities and towns.

THE PEOPLE

Slightly more than three-fourths of South Australia's population was born in Australia, and virtually all others were born in the British Isles or continental Europe. About 10,000 persons report having 50 percent or more Aboriginal blood, and few of these had been assimilated into the processes of modern Western society prevalent throughout the state. The Anglican Church has the largest number of adherents, followed by the Methodist and Roman Catholic churches, with approximately equal numbers.

Between 1945 and 1965 the state's population grew at a faster rate than that of Australia as a whole, mainly because of a disproportionately large influx of immigrants. After 1965, however, the rate of growth fell sharply and the immigration rate fell below the Australian average because of a deteriorating employment picture. The birth and death rates have declined since World War II.

Population distribution

South Australia's population distribution is extremely skewed, plunging from Adelaide's concentration of nearly 1,000,000 inhabitants to the less than 50,000 in Whyalla, the next largest urban centre. The only other centres with 10,000 or more are Mount Gambier, Port Pirie, and Port Augusta. The proportion of rural population, as in the entire country, has declined since World War II.

THE ECONOMY

South Australia's rate of economic growth exceeded that of the nation as a whole during the mid-20th century, and the labour market was strong, though a lower rate of growth in demand for motor vehicles and consumer durables, in later years, slowed the pace. Although the mean income is below the national average, so are living costs, putting the standard of living on a par with the rest of Australia. There were more automobiles and television sets per capita than in other states. In Adelaide the cost of land and building is below that of other Australian capitals.

Resources. More than one-quarter of the state's work force is engaged in primary production, including mining and quarrying. The state is not rich in minerals, but the Middleback Ranges produce the bulk of Australia's iron ore. Some of this is processed at the Whyalla blast furnaces, but most is carried by water to Newcastle and Wollongong in New South Wales. The state is a major world source of opals. Most of the nation's salt and gypsum come from South Australia, but the open-pit coal mines in the Flinders Ranges are not extensive. A pipeline to Adelaide from the large natural gas fields discovered in the north was completed in 1969.

Manufacturing. Motor-vehicle manufacture is a large component of secondary industry, which employs about one-tenth of the state's manufacturing labour force. General Motors-Holden Pty. Ltd., one of the nation's largest manufacturers, divides its operations between Melbourne and Adelaide. Consumer durables, metal pipes and tubes, chemicals, and textiles are produced in or near Adelaide. A soda-ash plant at Osborne, near Port Adelaide, is the major producer in the Australian alkali industry. Whyalla, in addition to its iron and steel industry, has Australia's largest shipyards. Sawmills and factories producing particle board and paper are located near the softwood plantations in the southeast. Port Augusta has a large power station using the low-grade coal transported by rail from the Flinders Ranges. Smelters at Port Pirie treat ores sent by rail from Broken Hill in New South Wales.

Agriculture. South Australia is a major producer of wheat and barley for export abroad, but it excels in its vineyards. Growing about a third of Australia's grapes, it produces about two-thirds of the wine and most of the brandy for the tables of the commonwealth. Tuna and crayfish are pulled from the coastal waters for export to other states and abroad, and its millions of sheep produce wool, meat, and hides. Dairy production is largely for local consumption, but a considerable quantity of vegetables is sent to other states.

Role of the vineyards

Transportation. Transport services within the state centre upon Adelaide, and there has been little planning to coordinate the transportation systems as a whole with the state's economic development in terms of geography. The railways are almost entirely owned by the state and commonwealth governments, most of the commonwealth operations being related to the transcontinental line linking Sydney and Perth. A direct service operates between Adelaide and Melbourne, the Victorian capital, and a line of five-foot three-inch gauge from Adelaide meets the commonwealth railway at Port Pirie, the latter a standard-gauge line. Operating losses have caused the closure of many rural lines and a diversion of traffic to road transport. Many of the roads are hard-surfaced or bitumin-sealed, but most are unsealed or of crushed stone. Port Adelaide and Whyalla are major harbours for trade with other states and foreign nations; Port Stanvac is an oil-discharging port near Adelaide. Direct air services operate from Adelaide to Sydney, Melbourne, and Perth and to several towns within the state.

ADMINISTRATIVE AND SOCIAL CONDITIONS

Government. The division of power between commonwealth and state parliaments results in commonwealth predominance in defense, foreign affairs, economic policy, welfare payments, and shipping and aviation. The main matters under state authority are education, hospitals, housing, police and prisons, and railways and roads. South Australia is represented in the House of Representatives and the Senate in the Commonwealth Parliament. The state parliament comprises a House of Assembly elected by universal and compulsory suffrage, and a Legislative Council elected by a restricted and voluntary franchise. Legislation must be passed by both houses.

The governor, the official representative of the crown, accepts the advice of ministers who, by constitutional convention, are collectively responsible to the House of Assembly. The principal minister, the premier, is usually the leader of the majority party in that house. Local councils are elected by property owners and tenants in local-government areas.

Political parties

The two main political parties are the Australian Labor Party (ALP) and the Liberal–National Country Party (LCP). The ALP usually receives a majority of the votes, but because the distribution of representatives favours rural areas the party is not always assured of a majority in the House.

The governor appoints the members of the state's Supreme Court, from which appeals lie to the High Court of Australia and the Privy Council. Among the lower courts are those dealing with such specific matters as licenses for the sale of liquor. The Industrial Commission is responsible for labour arbitration, but most workers are covered by decisions of the Commonwealth Conciliation and Arbitration Commission.

Education. Although education is compulsory from six through 15 years of age, most children enter at five, and the rates of retention in secondary education are rising. All state schools are tuition-free and administered by the Department of Education with a cabinet-level minister. The secondary system comprises high schools offering general academic training, vocationally oriented schools, and the "area schools" that serve rural populations in primary and secondary grades. Private schools enroll a small percent of primary pupils and secondary pupils. Apprenticeships are controlled by a state commission.

The institutions of higher learning in the state include the University of Adelaide, founded in 1876; the Flinders University of South Australia, which opened on the outskirts of the capital in 1966; and the South Australian Institute of Technology, dating from 1889. There are, in addition, an agricultural college, a school of art, and several teacher-training colleges.

Health and welfare. South Australia's very high health standards are reflected in the figures for infant mortality, which are among the lowest in the world (although the Aboriginal infant mortality rate is high). Hundreds of public and private hospitals are in service throughout the state. There are few slums, and a large proportion of housing is built by the South Australian Housing Trust.

Department of Aboriginal Affairs

The Aborigines and the state. In 1973 the Commonwealth assumed responsibility for Aboriginal affairs. The Department of Aboriginal Affairs administers a variety of educational and social welfare programs throughout the state, and the Aboriginal Land Fund Commission assists Aboriginal communities in the purchase of land and homes.

CULTURAL LIFE

The arts. South Australia's contributions to the arts are few, but it provides facilities for its citizens to participate in them. The Art Gallery of South Australia has an interesting collection of Australian art. A professional symphony orchestra is maintained by the Australian Broadcasting Commission and several theatres provide amateur and professional drama. Visiting artists present frequent concerts and recitals, and every second year a Festival of Arts is held, modelled on the Edinburgh International Festival of Music and Drama in Scotland. In addition to the art school, there is a conservatorium of music at the University of Adelaide, and Flinders University offers courses in theatre, film, and television.

Sports. The main spectator sport is Australian rules football, played during the winter from April to September. Cricket takes over during the summer. Year-round horse racing is popular, and in the summer the Adelaide showgrounds are given over to night trotting races.

(K.J.H.)

History

EXPLORATION

The South Australia coast was first discovered and explored in 1627, when the Dutch ship "Gulden Zeepaard" sailed the coastline from Cape Leeuwin to the islands of St. Francis, under the command of Francois Thyssen. Since South Australia did not look promising for trade, which was the prime interest of the Dutch, the region remained unexplored until the late 18th century.

The Frenchman Antoine Bruni d'Entrecasteaux, sailing along the coastline in search of another French explorer, sighted land south of Cape Leeuwin in 1792. He sailed from this point to the head of the Great Australian Bight and then south to Tasmania.

The first Englishman to sail and explore the coast of South Australia was Lieut. James Grant, who sighted the extreme easterly shores of the region in about 1800. During his voyages around the Australian continent, Matthew Flinders explored the coast of South Australia between January and April of 1802. Flinders named many places along the coast after places in his home county of Lincolnshire. While exploring the area of Encounter Bay, Flinders met a French explorer, Capt. Nicholas Baudin, in his ship the "Le Geographe." Baudin was in the process of charting the coastline of South Australia.

From 1804 to the first settlement in 1836, exploration continued in South Australia by several persons, such as captains Dillon (1815–16), Goold (1827–28), Hart (1831–33), and Jones (1833–34). The most important of these explorations were journeys of Capt. Charles Sturt, who explored the Murray River as it flowed from its source to the mouth in South Australia during 1830. Sturt's glowing reports of the region were influential in the decision to establish a colony. In fact, he recommended the site of present-day Adelaide as the best place for the capital of the colony.

Murray River explorations

THE FIRST SETTLEMENT

South Australia was officially settled as a new British province on December 28, 1836. Proposals for settlement had not been made until 1830, when plans made by the National Colonization Society were placed before the Colonial Office. Later the South Australian Land Company and the South Australian Association also put forward plans. Each proposed to apply the Wakefieldian theory of systematic colonization in a new territory where settlers might enjoy civil and religious liberties, but each in turn was rejected by the government as "too republican." These frustrating negotiations ended in 1834 when an appeal to Parliament brought the South Australian Act, which guaranteed a fixed price for land and ensured that no convicts would be sent to the colony. To make certain that no cost would be borne by the mother country, the Colonial Office was to have charge of all government affairs except land sales, emigration, and fund-raising, which were to be controlled by an independent board of colonization commissioners.

When settlement began, this division of authority led to bitter factional strife, exacerbated by the ill-considered rules of the commissioners in London. A site for the capital, Adelaide, was chosen and soon divided into quickly sold lots, but the survey of the country sections was unduly delayed, and very little food was produced until 1840. Some settlers thrived on speculation in town land and imported supplies, while thousands of poor immigrant families, heedlessly sent out by the commissioners, remained idle. As the commission had ordained a free market for labour, the first governor, Capt. John Hindmarsh, was powerless; the second, Col. George Gawler, exceeded his instructions by employing immigrants on costly public works. Before his unauthorized bills reached London, the commission was already bankrupt, its funds frittered away in "puffing" the new province. However, a parliamentary committee that examined South Australian affairs in 1841 blamed Gawler for the province's disastrous financial state. During the next year, Parliament paid most of the debts, stopped emigration, removed the colonization commissioners, and placed the province wholly under Colonial Office control.

PROGRESS TOWARD SELF-GOVERNMENT

The third governor, Capt. George Grey, was pledged to economize. He arrived to find land surveyed and settlers at work on their country sections. By 1844 the colony was paying its way. Although wool was the chief export, farmers were growing more wheat than the population of 17,000 could consume. In 1845 rich copper discoveries brought a spectacular mining boom that made many settlers independent of the absentee investors who had hitherto financed and directed them from England. Local politics thus gained new interest and became more controversial. The Colonial Office was defied over mining royalties and road taxes, and when it insisted that state aid be offered to all churches, only four denominations accepted it. The vigorous protests of Nonconformists brought an end to government grants in 1851, making South Australia the first part of the British Empire completely to separate church and state. At the same time, four years of heavy immigration increased the population to 67,000 and qualified South Australia for the grant of a partially elected legislative council with the right of drafting a constitution of self-government.

Discovery of copper

The gold rush to Victoria had drawn thousands of South Australians when the first constitution was prepared in 1853. To returning diggers and irate citizens alike the constitution was "a caricature of self-government." Their petitions led to a second constitution, which provided two elective chambers: a House of Assembly with triennial elections, adult male suffrage, secret ballot, and no property qualifications; and a Legislative Council with a modest property franchise to protect landed interests. To the Colonial Office this was "the only thorough Benthamite constitution in the empire," but royal assent was given, and South Australia's first Parliament with responsible government met in 1857.

Political independence seemed to exhaust the colonists' ideas but not their powers of resistance. During the first 30 years of self-government there were 34 changes of ministry and very little experimental legislation.

One important early exception was the Real Property Act (1858), subsequently copied by more than 50 countries and sometimes called the Torrens Title System after its sponsor, Sir Robert Torrens. To make conveyancing simple, quick, and cheap, the act replaced the unwieldy system of title deeds by one of certificates guaranteed by the state. In the courts the act was persistently challenged by Judge Benjamin Boothby, who claimed that it and other colonial enactments, including the constitution, were repugnant to English law. Boothby's stubborn stand soon

led to his removal from the bench. To meet his objections the Imperial Parliament passed the Colonial Laws Validity Act (1865), which, by defining "repugnancy," became the legal keystone of the empire until the Statute of Westminster of 1931.

ECONOMIC EXPANSION

By making titles secure the Real Property Act also simplified mortgaging, especially for wheat growers, whose small freeholds were beginning to force livestock owners farther afield in search of leases. Because pastoral expansion seemed to be checked by a barrier of salt lakes, exploration was stimulated. Slow-moving government parties found gaps between the lakes, and more mobile private explorers passed through them to find vast new pastures. The most notable of these pioneers was John McDouall Stuart, who finally crossed the continent in 1862, paving the way for South Australia's acquisition of the Northern Territory (1863) and for the overland telegraph, which in 1872 linked Adelaide with the ocean cable terminus at Darwin. Other explorers found traces of minerals. In 1860 the discovery of rich copper lodes near Moonta aroused hopes of general prosperity and encouraged many pastoralists to move their stock beyond the salt lakes. Disaster followed, for the country was poorly watered; in the drought of 1864–65, thousands of sheep and cattle perished. As the drought did not extend very far to the south, Parliament was easily persuaded that all the southern areas were safe for agriculture.

1864 drought

During the gold rushes South Australian grain had fed the diggers in Victoria and followed them into New South Wales and Queensland. From 1853 steamboats carried flour far up the Murray and Darling rivers and returned with cargoes valuable enough to justify a railway and new port at Encounter Bay. In the mid-1860s an even greater market for grain was found in England. To the Parliament in Adelaide the prospect seemed brighter for wheat than for wool. In 1869 the Strangways Act introduced "credit land sales," by which the land hungry might acquire freeholds on expired pastoral leases on easy terms. A decade of prosperity followed. Immigration, heavily subsidized by the state, provided agricultural recruits and labour for the government's new railways. Each year brought new records in cultivation; yet South Australia, with one family in three on its farms, depended too much on agriculture for its economic well-being.

Hope and inexperience carried the farms into unsafe districts, and old overworked districts became "wheat-sick." Though prosperity lasted until 1884, two droughts heralded collapse. Bankruptcies mounted as prices fell, banks failed, and copper mines showed signs of exhaustion. Average yields of 15 bushels per acre dropped to less than two bushels. Long before the economic depression ended, South Australia lost its lead in wheat growing to Victoria and New South Wales and fell behind Queensland in population. Numerous families migrated to eastern states, and many more were attracted to the towns even though urban industry had been neglected and employment was insecure. Demand for liberal reform stirred Parliament into action. Land in safe districts was offered on easy terms to homesteaders and workingmen. Promises of compensation encouraged pastoralists to improve their ranges and multiply their flocks. Water conservation projects were started; southeastern swamplands were drained; and irrigation and village settlements were begun on the Murray River. The opening of Roseworthy Agricultural College introduced farmers to the use of superphosphates, while scientific techniques began to produce better strains of wheat, freedom from disease, systematic fallowing, and more varied crops. Assisted immigration stopped, but railway building continued. To avoid further borrowing in London—the public debt was already close to £50 per head—in 1884 the government introduced for the first time in Australia direct taxation on incomes and land. When this proved insufficient, tariffs were sharply increased to curtail imports and foster urban industry. The opening of Broken Hill in New South Wales helped South Australia by providing employment for its miners, freight for its railways, and smelting works at Port Pirie, but these gains were at first offset by heavy migration to Western Australian goldfields. The high levels of production in 1884 were not regained again until 1906. Although the population had passed 300,000 by the mid-1880s, it did not reach 400,000 until 1910.

Roseworthy Agricultural College

POLITICAL PARTIES

The most marked advances of the depression years were in politics. Growing discontent brought the eight-hour working day (1884), payment of members of Parliament (1887), workmen's compensation (1889), industrial arbitration and votes for women (1894), and the first Australian use of a referendum (1895). Primary education, which had become compulsory and secular in 1875, was made free in 1891. The depression years also hastened South Australia's progress toward federation. The tentative beginnings of a Labor Party during the 1890s forced its opponents into closer alliances and more stable ministries. Leading politicians, perplexed by inland tariff barriers, problems at Murray River irrigation, and annual deficits in the Northern Territory, were eager for intercolonial cooperation. In the federal conventions South Australians played a leading part. Later, when the ablest leaders were drawn into the federal Parliament, local politics became more subdued. The Labor Party grew in strength and in 1905 took office in a coalition government. Its main concern was the reform of unequal electoral constituencies: Adelaide, with half the population, had only one-third of the seats in Parliament. Reform was resisted by the Legislative Council, whose abolition, though advocated by Labor, was not popular with electors. Nevertheless, the Labor Party did form occasional governments, most of them ending in dissension.

The years following World War I brought the organization of the Farmers and Settlers' League, the Progressive Country Party, the Nationalists, and the Liberal Union. From these emerged the Liberal Federation and Country Party Alliance, which by 1933 forced Labor into opposition except in 1965–68.

ECONOMIC DEPRESSION

During the first decade of federation South Australia seemed to prosper. Its share of federal customs revenue paid the annual interest on the public debt, while bumper harvests and rising prices renewed enthusiasm for agriculture. Aided by new railways and superphosphates, farmers opened fresh wheat belts on Eyre Peninsula and in the southeast. Assisted immigration was resumed. Valuable iron-ore deposits at Iron Knob were linked by rail with the new port of Whyalla. Port Adelaide was enlarged and an outer harbour added for mail steamers. Electric trams and suburban railways revived land speculation, while government high schools and free secondary education attracted families to the city.

The tide turned in 1910. Although the costly Northern Territory passed under commonwealth control, a new agreement for distributing federal revenue lowered South Australia's income. From 1914 to 1926 the government had to budget for eight deficits. State taxation quadrupled and government borrowing increased the public debt to £A136 per head (compared with Victoria's £A88). Large numbers of workers and some entire industries migrated each year to the eastern states where bigger markets, cheaper power, and lower taxes offered better opportunities. After 1928, four dry years coincided with worldwide depression to reduce the state to a worse plight than any of its neighbours. In the winter of 1931 more than 70,000 unemployed out of a population of 575,000 were dependent on government relief. Thousands of the unemployed sought work in the country, enabling a record crop of 8,000,000 acres (3,237,600 hectares) to be sown, although depressed prices gave farmers an average of less than 20 shillings per acre for their harvest.

Public debt

Appeals for federal aid led to the establishment of a Commonwealth Grants Commission and brought substantial assistance earmarked for developmental works. However, most politicians still looked to the land for salvation. To the auditor-general, W. Wainwright, this dependence on rural output seemed disastrous. He advocated more sec-

ondary industry; by judicious government guarantees to suitable enterprises the state would attract overseas investment and unlock its own private capital tied up in land. The threatened departure of a large motor-body-building industry forced the government to support Wainwright's plan. Certain wharf dues were abolished, and company tax was halved. The first large project was the planting of pine forests in the southeast, for the state had always depended on imported building timber. Guarantees were made to the cement and chemical industries at Port Adelaide. Elsewhere, industries were encouraged to make equipment for the expansion of roads, water supply, sewerage, and building.

To supplement private efforts the government built large reservoirs for Adelaide, expanded technical education, and created a trust for housing industrial workers.

WORLD WAR II AND AFTER

In 1938 Thomas Playford began his record term of 26 years as premier, and his accession infused new life into industrial policy. Some of his plans had strategic importance and were implemented during World War II. The Broken Hill Proprietary's new blast furnace and shipbuilding yard led to great expansion at Whyalla which had to be supplied with water by pipeline from the distant Murray River. Near Adelaide large munition works built by the commonwealth government were taken after World War II by overseas firms, some for new industries and others for the development of long-range weapons. The need for an isolated testing site for the weapons brought the town of Woomera into existence, its water piped from the Murray. In 1946 Playford's government took over the private Adelaide Electric Supply Company, turning it into an electricity trust that cut prices and greatly expanded consumption. To make the state less dependent on imported fuels, the government opened a brown coal mine at Leigh Creek, and a large supplementary power station soon followed at Port Augusta. The postwar years also brought great rural expansion. Scientific testing of inferior soils revealed the absence of trace elements, which, when restored, gave fertility to much unused scrubland. High returns for wool and barley freed farmers from the vagaries of wheat markets, while road transport, diesel rail engines, and bulk handling of grain helped to produce new records in cultivation, pasture improvement, and sheep numbers.

Postwar rural expansion

The most spectacular expansion, however, was in Adelaide. Its new and growing factories attracted a great influx of labour, which in turn demanded new homes and services. The city itself became independent on water from the Murray. As the Adelaide Plain rapidly filled with buildings, a new satellite town was created nearby at Elizabeth and other new southern suburbs.

In 1973 the South Australian Parliament underwent a fundamental reform as the special voting franchise was abolished and replaced by an 18-years-of-age adult requirement for all voters. Also under this reform, proportional representation was introduced for parliamentary elections. The first election held under these reforms occurred in 1975.

The South Australian Land Commission was established in 1973 to stabilize the price of urban land and stimulate urban development through land acquisition and management. Financial assistance for land acquisition was provided through the Urban and Regional Development Act of 1974. In 1980 the name of the commission was changed to the South Australian Urban Land Trust and its role was altered, transforming it to an urban land bank.

(D.H.Pe./Ed.)

TASMANIA

Location and general character of the state

The island state of the Commonwealth of Australia, Tasmania, lies about 150 miles (240 kilometres) south of the State of Victoria, from which it is separated by the relatively shallow Bass Strait. Physically, Tasmania forms part of the Great Dividing Range; culturally, it is part of Melbourne's metropolitan sphere of influence. The state comprises a main island called Tasmania; Bruny Island, nestling close to the southeast coast near Hobart, the capital; King and Flinders islands in Bass Strait; numerous smaller islands off the coast of the main island; and sub-Antarctic Macquarie Island, some 1,000 miles to the southeast. The main island is roughly heart-shaped, and its latitude and climate are broadly comparable to those of northern California and northwestern Spain. Its 26,383 square miles (68,332 square kilometres), although slightly larger than the area of Sri Lanka, comprise less than 1 percent of Australia.

The state owes its name to the Dutch navigator-explorer Abel Tasman, the first European to discover the island, in 1642, though until 1856 it was known as Van Diemen's Land, after the governor of the East Indies who had sent Tasman on his voyage of exploration. The island of Tasmania contains some of the most spectacular mountain, lake, and coastal scenery in the continent, has much of the country's hydroelectric-power potentiality, and displays a great diversity of natural resources. Throughout much of its history, Tasmania has experienced a net out-migration that has deprived the state of a wealth of talent but has contributed, perhaps disproportionately, to the leadership of the nation. Although insularity renders much of the political, economic, and social life distinctive, proximity to Melbourne and modern air travel make the island less isolated and more progressive than is often assumed in other states.

Physical and human geography

THE LAND

Relief. Tasmania is essentially a mountainous island. In the west, where the highest peak on the island, Mt. Ossa, reaches 5,305 feet (1,617 metres), the landscape comprises several parallel northwest–southeast ridges and valleys; eastward lies a series of plateaus at various altitudes, the highest point being Ben Lomond at 5,160 feet (1,573 metres) in the northeast. But the dominant feature of Tasmanian geography is the glaciated, lake-studded Central Plateau, bounded on the north and east by a 2,000-foot fault scarp and sloping gently southeastward from 3,500 to 2,000 feet (1,100 to 600 metres). Much of the east is made up of a low, dissected plateau averaging about 1,200 feet (365 metres). Extensive plains are confined to the far northwest, the lower South Esk River Valley, and the northeast. The Bass Strait islands represent outliers of the northern coastal platforms. In the southeast, postglacial submergence has produced one of the finest drowned coastlines in the world.

Major river systems

Drainage. There are two major river systems—the Derwent in the southeast and the South Esk in the northeast—many small systems, especially westward flowing to the west coast, and innumerable lakes. The Central Plateau is studded with more than 4,000 lakes in a landscape similar to northern Canada and Finland; almost all, including Great Lake, are shallow. Lake St. Clair, the deepest lake in Australia (over 700 feet [215 metres]), is a piedmont lake similar to the lakes of northern Italy. Several lakes, notably Lake King William, have been created by hydroelectric-power development.

Soils. Most Tasmanian soils are leached, acidic, poorly drained, high in humus, and low in fertility. Least fertile and most extensive are the soils, especially the moor peats, of the west and northeast. Fertile areas occur extensively in the northwest and locally elsewhere, notably in the northeast. Brown earths occupy the drier areas east of the Central Plateau; black earths, the southeast; and alluvial soils, the narrow valley floors to the east. Other fertile soils are those of former swamps in the far northwest and the Bass Strait islands.

Climate. Tasmania, located in the mid-latitude westerly wind belt and dominated by southern maritime air masses, enjoys a moist, equable climate, with mild to warm

summers, mild winters in most settled areas, and rain during all seasons. Yet the occasional incursion, in summer, of tropical continental air masses and, in spring and autumn, of tropical Tasman air masses along the east coast, together with the mountainous surface, results in greater annual rainfall variability, seasonal moisture deficiencies, and temperature changes than are the norm in similar climates elsewhere. Climatic averages thus mask considerable variations. The average annual precipitation exceeds 100 inches (2,500 millimetres) on the western ranges and declines eastward to under 20 inches in some places; along the north coast it exceeds 30 inches in all locations. The seasonal incidence in the north and west is greatest in winter, and in the south and east it is greatest in spring. Summer rainfall may vary markedly from year to year, especially in the drier east. Mean January temperatures are higher in the north and east than elsewhere, reaching 64° F (18° C) at Launceston; mean July temperatures are 46° to 49° F (8° to 9° C) in all coastal stations, declining sharply with altitude. Almost everywhere, the mean daily range of temperature exceeds the mean annual range.

Rainfall and temperature ranges

Plant and animal life. In general, the wettest areas have temperate rain forest, especially the beech or myrtle; areas having 30 to 60 inches of rain annually carry good-quality eucalypt forest; and the drier areas carry poor-quality eucalypt forest or savanna woodland. In certain areas, particularly in the forest of the south and southwest, an almost impenetrable thicket known as horizontal scrub develops. This is caused by the growth of a remarkable small tree called the horizontal (*Anodopetalum biglandulosum*). The slender trunk of the tree falls over under its own weight, and from it branches arise that behave in the same way. On the mountain plateaus are found many plants having subantarctic affinities. These include Tasmania's only deciduous tree or shrub, myrtle beech, and certain cushion plants. Rain forest would be more widespread in the absence of fires. Other vegetation includes the sedgeland along the west coast, the high moorlands, and the coastal heaths of the far northwest, the far northeast, and the Bass Strait islands.

Animal life is virtually absent from the true rain forest but abounds in the extensive eucalypt forest. The avian fauna includes the honey eaters, the black jay, black magpie, the black cockatoo, and various parrots. Among the mammals are the wallaby, brush and ringtail possum, and the marsupial carnivores—the native cat, the tiger cat, the Tasmanian devil, and the rare thylacine. The sedgeland and moorland are distinctive for the wombat, and the coastal heath for the green rosella, the platypus, and the short-nosed echidna.

Mammals

Settlement patterns. The inhospitable terrain of much of Tasmania naturally has had much influence on settlement patterns. The nomadic original Tasmanians have left a few archaeological traces, including geometric designs on exposed rock surfaces and evidence of cremations and corroborees, or ceremonial gatherings. The European settler has left his imprint according to economic activity, whether it be in the mining settlements of the west; the intensive cropping or dairying of the northern coastal belt, or of the southeast lowlands; the intensive beef fattening of the northwestern tip; or the dryland sheep farming of most of the eastern sector. The southwestern segment of the island is scarcely utilized by humans.

Traditional regions. A widely recognized regionalism bears witness to the diversity of the Tasmanian landscape. In the north the distinctive regions are the far northwest around Smithton; the northwest coast extending along the north coast west of Port Sorell; the Tamar Valley; the northeast around Scottsdale and Lilydale; and the far northeast from Ringarooma eastward. In central Tasmania the regions are the west coast, usually comprising only the Queenstown and Rosebery area; the Central Plateau and the Derwent Valley; the midlands, comprising the area east of the Central Plateau between Hobart and Launceston; and the east coast. Southern regions are the southwest, the Huon Valley, and the southeast. Numerous subdivisions, such as the north midlands comprising the South Esk plains around Launceston and the lower midlands around Tunnack, are also recognized.

Rural and urban settlement. While the pattern of rural settlement differs strikingly by region, the basic contrasts stem from farm size and the length of settlement. In general, the older settled areas, including the midlands, the central north, the east coast, and the southeast, have larger properties, dispersed homesteads, buildings often of stone or brick, architecture that typically is early Georgian, and nucleated villages laid out on a grid. Higher rainfall areas settled since 1850, chiefly the northwest, the northeast, and the Huon Valley, have generally small farms, buildings mainly of weatherboard, and houses and villages mainly aligned along roads. In all areas, villages normally contain a post office and store, but most villages also have a primary school, public hall, church, service station, and transport services.

Rural building and town patterns

Tasmania is less urbanized than other Australian states and exhibits a dispersed pattern of peripheral growth that falls into three urban regions. Hobart, located at the foot of Mt. Wellington on the Derwent estuary, is not only the capital city, the premier port, and leading industrial centre but the metropolitan focus for the southeast, the upper Derwent, the Central Plateau, the midlands south of Woodbury, and the east coast south of Swansea. Launceston, at the head of the Tamar Valley, is a secondary administrative centre and hub of the state's transport network, with important textile and engineering industries. Its sphere of influence extends westward to Elizabeth Town and incorporates the entire north and northeast. The third region centres on both Burnie and Devonport; it includes the northwest and the west coast. Significantly, the Hobart region contains only one full-fledged town; the Launceston region, three; and the Burnie–Devonport region, five. Each, however, has six or seven minor towns.

THE PEOPLE

Ethnic and religious groups. The original Tasmanians were an anthropologically interesting Negritoid people, with the widest nasal index ever recorded and shorter and broader heads than the Aboriginal peoples of the continental mainland. They may originally have drifted across from the mainland or arrived from as far as the New Hebrides, several thousand miles to the northeast. Estimates of their numbers at the onset of European settlement vary considerably, some investigators claim a population of about 1,200, but most estimates range from 3,000 to 5,000 among several tribes.

The Aboriginal Tasmanians

The last full-blooded Tasmanian on the island died in 1876. As a consequence of the "black war" and attacks from white outlaws, they were removed in the interim to Flinders Island, off the northeast tip of Tasmania, where the survivors languished and died. Within Australia as a whole, the population of Tasmania has a distinctive composition both by birthplace and by nationality. Of all the Australian states Tasmania has the highest proportion born in Australia, the lowest proportion born in the British Isles, the lowest proportion born in continental Europe, and the lowest proportion born elsewhere in the world. Nationality is thus overwhelmingly British by descent, the proportion being higher than that of any other state.

The origins and nationality of the population are reflected in its religious affiliations. Of the four major denominations, Tasmania has, relative to its population, more Anglicans and Methodists but fewer Roman Catholics and Presbyterians than Australia as a whole. In the postwar years, the only major denomination to show a relative increase has been Roman Catholicism, partly because of immigration from Catholic countries and partly because of stringent religious beliefs regarding the use of modern birth-control methods.

Contemporary demography. Tasmania continues to display a higher birth rate than most other states, but, as in Australia generally, crude birth rates have tended to fall, average birth rates declining in the decades of the middle and late 20th century. Birth rates are lower in the cities than in the smaller towns and rural areas. Death rates remain fairly constant; partly because the ratio of children to adults in the population is high, death rates tend to be among the lowest in Australia. Infant mortality rates are also among the lowest.

Birth and death rates and migratory patterns

After World War II Tasmania experienced a net in-migration of population from other states and overseas, but since 1959 there has been, in most years, a resumption of net out-migration to the mainland, particularly of young persons, chiefly men, entering the work force. This out-migration has been accompanied by a pronounced internal rural-to-urban migration, largely through the increasing scale of farming and the substitution of capital for labour in the farming sector. As a result, Tasmania differs from the mainland in having the smallest proportion of population in the labour force and the lowest growth rate of any state.

Since more than two-fifths of the island—comprising areas in the west and the south—is too rugged and too wet for agriculture, the population is largely confined to the north and the southeast, with the sparsely settled midlands region connecting them with an isolated cluster on the west coast. Hobart has one-third of the state's population in its metropolitan area and nearly one-half within its sphere of influence; the rest of the people are distributed more or less equally between the respective spheres of influence of Launceston and the Burnie–Devonport area. No other Australian state has so equitable a balance of population among the capital city, other urban centres, and the rural areas, the nearest equivalent being Queensland. One-quarter of the population live in centres of less than 1,000 inhabitants and on farms.

THE ECONOMY

The Tasmanian economy enjoys many comparative advantages, notably in its mineral, water, and tourist resources, in the diversity of the economic activity, and in fairly stable labour relations. It suffers markedly, however, from the small scale of much of its resource base, from restricted local markets, and from problems of transport to external markets. About one-eighth of the work force is in the primary sector (the production of raw agricultural and mineral resources), one-quarter in the secondary (manufacturing and processing), and the remainder in the tertiary (services).

Resources and primary industries. Resources are highly diversified. Important mineral deposits include iron at Savage River, zinc, lead, and copper at Rosebery, copper at Queenstown, and tin and tungsten in the northeast; poor-quality coal abounds in the Fingal Valley but is of declining importance. The heavy, well-distributed rainfall and the rugged terrain in the centre and the west facilitate hydroelectric-power development. The western forests furnish not only excellent hardwoods but also the raw material for pulp-and-paper industries, while the drier eastern forests yield wood chips. Dairying and mixed farming characterize the wetter north; extensive sheep grazing, the drier midlands and the east coast; and specialized horticulture, notably for apples and hops, the southeast. Apples and pears are also locally important in the Tamar and Mersey valleys, potatoes and canning peas in the northwest, and vegetables in the northeast. Fisheries occur in all coastal waters, the catch mainly comprising crayfish, abalone, shark, barracouta, and scallops. A tourist potential lies in variety of landscape.

Agriculture and fishing

The high degree of diversification in resource development is reflected in the relative contribution of each primary industry to the net value of primary production: mining generally contributes about one-quarter, dairying one-fifth, wool growing one-sixth, fruit growing one-seventh, and forestry one-eighth. Iron ore mined by the opencut method at Savage River is converted to slurry, piped to Port Latta near Stanley, pelletized, and exported to Japan. Copper ore mined underground at Queenstown and Rosebery is shipped directly to Japan, Tasmania accounting for nearly one-sixth of the total Australian copper output. Tin production fluctuates with market conditions. Butter, wool, and fruit are exported mainly to Great Britain. Tasmanian hops meet almost all of Australia's requirements. In timber production, Tasmania furnishes about one-sixth of the Australian total; wood chips are exported to Japan. The fishing industry depends largely on exports, which are fairly evenly distributed between mainland and overseas markets.

Manufacturing. Secondary industries are twice as important in net value of production as are the primary industries. Foremost among them are the electrometallurgical and electrochemical specialties dependent on cheap, bulk power provided by the Hydro-Electric Commission. The resultant industries include the electrolytic refinery in Risdon, which treats zinc concentrates from Broken Hill (New South Wales) and Rosebery, the aluminum and ferro-manganese plants at Bell Bay, near George Town, pulp-and-paper mills in Burnie and Boyer near New Norfolk, a pulp mill in Port Huon, and a cement plant in Railton. Other important industries include the manufacture of chocolate near Hobart, titanium pigments (Burnie), textiles (Launceston and Devonport), carpets (Devonport), and alginate from sea kelp (Orford).

Administration of the economy. Government intervention in economic matters takes place largely at the federal level in the primary industries and at the state level in the secondary industries. The Directorate of Industrial Development seeks to foster manufacturing growth by the provision of financial and other assistance. The state government also is active in promoting tourism and trade. Tourist development is favoured by Tasmania's highly diversified and spectacular scenery, its equable climate, its coastal, lake, and mountain recreation areas, and its colonial heritage of Georgian architecture. As a result, Tasmania is visited by more than 100,000 tourists each year. The private sector supports well-established chambers of commerce, a chamber of manufacturers, a tourist council, and more than 100 trade unions.

Transportation. Given its island setting and dispersed pattern of development, transportation, both internal and external, is especially important to Tasmania. The Transport Commission regulates and, where necessary, provides internal transport. Tasmania has the highest density of roads per square mile of any Australian state. The government railways comprise more than 500 miles (800 kilometres) of line. Of the intrastate freight movement, the majority is by road, about one-third by rail, and about one-tenth by sea; virtually all passenger movement is by road. Almost all interstate and overseas freight moves by sea and almost all passengers by air. Of the major seaports, Hobart has about one-third of the trade and Launceston, Burnie, and Devonport each about one-fifth; port administration is decentralized. Minor ports are Stanley, Currie, and Strahan. Of the airports, Hobart handles the most passengers (with Launceston close behind) and Launceston the most freight. Regularly scheduled air services also operate from Devonport, Wynyard, King Island, and Flinders Island.

ADMINISTRATIVE AND SOCIAL CONDITIONS

Government and politics. The state parliament comprises a House of Assembly and a Legislative Council, the latter by tradition a largely nonparty house. The system of elections for the House of Assembly is proportional representation by the single transferable vote; the system for the Legislative Council is the preferential system, with an obligation to record preferences, used for the federal House of Representatives. The main political parties active in Tasmania are the Australian Labor Party and The Liberal Party of Australia; the Australian Centre Party affiliated with The Australian Country Party dates from 1966. The Labor Party was in power from 1934 to 1969, when it was replaced by a Liberal–Centre party coalition; in May 1972 the Labor Party once more regained office. Executive government is by the Cabinet system, the governor representing the British monarch and the Executive Council of Ministers of State giving Cabinet decisions where necessary to legal form. Local government is exercised through cities and municipalities, which comprise the 49 divisions of the state.

Legislative, executive, and judicial institutions

Although the state now administers some of the lower courts, most are still administered by municipal authorities, a feature unique to Tasmania; in other states the administration is vested in a state department. Courts of petty session have jurisdiction over all summary offenses and certain indictable offenses at the option of the defendant. Minor civil proceedings are dealt with by courts of

request in the cities and some municipalities or by courts of general sessions. The Supreme Court of Tasmania sits regularly in Hobart, Launceston, and Burnie; it has jurisdiction over all causes save those reserved to the high court of Australia under the commonwealth constitution. The high court normally sits in Hobart each year, but Tasmanian cases are also heard in Melbourne and Sydney. Children's courts have jurisdiction over persons under the age of 17.

Social welfare. Wages and working conditions of Tasmanian employees are regulated either by awards of the Commonwealth Conciliation and Arbitration Commission or by state wages boards. Tasmania tends to have the lowest average weekly earnings per employed male of all the states except Queensland and, usually, South Australia. That weekly earnings are not the lowest is in line with the cost of living, which is relatively high, partly because of freight charges on imported goods. Although the standard of public health is good, there is a high incidence of hydatids in some rural areas, and goiter is endemic locally. Social status, as in Australia generally, derives primarily from occupation, but Tasmanians accord a lower ranking to professionals and a higher ranking to businessmen than do Australians as a whole. The standing and numerical importance of small and medium property owners is reflected in the relatively conservative tendency of the two main political parties. There are no great extremes of wealth or poverty.

The general standards of living

Social-service benefits provided by the Commonwealth include maternity allowances; child-care payments; unemployment, sickness, and special benefits; age, invalid, and widows' pensions; and funeral benefits. The commonwealth also provides paid employment for the disabled, a rehabilitation service, and subsidies for aged persons' homes. The state Department of Social Welfare provides assistance to deserted wives or husbands with children, wives with husbands in prison, and neglected or deserted children. It also maintains homes for maladjusted and delinquent children and wards of the state.

Education. School attendance is compulsory for all children between the ages of six and 16. Government-supported schools include infant, primary (some with preschool facilities), area, district, and high schools (nonselective, comprehensive, mostly coeducational), together with matriculation colleges (a Tasmanian innovation) and special schools. Independent (nongovernment) schools, which enroll about one-sixth of the school population, are mostly controlled by religious denominations, the great majority being Roman Catholic; since 1967 independent schools have received state aid. Institutions of higher learning include technical colleges, teachers colleges at Hobart and Launceston, the University of Tasmania (founded 1890) at Hobart, and the College of Advanced Education (opened 1972), providing technical, education, liberal arts, and professional courses.

Health. The state government controls directly or through hospital boards the general hospitals at Hobart, Launceston, Latrobe, and Burnie–Wynyard; numerous district and district nursing hospitals; a mental hospital at New Norfolk; maternity hospitals at Hobart and Launceston; a chest hospital at New Town; and hospitals for the aged at Hobart, Launceston, and Wynyard. It also provides district medical officers for the more remote areas, a district nursing service, a school medical service, a school dental-health service for which it trains dental nurses, and a child-health service. Private institutions include seven hospitals and numerous nursing homes. Annual X-ray examinations for adults are compulsory by law. Public water supplies are fluoridated at Hobart, Launceston, and Beaconsfield.

Housing. Three-quarters of all Tasmanian houses are owner occupied, and two-thirds are built of weatherboard; apartments make up less than one-tenth of all dwellings. The Commonwealth Housing Commission administers the homes-savings-grant plan, which makes a gift to young persons with certain minimum savings for a first home, a housing-loans-insurance plan, and assistance to certain former members of the armed forces. The State Housing Commission caters to the lower income groups, providing mainly three-bedroom timber units for families to rent or buy and apartments for elderly persons to rent. Rental rebates are allowed as necessary. Advances for home building are made available by the Agricultural Bank of Tasmania and cooperative building societies.

Public aid for housing

CULTURAL LIFE

The arts. For the smallness and dispersion of its population, Tasmania has a remarkably vigorous cultural life. At the amateur level, there are numerous musical groups, ranging from the full orchestra to the ensemble, as well as several choral societies and repertory companies. The College of Advanced Education incorporates a conservatorium of music and a school of art. The Australian Broadcasting Commission, with financial support from the Hobart and Launceston city councils, maintains the Tasmanian Symphony Orchestra, which gives regular concerts in the main centres, often with visiting artists from the mainland or overseas. The National Theatre and Fine Arts Society of Tasmania, in association with the Elizabethan Theatre Trust, is responsible for bringing drama and opera from the mainland to the Theatre Royal in Hobart, Australia's oldest theatre (1834), and the National Theatre in Launceston. Both cities have museums and art galleries that exhibit paintings, pottery, and sculpture. Tasmania has well-established film festivals and eisteddfods (Welsh poetry festivals); in 1972 it also held its first arts festival. At Hobart's Battery Point, Narryna, a Georgian town house dating from 1839, has been preserved as the Van Diemen's Land Folk Museum, with furniture and furnishings of the early 19th century. Other Georgian houses that have been restored for public benefit include Franklin House and Entally House, both near Launceston, and Runnymede, at New Town, Hobart.

Library services. The state provides lending-library services to adults and children, including a central service with film and recorded music in Hobart, regional services in Launceston and Burnie, and country services in almost all municipalities. Bookmobiles operating from Hobart, Launceston, and Burnie serve schools and other institutions as well as rural areas that lack library facilities. A state reference library and the state archives are located in Hobart.

The communications media. Tasmania has daily newspapers published in Hobart, Launceston, and Burnie and receives national dailies from Sydney and Melbourne. Each of the smaller centres also publishes a weekend paper, and weeklies of regional interest are published in Queenstown, Smithton, George Town, Scottsdale, New Norfolk, and Huonville.

In Tasmania, as throughout Australia, broadcasting and television services are produced by both the Australian Broadcasting Commission and by commercial transmitters. National radio stations operate from Hobart, Launceston, and Queenstown, while commercial radio stations relay from Hobart, Launceston, Devonport, Burnie, Queenstown, and Scottsdale. There are both national and commercial television stations in Hobart and Launceston. Tasmania's rugged relief necessitates the provision of an elaborate network of translator stations to ensure adequate reception in all districts. The Australian Broadcasting Control Board exercises control in certain matters over the commercial private-enterprise services. The island is connected with the mainland by a broad-band radio link that makes it possible for television programs to be relayed directly from other states.

(P.Sc.)

History

EARLY EXPLORATION AND SETTLEMENT

The Aborigines of Tasmania, like the Aborigines of the Australian mainland, were hunter-gatherers when first contacted by Europeans during the late 18th century. They grew no crops, raised no domesticated animals, had no metal tools, and possessed no system of writing. They had adapted for many years to the Tasmanian environment, prior to their confrontation with 19th-century Europeans and the metal-gunpowder technology. Radiocarbon dating from two sites in northwest Tasmania has shown that

Aboriginal people occupied the region about 8,000 years ago.

Origin of the Aborigines

Two main theories of their origin are held: (1) They formed the original population of eastern coastal Australia; then, yielding to Australoid immigrants, they crossed into Tasmania. (This view implies a land corridor or shorter stretches of ocean and also winds and currents in the opposite direction to those now prevailing.) (2) They drifted to Tasmania in canoes or rafts from the New Hebrides region where there are similar Negritoid types; winds and currents favour such voyages as well as possible landings on the Queensland coast.

The most reliable assessment of Tasmanian population before European settlement in 1803 is 2,000 or less; the last full blood is recorded as having died in 1876. This peaceful people became fearful and resentful when faced with dispossession and with cruelty on the part of some European convicts and settlers. By 1826 the clash of cultures was officially recognized; by 1831 the remaining Aborigines were removed first to Gun Carriage Island and then to Flinders Island, where, in spite of care, they languished and died out.

The Tasmanians were organized as five tribes, each of which spoke its own dialect or language. They usually travelled in small groups, gathering food and hunting with sticks and clubs; unlike mainland people, they had neither spear-thrower nor boomerang. They held corroborees, cremated the dead, scarred and red-ochred their bodies, and engraved lines and circles on rocks.

After the initial exploration undertaken by Abel Tasman, Tasmania (then Van Diemen's Land) was explored by the French and British navigators Marion Dufresne (1772), Tobias Furneaux (1773), James Cook (1777), William Bligh (1788 and 1792), Bruni d'Entrecasteaux (1792–93), John Hayes (1793), George Bass (1798), Nicolas Baudin (1800 and 1802), and Matthew Flinders (1798 and 1802). In 1803, a party of 49 persons, led by John Bowen, settled at Risdon Cove. The settlement was abandoned and relocated at the new settlement of Hobart in 1804. The British settlement at Hobart was followed by another that was to become Launceston (1806); most of the first settlers, both free and convict, were transferred from Norfolk Island. The settlement, under resident lieutenant governors, was governed from New South Wales. Free immigrants were attracted by land grants and by the development of the sheep and whaling industries.

INDEPENDENT SETTLEMENT

Van Diemen's Land was made a colony by order in council in 1825, and henceforward its government was practically separated from that of New South Wales and until 1850 was exercised through nominated executive and legislative councils. Law and order and a sound administration were introduced by Lieutenant Governor William Sorell (1817–24) and were developed by Lieutenant Governor George Arthur (1824–36), with wider powers, into a detailed and efficient control.

"Black war"

In this period, despite the depredations of the outlaw bushrangers and the "black war," which virtually extinguished the Aboriginal people, great material progress was made. In 1828 the total white population was 17,000, of whom 7,500 were convicts under restraint, though a majority were assigned as servants to free settlers. Grain was exported regularly to New South Wales, while rising wool prices brought general prosperity and stimulated much speculation in land. The adoption of the probation system for convicts in 1840 marked a change in British policy on transportation, depriving free settlers of assigned servants and transforming the colony into a penal settlement. The influx of population was rapid until 1847, growing to 70,000, but the ratio of free to unfree was reversed. In the late 1840s the boom gave way to an economic depression, and the colonial government faced a deficit, which local opinion attributed to the cost of police and jails. Belatedly, the home authorities offered some relief but not before public opinion was roused to demand representative institutions and the cessation of transportation of British convicts. Representatives were first elected to the Legislative Council in 1851, and the transportation policy ceased in 1853.

The constitution was settled more definitely in 1856 with the election of a two-chamber Parliament and with the appointment of the first responsible ministry. From 1856, the colony was known as Tasmania.

The gold discoveries in Victoria in 1851 caused inflation and labour shortage in Tasmania, and the economy was unable to support the strain. The whaling industry declined and with it shipbuilding. The adoption of a protectionist policy by Victoria further injured Tasmanian manufactures. After 1857 a prolonged depression, lasting 25 years, was marked by deficits, reduction in services, opposition to all forms of taxation, political feuds, and frequent changes in ministries. Under pressure of competition from continental Australia, wheat production declined to its present insignificance.

After 1880 economic conditions improved, largely because mining industries were opened up in the west and northeast. The hardwood forests were exploited, and fruit growing in the south became the basis of small manufacturing and export industries. Railways, roads, and harbour facilities were developed, while relative political stability was achieved, particularly during the Fysh ministry (1887–92). After the recession following on the failure of the Bank of Van Diemen's Land in 1891, there was from 1894 to 1899 a slow recovery during Sir Edward Braddon's ministry. In the 1890s a majority of Tasmanians supported the movement that ended in 1901 with Australian federation. Economic development slowed down after 1910 as the mining boom passed, and during the next 25 years Tasmania was relatively depressed, the worst years being 1922–26 and 1931–35. Population grew slowly, the birth rate fell, and the exodus of the young and more vigorous people was resumed.

Labor Party

State politics entered a new phase with the founding of the Labor Party in 1903. Labor first took office in 1909 and was again returned to power in 1914. From that time both the Labor and the Liberal (Nationalist) parties sponsored social and industrial legislation. All the administrations faced financial difficulties, which were not resolved until in 1933 the commonwealth government set up the Grants Commission, through which special financial assistance is given to the state. This enabled Tasmania to bring the level of its social services up to the national average. After 1942 uniform taxation throughout the commonwealth and virtual commonwealth control of the main sources of revenue, together with the provision of reimbursement grants to the states, favoured Tasmania, which received more per capita in such grants than any other state. With increasing revenue, supported by loan funds backed by the commonwealth, Tasmania has invested since 1950 in a large capital works program, especially in the development of hydroelectric power, which in turn has stimulated the growth of much secondary industry.

A Labor government was formed in 1934 and won eight successive elections, continuing without a break to govern the state until the 1969 election when it was replaced by a Liberal–Centre Party coalition. Liberal–Centre Party rule ceased in 1972, following Cabinet resignations, and the Labor Party returned to power in the elections of that year. The Labor Party has remained in power since 1972, winning majorities in the House of Assembly in 1976 and 1979.

During the mid-20th century the Tasmanian economy went through a transformation. Secondary industry expanded because of increases in hydroelectric energy. Some of the major industrial operations established since 1945 include the production of aluminum at Bell Bay in 1955, the discovery of uranium at Mt. Balfour and Royal George the same year, the Savage River mining operations along the west coast and the associated iron-ore pelletizing plant at Point Latta, and acid production for the processing of pyrites at Burnie in 1970.

In February 1980, the state government approved a new system of government by consensus in which the opposition and the Legislative Council would be allowed a voice in the governmental decision-making process.

(A.P.E./W.A.T./Ed.)

VICTORIA

Among the Australian states, Victoria is second to New South Wales in terms of population, production, and party strengths in federal politics, but, since 1945, as the traditional financial hub of Australia, has sustained the faster rate of economic advancement. Thousands of immigrants have arrived, and manufacturing industry has expanded to the point at which the economy is now broadly and soundly based. Two recent developments—the exploitation of the gas and oil fields of Bass Strait, which will provide both cheap fuel and state revenues, and an integrated steel plant—were, in the late 20th century, further narrowing the economic gap between Victoria and New South Wales.

Victoria, which lies at the southeastern tip of the continent, was politically separated from New South Wales in 1851, and responsible government was conferred in 1855, with power vested in a bicameral legislature. On January 1, 1901, Victoria and the five other Australian colonies became states and formed the Commonwealth of Australia. Only Tasmania is smaller than Victoria's 87,884 square miles (227,620 square kilometres), but only New South Wales has a population greater than that of Victoria. Victoria is separated from New South Wales to the north by the River Murray (for a length of 1,065 miles) and also by a straight boundary (110 miles long), linking Cape Howe and the nearest source of the Murray. The meridian marking 140°58′ E forms the western boundary with South Australia for a distance of 280 miles. The coastline, which completes the land boundaries, stretches for 980 miles. It includes the 164-mile shoreline of Port Phillip Bay, at whose head lies Melbourne, the capital and major city. The maritime boundary with Tasmania is latitude 39°08′ S.

Physical and human geography

THE LAND

The rich variety of landscapes in Victoria includes both alpine plateaus in the northeast, around Bright, and sandy deserts in the west, near Lake Hindmarsh. This wide range results from a complex geological history and from variations in the weather as it is experienced in particular areas. These dominant factors have created distinct regions, which create different opportunities and problems. The main upland areas are a continuation of the Dividing Range of eastern Australia. Starting with a width of 190 miles on the New South Wales border, these uplands arc westward across the state, becoming narrower and lower for 400 miles before terminating in the Grampians and Dundas Highlands, 25 miles east of the South Australian boundary. Plains surround this upland core on the north, west, and south. The southern plains are divided into smaller areas by the Otway Range, the South Gippsland Highlands (north of Wilsons Promontory), and Port Phillip Bay, which was formed by the invasion of the sea after a downward movement of the earth's crust.

Climate. The easterly passage of anticyclones (high-pressure areas) and depressions is the main determinant of weather for most of Victoria. Their track lies overland during winter (generally the wettest season), and during summer a more southerly oceanic course reduces the frequency of rain days. East Gippsland is an exception, since most of its rain results from intense depressions centred east of Bass Strait in the summer. There is a close correlation between annual rainfall and elevation, and there is a clear decline in annual rainfall toward the northwest.

Relief. *The central uplands.* The upland core of Victoria is divided by the low, wide Kilmore Gap into two distinct sections. The Eastern Cordillera are more extensive and higher, with several peaks over 5,000 feet, culminating in Mt. Bogong (6,516 feet [1,986 metres]). There are also some high plateaus. The varied geological structure has been heavily cut into by perennial streams, fed in spring by melting snow and ice. The steeper slopes and longer winter of the Eastern Cordillera have discouraged cultivation except in sheltered valleys, where some hops and tobacco are grown. Beef-cattle breeding and fattening, the raising of prime lambs, and the production of crossbred wool are the main rural activities. This area also contains the most extensive forest in Victoria, the main species being snow gum, Alpine ash, and mountain ash. Its beautiful scenery makes it a popular year-round attraction for tourists.

The highest of the peaks in the Western Highlands is Mt. William (3,829 feet [1,166 metres]), in the Grampians. The low relief, gentler slopes, and milder winter makes the area suitable for farming, and much of the forest has been cleared. Most farmers rear sheep and beef cattle.

The northern plains. The plains north of the uplands can be divided into three distinct areas. Apart from a distinctive narrow strip adjoining the River Murray, where conditions for farming are more reliable, the entire northwest area, bounded by 36° latitude and the Avoca River, is known as the Mallee. This name is derived from a type of eucalypt that sends up a number of slender trunks from a single large, underground source; the trunks rarely exceed 30 feet in height. The region has unreliable annual rainfall, generally less than 15 inches. The undulating surface of broad, low ridges and depressions reflects the faulting and folding of the underlying sedimentary rocks, which in turn have been overlain by windblown deposits that, except in the region known as the Big Desert, have been fixed by drought-resisting vegetation. No streams rise in the Mallee, and those that enter from the south fail to reach the Murray, terminating instead in such salt lakes as Lake Tyrrell. Apart from the extreme northwest and the Big Desert, where lack of water and wind-erosion hazards make conditions too difficult for farming, the area has been only lightly settled. The light soils are easily cultivated, and the development of suitable fertilizers has allowed better crop rotation and the production of improved pasture. Wheat and fodder are the main crops, supplementing the production of prime lambs and wool.

Erosion hazards

The plains that lie between the Mallee and the uplands have a uniform rainfall, which varies from 15 inches (381 millimetres) on the northern edge to 24 inches near the foothills. In the area west of the Loddon River, known as the Wimmera, the soils are mainly gray clays that swell when wet and crack open when dry. Wheat growing and the raising of prime lambs and Merino sheep form the basis of rural industries there, except in the Little Desert, where deep sands, deficient in zinc and copper, make the land unsuitable for settlement.

The plains east of the Loddon River, which narrow toward Albury, are irrigated and support a wider range of farming activities. In recent geological periods the forerunners of present rivers laid down the alluvial deposits, which reveal considerable diversity in sand and clay content. When irrigated, the differing soils allow the rearing of sheep and dairy cattle and the cultivation of wheat and fruit.

The southern plains. Port Phillip Bay divides the southern plains into two distinct regions. To the west, most of the surface is formed from ancient basaltic flows, above which stand some of the original volcanic cones. The flows covered 7,000 square miles and stretch 190 miles west of Melbourne. The uniform origin of this plain has not provided identical soils, as there are local differences in the colour, texture, and stone content of the soils. These variations were, however, masked by the grasslands that originally covered the area, and today there is a uniform land use. Livestock are preeminent here: fine-wool-sheep breeding and beef-cattle breeding and fattening are general, with dairying around the towns of Colac, Camperdown, and Koroit.

The Gippsland Plains are located east of Port Phillip Bay. The larger East Gippsland Plain experiences maximum rainfall during the summer. It has fairly mild winters. These conditions are excellent for dairy farming, which is the main rural activity, although the production of veal and pork supplements the yields of milk and cheese. Inten-

sive settlement of this area has resulted in the destruction of most of the original forest cover.

THE PEOPLE

Before 1939 the majority of Victoria's population had been born in Australia, and in 1947 only 8.7 percent of the population was foreign-born, most of it British. Since 1945, however, Australia in general and Victoria in particular have encouraged large-scale immigration from Europe in order to make the country stronger strategically, to assist many European refugees made homeless by the war, and to reduce the economic problems caused by the low Australian birth rate during the interwar depression. By the latter decades of the 20th century more than one-fifth of the Victorian population was foreign born. The main national groups, apart from Australians and British, were Italian, Greek, Dutch, and German. In some sections of Melbourne, Greek or Italian communities predominate, but the "new Australians" mix easily with the rest of the population, and national friction is at a minimum. The benefit of these groups to the state's economy is incalculable. European migrants have also made a distinctive contribution to the cultural life of Victoria—in architectural style and house decoration, in sport, with soccer growing in popularity, and in culinary variety. Estimates indicate that there are about 6,000 Aborigines in Victoria. By the Aborigine Land Act of 1970, title to a few thousand acres was transferred to a Trust of Aborigines. Some Aborigines face problems of adjustment to a white society, and government policies are designed to reduce difficulties associated with housing, employment, and education.

The role of immigration

The diverse origins of the population are reflected in the variety of religious faiths found in the state. These include, most importantly, the Anglican Church, followed by Roman Catholic, Presbyterian, Methodist, Greek Orthodox, Baptist, Churches of Christ, Lutheran, and Jewish groups.

Almost three-quarters of the population live in the Melbourne metropolitan area, about one-tenth live in six other urban areas (Geelong, Ballarat, Bendigo, Albury-Wadonga, Morwell, Shepparton-Mooroopna), and the rest of the population resides in towns of less than 20,000 and in rural areas. Population distribution outside the metropolitan area reflects the qualities of the landscape. Geelong is the second port of Victoria; Ballarat and Bendigo originally grew up around goldfields, now largely worked out, and Moe-Yallourn stands on brown-coal beds, used for electricity production. The densest rural settlements are in fertile sections of the irrigated Murray Valley and the dairying areas of Gippsland, and the lowest are in the Alpine sections of the Eastern Cordillera and the dry Mallee.

THE ECONOMY

Victoria has a broadly based economy with well-developed primary, manufacturing, and service sectors. The lack of major mineral deposits was balanced by the discovery of huge natural-gas fields in East Gippsland waters in 1965 and of major oil fields in eastern Bass Strait in 1967.

Most of the main farming areas are used for improved or natural pastures or cultivation, which involves mainly wheat and fodder crops. The main categories of productive holdings are dairy farms, sheep stations, mixed sheep and cereal farms, cereal farms, and beef cattle.

Resources and manufacturing. Prior to the discovery of oil and natural gas, the brown-coal deposits near Moe-Yallourn were the mineral deposits of greatest value to the state. The focus of attention has shifted to Barracouta, a gas and oil field in 150 feet of water, 14 miles offshore; Marlin, a gas field in 195 feet of water 32 miles offshore; and the Halibut and Kingfish oil fields in 250 feet of water 40 and 48 miles offshore, respectively. All these fields have been linked by pipelines to the Longford gas-processing and crude-stabilization plant in Gippsland and the Long Island Point fractionation and crude-storage plant on Western Port Bay.

Oil and gas discoveries

Victoria's factories employ several hundred thousand workers: the vast majority of the factories and workers are employed in Melbourne, and Geelong and the coalfield centres of the Latrobe Valley. The original industrial suburbs of Melbourne had a central location, but many new factories are now being constructed in peripheral areas, such as Altona, Dandenong, Broadmeadows, and Moorabbin, where larger areas of cheaper land are available. Geelong, like Melbourne, produces a wide range of products, but other centres tend to be specialized. The Latrobe Valley is noted for the generation of power, and other centres are concerned mainly with food and clothing manufacture, using local materials. Although only about one-tenth of the factories employ more than 50 workers, such factories employ two-thirds of the total work force. In terms of numbers employed and value of wages, the most important industries produce metals, machines, vehicles, clothing, foodstuffs, and paper.

Transportation. The major port and the focus of the rail, air, and road systems of Victoria is Melbourne. Melbourne and Geelong ports between them handle almost all of the cargo entering and leaving the state, and Melbourne is the dominant passenger terminal, as well. Victorian Railways serve all productive areas through the several thousands of miles of mainly single-track lines. Since 1962, narrow gauge tracks have linked Melbourne with the standard system of New South Wales at Albury. The capital's electrified metropolitan system carries thousands of passengers each working day. Melbourne Airport was opened to international flights in 1970 and to domestic flights in 1971. It stands on several thousand acres, 13 miles from the city, and includes an industrial estate served by taxiways. Through highways, with many divided sections, link all the major centres of the state.

ADMINISTRATIVE AND SOCIAL CONDITIONS

Government. The Parliament of Victoria consists of two houses: the Legislative Assembly and the Legislative Council. The leader of whichever party or alliance of parties forms a majority in the Legislative Assembly is requested to form a government by the governor, the titular representative of the English monarch. The premier-elect submits names of proposed ministers to the governor for appointment. These ministers become members of the Executive Council, which advises the governor. His position places him above party politics, and he is regarded as the trustee of the constitution. The governor summons and prorogues Parliament, outlines the government's legislative program at the beginning of each session, and gives assent to bills that do not have to be referred to the monarch.

The role of the governor

The Victorian Parliament legislates for those subjects not exclusively granted to the commonwealth Parliament by the commonwealth constitution. If there is any inconsistency between state and federal laws, the federal laws prevail, and the subjects granted to the commonwealth may be varied by an appropriate commonwealth act. The public service of Victoria is based on the departments of chief secretary, premier, and treasury, which serve a variety of purposes, and a number of other departments and ministries dealing with a single subject. There are also several public corporations, of which the most important are the Country Roads Board, the Gas and Fuel Corporation, State Electricity Commission, and State Rivers and Water Supply Commission.

The political process. The members of the Legislative Assembly are elected by universal adult suffrage in single-member electorates for a period of three years. Members of the Legislative Council are elected from two-member electoral provinces by universal adult suffrage for a six-year term. Elections to the council are arranged so that half the members, one from each province, retire every three years. Voting for both houses is compulsory.

In the late 20th century there were three major political parties in Victoria—the Liberal Party, the National Party, and the Australian Labor Party. The Liberal Party, predominant in recent decades, supports free enterprise and draws most of its support from middle- and upper-class voters in urban and some rural areas. The National Party represents rural interests and wins no urban seats. The Australian Labor Party finds its strongest support among the working classes and forms the official opposition when not in office.

Local government and the judicial system. Apart from

some lakes and small islands, the total area of Victoria is incorporated into more than 200 municipal districts for local government. There are, in addition, shires, boroughs, towns, and cities, which are distinguished by population and the annual revenue derived from rates (taxes). The towns contain at least 5,000 citizens each. The cities in Victoria have at least 10,000 residents each. Councils are elected each August by ratepayers and derive their revenue from rates, government grants, and certain license fees. Some councils levy rates on net annual value, while the remainder employ unimproved capital value.

Before 1973 local governments received no financial assistance from the commonwealth government. Financial assistance did exist, but it was filtered through the state governments. Procedures have been established whereby local governments can apply to the commonwealth for general purpose grants. Victoria local governments received several million dollars in grants from the commonwealth during the initial years. Victoria and its local governments also receive personal income tax funds that are redistributed by the commonwealth under the Local Government Personal Income Tax Sharing Act of 1976. The monies are distributed by the state of Victoria to the local governments based on a fixed percentage of income taxes collected within the municipality the previous year.

The courts of justice in Victoria are graduated in status according to the gravity of cases that they consider. Petty sessions courts, found in metropolitan suburbs and about 200 other towns, deal with less serious criminal offenses, hold initial inquiries into indictable criminal offenses, and adjudicate civil matters not involving more than A$1,000. A county court sits continuously in Melbourne, and visits 18 other circuit towns. It deals with civil matters in which the amount does not exceed A$4,000, and all criminal offenses except murder, treason, and other statutory exceptions. It may also act as an appeals court for petty sessions courts. The Supreme Court sits in Melbourne and visits ten other cities. It may deal with all matters not excluded by statute and act as appeal court for the county court.

Social welfare and education. Wages and working conditions in Victoria are supervised by the Commonwealth Conciliation and Arbitration Commission, for industries that extend beyond the state, and by the Victorian Conciliation and Arbitration Boards, for industries entirely within the state. In 1981 the national basic minimum wage, based on fluctuations in the Consumer Price Index, was abolished. Basic wages are now established according to specific industry or trade requirements. In 1982 a 38-hour workweek was implemented by numerous conciliation and arbitration boards.

The basic wage

Since 1872 every child in Victoria has been entitled to secular, compulsory, and free education to the age of 15 years. Both state and private schools operate. Primary schools offer seven years' education and teach general academic skills. Secondary schools offer six years' education and should fit the student for any tertiary course. Technical schools operate on a five-year course and qualify students for training in technical colleges. In addition, there are certain special schools for handicapped children. There are three universities in Victoria. The Victoria Institute of Colleges is an association of many technical colleges, which now awards degrees, as well as diplomas, in certain fields.

CULTURAL LIFE

The most important recent cultural development in Victoria has been the creation of the Victorian Arts Centre, on land near the centre of the city of Melbourne. The entire project includes art galleries, courtyards for theatrical productions and displays of sculpture, underground theatres, a convention and concert hall, a display centre, and a hall for state receptions. There are several other art galleries at such centres as Ballarat, Bendigo, Geelong, Warrnambool, Hamilton, and Shepparton, and many of them award annual prizes to be won in competition. A folk museum of buildings, machinery, and paddle steamers has been built on a large site at Swan Hill. The Library Council of Victoria, created in 1966, manages the important State Library of Victoria and advises the government on the promotion of library services throughout the state. During the last century the State Library built up strong collections in a variety of fields, but more recently a shortage of funds and rising costs have limited the areas where collections are made in depth. Fields of current interest include historical bibliography, fine arts, biography, military history, and British typography. The collection of Victoriana in the library and the state archives are valuable historical sources.

The Victorian Arts Centre

Melbourne's orchestra has made successful overseas tours, and the city has many cinemas, theatres, and restaurants. The National Trust of Australia (Victoria) plays an important part in stimulating the public's interest in old buildings and places of scenic beauty and with rare flora and fauna. A number of old buildings have been acquired and restored, and a schedule of other existing buildings has been prepared to avoid major cultural loss through demolition. Melbourne's daily papers have a combined daily circulation exceeding 1,300,000. Most of the other urban centres throughout Victoria have their own local papers, and other daily papers circulate in the state. The Australian Broadcasting Commission, financed by the government, provides a comprehensive radio and television service to the entire state. There are also many commercial radio stations and several commercial television stations. Victoria has numerous national parks with a total area of several hundred thousand acres. (J.R.V.P.)

History

EXPLORATION AND SETTLEMENT

Victoria was founded by groups of pastoral pioneers who crossed Bass Strait from Launceston in the 1830s in search of fertile grazing land. The occupation of the area was made in defiance of a British government edict forbidding settlement in the territory, which was then part of the colony of New South Wales.

In November 1834 Edward Henty and his father and brothers landed stock and stores at Portland, on the south coast, and established the first settlement. In 1835 John Batman, backed by a Launceston syndicate, landed at Port Phillip. Batman's venture led the way to the pastoral occupation of Victoria. In the same year John Pascoe Fawkner established a small colony of settlers on the banks of the Yarra River. From Batman's colony grew Victoria's capital city, Melbourne.

Early settlement of Melbourne

Exploration by sea and land had preceded settlement. George Bass, Matthew Flinders, and John Murray had established the coastline and the main harbours by the beginning of the 19th century. In the 1820s and 1830s a series of overland expeditions from New South Wales had opened up the hinterland. Hamilton Hume and W.H. Hovell struck south and reached the coast of Port Phillip in 1824; Charles Sturt plotted the full reach of the Murray in 1829; Maj. Thomas L. Mitchell crossed the central and western plains in 1836; and several parties had penetrated the mountainous Gippsland district to the east by 1840. Early attempts to establish convict settlements on the south coast (near Sorrento in 1803 and on Western Port in 1826) failed. But the Port Phillip settlement developed rapidly. In 1836 the British government authorized the settlers but gave no title to their land. In December of that year Capt. William Lonsdale was appointed first resident magistrate.

During the 1840s and 1850s Victoria became a prosperous pastoral community. The squatters extended their grazing runs to the boundaries of the territory. The population rose rapidly as British migrants arrived and more settlers crossed from Van Diemen's Land (renamed Tasmania in 1856) or drove their flocks and herds south from New South Wales. By 1850 Victoria had 76,000 people and 6,000,000 sheep. Melbourne (pop. 23,000), Geelong, and Portland were its main urban centres.

INDEPENDENT SETTLEMENT AND DISCOVERY OF GOLD

Dissatisfied with their limited representation on the Legislative Council of New South Wales, the Port Phillip pastoralists agitated for separation. In 1851 Victoria became a separate colony, with an Executive Council appointed

by the crown and a Legislative Council, partly elected and partly nominated, effectively dominated by conservative landed interests.

Victoria's establishment as a separate colony coincided with the discovery of gold. An early find at Warrandyte, 16 miles from Melbourne, led to a dramatic rush. By the end of 1851 half the men of the colony were working on the goldfields. In 10 years, more than £105,000,000 worth of gold was won from fields of which Ballarat and Bendigo were the most important. More than 200,000 migrants arrived there from Britain and 25,000 from China. By 1860 the population of Victoria had passed the 500,000 mark and was 46 percent of the Australian total.

Gold transformed Victoria from a pastoral backwater into the most celebrated colony of the empire. The comparatively well-educated and skilled artisans lured by gold produced a society renowned for its attachment to 19th-century middle-class values and institutions. Although actual opportunities soon contracted and a Melbourne proletariat rapidly emerged, Victoria was noted for its economic individualism and opportunism, its material progress and financial speculation, as well as its imperial loyalty and political pragmatism.

The gold rushes produced a spectacular, but short-lived, boom. By 1854 Melbourne was suffering from a severe depression. Financial stringency aggravated discontent on the goldfields. The miners strongly resented the fee demanded for a mining license and the brutal fashion in which it was collected by the goldfields' police. This discontent culminated in a minor rebellion at Eureka, near Ballarat. Mining licenses were burned and a republican flag was hoisted. On December 3, 1854, police and troopers stormed the rebels' stockade. Thirty miners and four soldiers were killed. But the incident hastened the redress of the miners' grievances and gave colonial radicals symbols and martyrs.

Miner's revolt

Victoria attained self-government in 1855. The new constitution set up two Houses of Parliament—a Legislative Council of 34 members, elected on a limited property franchise, and a Legislative Assembly, elected on a wider property and income franchise.

The Legislative Council remained the stronghold of the rich conservative landowners and the main obstacle to land reform. But in the 1860s a series of land acts, designed to encourage small freeholders and "unlock" the large grazing leases of the pastoralists, helped establish small wheat farmers in the Mallee and Wimmera. Victoria, previously an importer of flour, became Australia's largest wheat producer by the end of the century.

In other districts, however, the wealthy pastoralists managed by guile and financial manipulation to evade the land acts, and very many of them acquired freeholds to large estates at low prices, especially in the fertile Western District.

In 1871 the property qualification for the Council was reduced and the tenure of Council members shortened. In 1888 additional electoral reforms for both houses were passed. With a few exceptions, single-member constituencies became the rule for the Assembly. In 1899 plural voting for the Assembly was abolished and in 1900 postal voting introduced. A free, compulsory, and secular educational system was established in 1872. The introduction of an eight-hour working day in 1856 began a series of social and industrial reforms, which produced a minimum wage and standard hours and conditions of employment in the 1890s.

By the end of the 1880s the state's growing prosperity had expanded into a speculative boom. The crash came in 1891 and was considerably aggravated by a sharp fall in the prices of Victoria's main exports, wool and wheat. In 1891–92, 21 finance societies and land banks collapsed and in 1893 all but three of the trading banks closed their doors. The depression that followed was marked by high unemployment and considerable industrial unrest; it did not completely end for almost 20 years. These disasters transformed Victorian politics and socioeconomic attitudes and behaviour. The confidence of the bourgeoisie was completely destroyed. Men of property rallied, but without intellectual or even spiritual convictions, to defend the old order against those who proposed change. In this they were overwhelmingly successful. From Australia's most radical and progressive colony, Victoria, almost overnight, became the bulwark of conservatism. In consequence, although all parties were forced in some measure to recognize the peculiar problems of the environment by introducing a form of state Socialism, Victorian politics in the 20th century were generally characterized by greater ideological cleavages than usually appeared in the other states and by the monopoly of effective political power by interests committed to the preservation of 19th-century notions of property and social conformity.

FEDERATION AND THE STATE OF VICTORIA

In 1891 the first Australian National Convention met in Sydney to consider proposals for the creation of an Australian federation. On May 9, 1901, the first federal Parliament was opened in Melbourne. It was moved to Canberra in 1927.

First Australian National Convention

Following federation, Victoria's final legacy to the Australian colonies, there was a need to rationalize the constitutional structure of the states. In Victoria, the number of members was reduced in 1903, the Council franchise and property qualifications for membership were liberalized, and adult suffrage was introduced in 1908. In 1923 women candidates were admitted for election to both houses. Preferential voting was introduced for the Assembly in 1923 and for the Council 10 years later, while voting was made compulsory for the Assembly in 1926, and for the Council in 1935. Full adult suffrage was not introduced for Council voting until 1950.

Political developments after federation were marked by the rise of the Labor and Country parties, leading, in the 1920s, to the creation of a three-party pattern in place of the former Liberal–Conservative two-party system. The existence of three parties partly accounts for the apparent instability of Victorian government. Between 1901 and 1955 there were 31, mostly composite, ministries. Labor first won office, as a minority government, in 1913 and formed coalition governments on five subsequent occasions. It succeeded in winning political power only in 1952, when it gained a majority under John Cain. In 1914 the Victorian Farmers' Union was founded, sending four of its members to Parliament in 1917. In 1926 it changed its name to the United Country Party, and from 1935, led by Albert Dunstan, governed for eight years with the support of the Labor Party.

The Country Party's influential role in Victorian politics was partly explained by the unequal size of rural and urban electoral districts. Between 1924 and 1945 the unequal size of electorates made each rural vote equal to more than two urban votes. In 1953–54 an electoral redistribution, planned by a section of the Liberal Party and carried through by a Labor government, restored comparative parity between rural and urban electorates. This destroyed the crucial influence of the Country Party in the Assembly. The rebel Liberal Party was annihilated at the 1955 election. But a breakaway Labor Party, known eventually as the Democratic Labor Party, while losing Parliamentary representation, survived with Roman Catholic lay support and backing by the hierarchy. However, the Legislative Council, which retained its gerrymandered electorate despite the introduction of adult suffrage, continued to provide opportunities for political intrigue and remained the main divisive force in the Victorian political system, often vetoing bills originating in the Assembly.

Labor Party government

From 1955 to 1981 Victoria was governed by the Liberal Party under Sir Henry Bolte, a shrewd, earthy, and assertive leader, and the state's most successful 20th-century politician. His administration coincided with a lengthy period of general Australian prosperity. In the 1982 elections, the Labor Party won the majority, ending more than one-quarter of a century of Liberal government.

The major concerns of the state government are the administration of the police force and of justice, education, health services, agriculture (including marketing boards), transport, and the utility and development services associated with electricity, irrigation, water supply, ports and harbours, sewerage, forestry, and country roads. Govern-

ment departments or agencies are also vested with responsibility for the regulation of primary production, industry and commerce, labour, professional and occupational standards, education, social welfare, and public health.

The development of Victoria during the 1950s and 1960s was marked by two main trends. First, an increasing proportion of the population became urbanized. At the beginning of the century 58 percent of the people lived in urban areas. By the mid-1960s the proportion had risen to more than 85 percent. After the mid-1950s, however, "development" and private economic opportunism, fostered by liberal concessions to business and industrial enterprises, produced both crisis and expansion in Victorian affairs.

The discovery of natural gas and oil on the Gippsland Shelf in Bass Strait in February 1965, the increasing industrialization of the state, the high rate and proportion of migration which resulted in 20 percent of all Victorians having been born overseas (many in Italy and Greece), and the creation of two new universities, Monash and La Trobe, have given Victoria the basic resources, people, and skills it requires to overcome the lengthy inertia following the depression of the 1890s. The opening of the Victorian Arts Centre in August 1968 partly remedied 50 years of official cultural indifference and neglect.

Nevertheless, although the electorate's public devotion to mid-19th-century individualism and provincialism remained unshaken, the problem of commonwealth–state financial relationships, the social demands of a predominantly urban community, and the general array of problems that afflict federalism in Australia suggested that Victoria's individuality would become increasingly submerged beneath the general Australian suburban patterns of life, attitudes, culture, and thought.

In 1972 the Ministry of Conservation was created by an act of Parliament as an umbrella organization to bring together other government agencies and departments responsible for conservation and environmental protection. The ministry is responsible for achieving the aims of the Ministry of Conservation Act, the protection and preservation of the environment, and the proper management of Victoria's land and water resources.

Victoria Constitution Act

In 1975 the Victorian Parliament adopted the Constitution Act, defining the basic laws governing the state, outlining the state's relationship to the commonwealth, enumerating the powers of the state Parliament, the Supreme Court, and the executive. Prior to this, the Constitution was contained within an act passed by the United Kingdom Parliament in 1855. This document served as Victoria's constitution for some 120 years, although the state did have the power to adopt its own constitution during this period.

The National Parks Act of 1979 increased the acreage of national parks in Victoria to almost three times the previous area. Four new national parks were created under the act in 1979, and two existing parks were brought under state control. In 1980 seven additional parks were declared under the jurisdiction of the National Parks Service, bringing the total amount of park land under government control to about 3 percent of the entire state's area.

(D.B.Wa./Ed.)

WESTERN AUSTRALIA

Western Australia, by far the largest state of the Commonwealth of Australia, comprising that part of the continent lying west of longitude 129° E, covers an area of 975,920 square miles (2,527,621 square kilometres), or one-third of the total area of Australia, and about eight times the size of the British Isles. Most of the inhabitants are clustered around Perth, the capital, which has been called the most isolated city in the world. Most of Western Australia is a low tableland called the Great Western Plateau and is bounded on the east by the Northern Territory and South Australia. It has a maximum northeast-to-southwest length of 1,480 miles (2,382 kilometres) and a long, sometimes jagged, coastline of 4,350 miles, which provides few good harbours except Albany. Broome, for instance—sometimes called the Port of Pearls—is the busiest port in the northern part of the state, but tides sometimes leave ships stranded beside the jetties. More serious is the almost complete lack of any large rivers that flow all year round. There are no large lakes, and the rainfall is insufficient everywhere except in the southwest corner, the warm, sunny climate of which is considered by some the best in Australia. Western Australia is correctly said to be separated by an "ocean of sand" from the more thickly populated eastern coastal belt, but the isolation has been diminished by rapidly improving communications. The continental shelf is very narrow in the south and very broad, sloping, and deeper than usual in the north. Subsidence is evidenced by numerous coral and limestone reefs in the north and west and by granite islands in the south.

Water problems

Physical and human geography

THE LAND

Relief. The Kimberley Block (70,000 square miles) is the northernmost region, bounded by King Sound and the King Leopold Range (Mt. Broome, 3,040 feet [927 metres]). It is a plateau incised by deep valleys and with a rugged coastline cut by drowned valleys. Tides are extreme (Hanover Bay, 38 feet [12 metres]).

The Canning Basin (150,000 square miles), immediately south of the Kimberley Block, has a low-lying and uniform coast known as the Eighty Mile Beach. There is artesian water; and most of the interior is covered with sand dunes.

Southward, the Nullagine Platform (200,000 square miles) is a plateau (1,000–1,400 feet) with hills, often flat-topped, giving way to dunes in the northeast. In the north is the Pilbara Block, with flat-topped hills. The Hamersley Plateau has been uplifted and dissected, residual mountains rising to more than 3,000 feet (Mt. Bruce, 4,024 feet [1,227 metres] is one of the highest peaks in Western Australia). Rocky headlands and coral islets line the shore.

The Carnarvon Basin (50,000 square miles) has artesian water and rises gradually to 1,000 feet from the hills near Exmouth Gulf and the lowlands around Shark Bay.

Southward, the Yilgarn Block (300,000 square miles) has very poor soils, and southwestward are small basins with coal deposits. To the south is the Stirling Range (Bluff Knoll, 3,640 feet). The southwestern edge of the plateau forms the Darling Range (Mt. Cooke, 1,910 feet), which rises above the great Darling Fault (600 miles north to south).

The Swan coastal belt covers about 16,000 square miles west of the Darling fault scarp. It is mainly lowland, forming an artesian basin.

Part of the Eucla Basin, which is artesian and underlies the enormous Nullarbor Plain, is in Western Australia; it slopes to the south from about 1,000 feet, ending in cliffs 200–400 feet high.

Climate. An anticyclone approaches Western Australia from the Indian Ocean with cool, humid, southwesterly air, usually polar maritime in winter, modified in summer. Inland the air becomes drier, and the southwesterly wind is gradually replaced by a northeasterly, with all the characteristics of tropical continental air.

Temperature ranges

In summer (December–February) the anticyclones travel along their southernmost path and bring an alternation of cool southwesterlies (70°–75° F, 21°–24° C) and hot, dry northeasterlies (80°–100° F, 27°–38° C) to the southern parts of the state. Occasional high temperatures reach 110° F (43° C). Humidity is very low, and there is no rain except rare local thunderstorms. The heat over the northwest allows equatorial air to flow in, bringing heavy rains to the north and sultry days to large areas. Temperatures in the northwest vary from 75° F at night to 110° F in the afternoon. Summer rains vary from one to three inches (25 to 76 millimetres) in the south to 20–30 inches in the north, with large dry areas in the interior.

In February and March, disastrous hurricanes ("willy-willies") sometimes hit the coast between Broome and Onslow and proceed southeastward, bringing torrential rains, perhaps two to three inches in one hour. Occasional hurricanes in the northwest occur during autumn (March–May), with the last rains in the north and the gradual northward migration of the anticyclones. Humidity varies at this time, and the far southwest receives its first frontal rains (six to nine inches for the season). In winter (June–August), cool, dry air (trade winds) blows offshore in the north, whereas there are heavy frontal rains in the south, ranging from 15–20 inches on the coast to eight to 10 inches in the interior. Temperatures are 30°–60° F (−1°–16° C) at night and 60°–75° F in the afternoon in the south and about 10 degrees warmer in the north. Spring (September–November) brings the last frontal rains to the south (four to eight inches) and the highest temperatures to the parched north, before the first monsoonal rains fall in November.

The rainfall is very reliable in the far south and the far north, but in between it is variable and scarce, 58 percent of the state receiving less than 10 inches per year. Reliable rains of 30 inches or more a year fall over only 5.4 percent of the state, mostly in the southwest.

Drainage. The rivers of the Kimberley Block flow only after the summer rains but have large permanent pools. The Ord (300 miles) has a large estuary, and the Fitzroy (325 miles) crosses the King Leopold Range. The northwestern rivers flow occasionally but may be dry for several years. They include the De Grey-Oakover (370 miles), the Fortescue (340 miles), the Ashburton (400 miles), the Gascoyne (475 miles), and the Murchison (440 miles). A very large inland area has sporadic drainage at best.

The rivers that rise on the Yilgarn Block and flow west through the Swan coastal plain are short, such as the Swan-Avon (240 miles), which is dry in summer, the Murray (70 miles), and the Collie (60 miles). Farther south the flow is nearly perennial with a strong winter–spring maximum: the Blackwood (190 miles) and the Frankland (80 miles); these rivers are more or less brackish. The rivers that flow to the south are short and intermittent and run mostly in winter.

Plant and animal life. Botanically, Western Australia may be divided into three provinces, each with its own distinctive vegetation. In the northern province, the flora is rich in Malayan (Palaeotropical) forms, whereas the native element is strongly represented in the Eremean (desert) and in the southwestern provinces; in the latter there are also distinct Antarctic affinities. Long isolation without interference has resulted in the evolution of many forms and a high degree of endemism (confinement to a particular area).

There are approximately 6,800 species of plants, the carnivorous pitcher plant, the kangaroo paws, the feather flower, and the Christmas tree being of special interest. The northern province is characterized by monsoonal woodland and by rich grassland in the Kimberleys, whereas south of the Fitzroy River harsh grassland is dominant. By contrast, the southwestern province has relatively little grassland, whereas the trees and shrubs have developed strongly. There the sand-plain vegetation shows exceptional diversity and floral brilliance. The Eremean, or desert province, covers a very large area of the state.

The main woodland is in the southwestern province. Jarrah, a type of eucalyptus, grows on the sandy coastal plain, and another eucalyptus, karri, occurs in the deep southwest. Of two other eucalyptuses, the tuart is restricted to the coastal plain, whereas the wandoo grows throughout the jarrah forest area. Yet another eucalyptus, the marri, is found with the jarrah. Drought-resistant woodlands of mixed *Eucalyptus* species occur farther inland, being succeeded in drier areas by mallee shrubs (a low-growing eucalyptus).

Eucalyptus types

Fauna are also grouped in three regions. Unique forms are the burrowing frog, found where there are termites, and a member of the butterfly and moth family, often found in the nests of the ant. The numbat (banded anteater) is now restricted to the wandoo forests. The northern province has a fauna related to that of Queensland and northern Australia. In the desert areas, among other animals, is found a grotesque lizard.

THE PEOPLE

Although Western Australia comprises almost one-third of the total area of Australia, it contains less than 8 percent of the population. When the colony was founded in 1829, the population was 1,003. Growth was slow until the latter part of the 19th century; at the census of 1881, the population was 29,708, but, by 1901, it had increased sixfold to 184,124. In the 20th century, the state's average annual rate of increase was higher than that of any other Australian state. The population of 332,732 in 1921 more than quadrupled by the late 20th century. Immigration in Western Australia comes largely from other parts of Australia. Of immigrants from abroad, by far the majority were of British origin. There is no state church, but there are several hundred thousand members of the Anglican Church and the Roman Catholic Church; other major religious denominations include Methodists and Presbyterians. Western Australia is a comparatively young state; in the late 20th century about one-third of the population was under 15 years of age.

The vast majority of the population of Western Australia is urban. Much of the interior of the state is sparsely populated by white Europeans, despite its considerable pastoral capacity and the phenomenal northwest iron-ore and other mineral developments of the 1960s. Most of the increase in population occurs on and near the coasts, chiefly in the temperate southwestern corner. In this zone of settlement, Perth and its port, Fremantle, hold a key position, strengthened by major industrial development in the neighbouring Kwinana–Cockburn Sound area from the mid-1950s onward. Other settlements of the southwest are mainly small agricultural and railway centres, with the beginnings also of some industries processing primary products.

Major cities and towns are Perth Statistical Division, including Fremantle, Rockingham, and Kwinana; Kalgoorlie–Boulder; Bunbury; Geraldton; and Albany.

THE ECONOMY

Agriculture, pastoral activities, and land settlement. The boom prices after World War II transformed Western Australia's economic outlook.

Agriculture is specialized and commercial in character. The farmer expects to produce sufficient supplies for sale to enable him to enjoy a standard of living comparable to that of the town worker. Farm labour is scarce and expensive, and full advantage is taken of mechanization, particularly for the growing of grain.

The three climatic economic zones

On the basis of rainfall distribution, the state falls into three well-defined zones: (1) the Kimberley division in the extreme north; (2) an area in the southwestern corner where the rainfall ranges from 50 inches on parts of the coastline down to 11 inches on its inland boundary; and (3) the large area lying between these two, where the annual rainfall is sparse (10 inches or less) and uncertain in its incidence.

In the southern part of the northern region, sheep are raised, and sugarcane, safflower (a yellow herb), and rice have emerged as potential commercial crops. The cattle industry also benefitted greatly from "beef roads" constructed in the 1960s with commonwealth and state funds. Pearling and the production of cultured pearls are other economic activities of note in this northern region.

The vast arid zone is bad for agriculture or grazing. A few hundred million acres, however, are occupied as pastoral leaseholds, and wool is produced under conditions of extensive grazing. In places, even pastoral activities are impossible, and the only agricultural settlement in the central and northwest region is at Carnarvon, on the coast, where bananas and vegetables are grown. Farther north, in the eastern Kimberleys, cotton growing was stimulated in the mid-1960s by plans for the construction of the Ord River Dam, which was completed in 1972.

In the southwest, agriculture extends over nearly 100,000 square miles. Most of the region has little value for grazing before it is cleared of timber and scrub and fertilized with

phosphates. By the late 20th century millions of acres had been cleared with government support.

Wool is the most widespread of the state's rural industries. The sheep population of Western Australia is several tens of millions, mostly Merinos. Nearly three-quarters of the sheep are in the agricultural districts in the southwest, where grain and sheep can be raised well together. In the northwest, fewer sheep per acre can be raised, and, despite pastoral properties of between 100,000 and 1,000,000 acres, the carrying capacity averages one sheep to 20–30 acres.

Farms in the wheat and sheep districts vary considerably in size. Although when judged by European standards the grain yields of Western Australian farms are low, a favourable combination of circumstances permits high output per person, and a large wheat and sheep property is usually a highly mechanized and efficient concern.

Meat production ranks second to wool in value. The only area in which it is a single specialized enterprise on a large scale is in the northern part of the state, where cattle raising is almost the sole occupation. Uncertain prospects handicap the production of high-quality beef. In most agricultural areas, meat production is mainly a sideline to wool production, wheat growing, and dairying.

Dairying is an important occupation in the more heavily timbered country of the extreme southwest. Irrigation and drainage projects on the coastal plain south of Perth have considerably increased the productivity of the area. Several thousand acres are irrigated each year as pasture for dairy cows, but physical conditions hinder expansion.

Forest products

Almost two-thirds of the prime-forest belt has been designated as state forest for the production of timber in perpetuity. Managed under the principle of sustained yield, these forests provide 80 percent of the state's total timber output. The remainder comes from private property.

A stable and thriving sawmilling industry produces sawed timber, mainly jarrah and karri. Jarrah has a worldwide reputation for durability and general utility, and karri is equally renowned for strength and size.

The dry country savannah forests to the east of the southwestern province are slowly giving way to agriculture in the more favourable situations but still cover several million acres. Tens of thousands of acres of pines had been established by the late 20th century. In a region that suffers from a long summer drought, protection from fire is essential for the preservation of the forest.

Orchards and vineyards occupy several thousand acres. The poultry industry produces a surplus of eggs, which are exported chiefly to the United Kingdom. Cereals are widespread, and flax is grown on a small scale.

Industry and trade. The state's predominantly agricultural and pastoral economy has twice been revolutionized by mineral discoveries—gold toward the end of the 19th century (chiefly on the eastern fields of Coolgardie and Kalgoorlie) and, since the late 1950s, by the rapidly expanding exploitation of one of the world's greatest iron ore deposits in the Ashburton and Pilbara districts in the northwest. Other minerals produced include manganese, ilmenite, monazite, rutile, leucoxene, and zircon. A large bauxite deposit has been developed in the hills south of Perth, and polar salt is produced in the northwestern area of the state.

In 1967, at Barrow Island, near Onslow, one of Australia's largest commercial oil fields began operations, and natural gas was discovered at Yardarino near Geraldton. Meanwhile, after 1971, a 225-mile pipeline carried natural gas from Yardarino to the Pinjarra alumina refinery in the southwest. One of Australia's largest nickel mines is at Kambalda near Kalgoorlie.

Manufacturing is confined mainly to the Perth metropolitan area. Until mid-century, most industries were related to primary production: sawmilling, bacon curing, dairy products processing, superphosphates manufacture, and railway engineering. Thereafter, substantial industrial development in the Perth–Fremantle–Cockburn Sound (Kwinana) area included a large oil refinery and a steel mill, and these became the nucleus of an industrial complex of allied industries. These were facilitated by new public works and services, including a deepwater supplement to Fremantle harbour for oil tankers; freeways; an additional power station; and standard-gauge railway communication (opened 1968) with iron ore deposits at Koolyanobbing near Southern Cross en route to Kwinana on the coast south of Fremantle. The standard-gauge railroad was eventually extended through Kalgoorlie to link with Sydney and Brisbane in eastern Australia.

Apart from its iron ore and other mineral exports to overseas countries, chiefly Japan, a large portion of the state's trade is still with other Australian states.

The main sources of state revenue are grants from the Commonwealth government. Other income comes from land, probate, betting, and payroll taxes; license fees and stamp duties; mining royalties; and departmental and utility charges including railways. State expenditure is chiefly on public utilities, law and order, education, health (including hospitals), and other social services, supplementing those of the Commonwealth government.

Transportation. Railways were first developed in the late 19th century. Apart from certain privately owned mineral lines of standard-gauge, all railways are state-owned and most are narrow gauge (3 feet 6 inches [107 centimetres]) except for the standard-gauge transcontinental line mentioned above, which is partially owned by the commonwealth government. Road transportation, now widespread, competes with the state government railways, and, in the northern pastoral and mineral areas, airplanes are used extensively. A network of flight schedules links Perth with these areas and with highly developed ports for mineral exports such as Dampier and Port Hedland. Of particular importance is the Royal Flying Doctor Service begun in 1935 to assure the isolated stations of the northwest of regular medical attention. Perth is on international air routes.

Air services

Shipping services include: (1) the main overseas lines, which make Fremantle their first and last port of call and help to make it the largest oil-bunkering port in Australia; (2) services to other states; and (3) Western Australian coastal services, mainly northward, extending also to Java and Singapore, where a trade in fat sheep and fruit has been developed.

ADMINISTRATIVE AND SOCIAL CONDITIONS

Government. The state Parliament of Western Australia, representing each province, consists of two members per province who serve terms of six years. One-third of the Legislative Council is elected every two years, whereas the Legislative Assembly members are elected for three years. Since 1963 a property qualification no longer restricts voting for the Council, but seat distribution ensures an anti-Labor majority there. Local government is conducted through city, town, and shire councils.

Education. Two systems of primary and secondary education exist side-by-side. The state schools are under centralized administration; the private, or independent, schools, serving between one-quarter and one-fifth of the children, are controlled for the most part by church organizations, among which the Roman Catholic teaching orders predominate. Education is compulsory for children between the ages of six and 15 years who live within three miles of an established school or within two miles of a school-bus route. In rural areas children attending schools at central points are taken there daily by bus at government expense. This system is supplemented by correspondence classes for more isolated children and by the use of itinerant teachers in the remote northwest.

The University of Western Australia (founded 1911) has several thousand students in the faculties of arts, law, education, economics and commerce, science, engineering, agriculture, architecture, dental science, and medicine. The state also possessed an autonomous institute of technology, established in 1967, and several technical colleges within the technical division of the state education department, as well as technical schools and centres scattered through the state. It has secondary and primary teacher-training colleges, one agricultural college, and four residential agricultural schools. A second metropolitan university, Murdoch, was opened in 1974.

Health and welfare. Old-age, invalid, and widows' pen-

sions are generally provided by the federal government, which is also responsible for unemployment and sickness benefits, as well as maternity, child endowment, tuberculosis, repatriation, and rehabilitation allowances. The state's social services include those of municipalities and state government departments, such as education, public health, police, mental health, child welfare, and native welfare. The Western Australian Workers' Compensation Act provides for compulsory insurance by employers to compensate for losses by industrial accident, including industrial diseases.

CULTURAL LIFE

Despite Western Australia's isolation from the more sophisticated east, Perth, the capital, has an annual five-week summer festival of the arts—drama, music, ballet, films, and art. Perth boasts theatres, concert halls, "art" cinemas, and many commercial cinemas.

The Australian Broadcasting Commission invites international orchestras, soloists, and singers to Perth for celebrity concerts and youth concerts. It also sponsors the West Australian Symphony Orchestra, which also enjoys state-government assistance. Australian composers are included in the repertoire.

The University of Western Australia is also concerned with music, offering degree courses for which the Wigmore Music Library is indispensable. University music societies of various sorts flourish. Another service the university renders is the study of Aboriginal musical forms. Other musical organizations include a youth orchestra, music clubs, and choral groups. The professional Western Australia Opera Company receives both federal and state subsidies.

The state has its own Ballet Company and the Western Australian Ballet Workshop. Regular tours of country areas take place, and international troupes visit Western Australia.

Perth has an art gallery and museum, which displays works from abroad as well as contemporary native art and Aboriginal bark paintings. There is a state library, and over half the population uses the Library Board of Western Australia's services, through a steadily increasing number of local public libraries. (F.A.)

History

EXPLORATION AND SETTLEMENT

Prior to European exploration, Western Australia was inhabited by Aboriginal peoples, who migrated to this part of the continent from southern Asia some 25,000 to 30,000 years ago. These peoples were nomadic hunters and gatherers, holding undisputed possession to the region for several thousand years. Occasional Asian visitors may have reached the coast of Western Australia, but the impact of such visitors was slight and was confined to the coastal regions.

The date of the first sighting by Europeans of the western shores of Australia is uncertain. It has been suggested that both the western and northern coasts of the state may be identified with the "Java le Grande" of maps said to date from around 1550. It is possible that Portuguese navigators sighted the northwestern coast on the way to the Spice Islands, as early as 1527. The first landing of which definite records exist is that of the Dutch captain Dirck Hartog in October 1616. To commemorate his discovery, Hartog left a pewter dinner plate, with an inscription scratched into it, on the island now bearing his name. Subsequent Dutch navigators, including Abel Janszoon Tasman, charted the coast as far east as Nuyts Land, between 1618 and 1644. Later comers included William Dampier (1688 and 1699) who visited the northwest coast but reported unenthusiastically on colonization and trade. Other early adventurers and explorers were William de Vlamingh (1697–99), George Vancouver (1791), and Antoine Bruni d'Entrecasteaux (1792). None of these early explorers ventured beyond the waterless and barren coastal hills, as the insects and heat discouraged excursions inland and prospects for trade looked slim.

Dutch exploration

The colony of New South Wales, established by the English in 1788, included only the eastern half of Australia. The remainder of the continent continued to be known as New Holland, the name given to the region by Abel Tasman in 1644. The western boundary of New South Wales was extended to longitude 129° E, leaving the area of Western Australian unclaimed.

The earliest settlement was made from Sydney, then a British possession, in December 1826. Because of suspicions that the French might establish a base on the west coast of Australia, Maj. Edmund Lockyer was dispatched to take formal possession of King George Sound with a party of convicts and soldiers. This establishment remained under the control of the government in Sydney until 1831. Meanwhile, in March 1827 Capt. James Stirling had made a survey of the Swan River and offered to lead an expedition to establish a settlement on the coast. His offer was at first rejected, but in 1828 the British government, uncertain of French intentions in the area and encouraged by the interest of a number of potential investors and colonists, sent Capt. C.H. Fremantle to take possession of the unoccupied part of Australia west of 129° E. In June 1829 Stirling arrived, as lieutenant governor, and shortly afterward founded the towns of Perth and Fremantle on the Swan. Several parties of migrants followed, including those brought out by Thomas Peel, the largest single investor. They were ill-prepared for pioneering; the coastal land was poor and labour to work the large and scattered land grants was unobtainable. Many left for the eastern colonies, but the discovery of better land beyond the Darling Range led to the extension and consolidation of settlement by 1835.

Western Australian Company

Despite Peel's failure, a second syndicate, the Western Australian Company, was formed in 1840 to acquire the Lautour estate, 90 miles south of Perth, and to establish a settlement there. Between 1841 and 1843, several parties were sent out to this new settlement, Australind, but loss of confidence among those financing the scheme contributed to its failure. The explorations of George Grey (1838) and the Gregory brothers (1845–46 and 1848) led to the settlement of the Champion Bay district, north of Perth, centred on the town of Geraldton.

The continued labour shortage in the colony led to a demand that convicts be imported from the United Kingdom for employment on public works and private farms. Consequently, an order-in-council authorizing the transportation of convicts to the colony was issued in 1849. The cheap convict labour, combined with grants-in-aid from the United Kingdom, gave an impetus to expansion before convict transportation ceased in 1868. The northwestern pastoral country was opened in 1863 after further explorations by the Gregory brothers. After the abolition of transportation, a partly elective Legislative Council was appointed in 1870, during the administration of Gov. F.A. Weld. Between 1870 and 1890 valuable work in opening up more land for pastoral and agricultural settlement was performed by the officers of the Survey Department, under Malcolm Fraser and John Forrest. The brothers John and Alexander Forrest crossed the continent between Esperance and Adelaide in 1870 and between Geraldton and the overland telegraph line in 1874. Other transcontinental crossings were made by P.E. Warburton in 1873 and E. Giles in 1875. Such explorations enlarged the area of pastoral settlement but demonstrated the desert nature of much of the interior.

Alexander Forrest opened up the Kimberley district in 1879 and there, in 1885, the government geologist reported indications of the presence of gold. A gold rush to Hall's Creek followed (1885–87), but this field had been largely abandoned by the time of the proclamation of goldfields in the Pilbara district in the north in 1888, and in the Yilgarin, 200 miles to the east of Perth, in 1887. Within five years, richer finds followed at Murchison (1891), Coolgardie (1892), and Kalgoorlie (1892–94).

Progress in other directions included the construction of a railway from Perth to Albany, completed in 1889, and the beginning of another to Geraldton. The improving prospects of the colony induced a demand for responsible government. A bill enabling Queen Victoria to grant a constitution to Western Australia received the royal as-

Representative government

sent on August 15, 1890. This provided for a governor, a nominated Legislative Council, and an elected Legislative Assembly. The Legislative Council was to become elective when the population reached 60,000—a change effected in 1893.

FEDERATION AND THE STATE OF WESTERN AUSTRALIA

Sir John (later Lord) Forrest, prime minister from 1890 until his entry into federal politics on the establishment of the Commonwealth of Australia in 1901, led a legislature at first dominated by primary producing interests. The advent of thousands of gold miners accelerated the progress of Western Australia and led to the provision of such public works as the Fremantle harbour and the Perth–Kalgoorlie railway. Forrest also sponsored the construction of a pipeline through which water was pumped more than 300 miles from Mundaring (new Perth) to the arid Coolgardie goldfields. During Forrest's later years of office, however, an opposition party arose among the goldfields men—many of them from Victoria and New South Wales—who strongly supported the movement to have Western Australia join in an Australian federation.

The agricultural interests hesitated to join the new Commonwealth of Australia without receiving a pledge for the retention of their own customs dues for five years, but after Forrest had made an unsuccessful attempt early in 1900 to secure this concession from the other colonies, a referendum in July of that year showed that a majority of more than 25,000 in Western Australia favoured federation. The Constitution Act ultimately provided that Western Australia should have the right to enact its own tariff as against the sister states for five years, decreasing annually at the rate of one-fifth of the amount of the original duty until it disappeared. A gentlemen's agreement among the prime ministers also provided for the construction of a transcontinental railway.

With the decline of alluvial mining, a policy of agricultural expansion was pursued and, between 1906 and 1914, the wheat lands were extended into the 10-inch rainfall area. The growth of the new wheat belt was the greatest achievement of the first 30 years of the 20th century. The federal government's policy of protecting secondary industries tended to react unfavourably on Western Australia, which emphasized primary production, so that during the Depression at the end of the 1920s a movement arose for secession from the federation. A referendum held in 1933 favoured secession, by 138,653 votes to 70,706, but in 1935 a joint committee of the British Houses of Parliament declared itself incompetent to consider the state's petition to give effect to this move. With the revival of prosperity and the coming of World War II, during which Western Australia faced a threat of invasion by Japan, the secession movement lost force. Good wheat and wool prices and a rapidly rising population, mainly the result of federally aided immigration from the United Kingdom and other European countries, ensured postwar prosperity. The potentialities of the northern part of the state were emphasized in the 1960s by iron-ore sales to Japan and by cotton growing, which was made possible by the damming of the Ord River.

Western Australia thus entered the second half of the 20th century with an awareness of impending economic and social changes. The material progress of the 1950s and 1960s in fact proved comparable with that which had followed the gold discoveries.

Liberal–Country coalition

In 1959 a Labor administration (1953–59) was succeeded by a Liberal–Country coalition government under the leadership of David Brand. Both governments were active in promoting housing programs and industrial development with overseas capital, but the Brand administration and its minister for industrial development and for the northwest, Charles Court, were particularly identified with the Northwest mineral and Ord River schemes noted below. Confirmed in office in 1962 and 1965, the Brand administration faced a strong challenge from a Labor Party led by John Tonkin on the retirement of A.R.G. Hawke in 1968, but was returned to power, with a reduced majority, in elections in March of that year. The Liberal–Country party remained the dominant political force in the later 20th century.

Technological advances, further agricultural mechanization, and irrigation from the Ord River have opened new areas for agricultural development in the southeast.

Since the end of World War II, Western Australia has attempted to attract outside capital for industrial development. These efforts created increased volume in manufacturing beginning in the 1950s. The opening of an oil refinery and steel mill in Kwinana (1955–56) was one of the early rewards. Emphasis shifted to mineral exploitation during the 1960s with the discovery of a commercial oil field on Barrow Island, vast iron-ore deposits, and extraction of bauxite and nickel. (F.A./Ed.)

AUSTRALIAN EXTERNAL TERRITORIES

Apart from claims in Antarctica, the external territories of the Commonwealth of Australia are made up entirely of islands. They comprise innumerable small reefs, cays, and atolls between the Great Barrier Reef of Queensland and longitude 157°10′ E, and several remote and diverse islands in the Pacific and Indian oceans. These latter oceanic outposts represent the tips of submerged mountain ranges, many of volcanic origin. Those in the tropics often support fringing coral reefs, or atolls, where the sea level has risen sufficiently relative to the basic rock. Each of the Australian external territories was uninhabited when it was first annexed by Great Britain.

The composition of the Australian external territories is as follows: Norfolk Island, in latitude 29°02′ S, longitude 167°57′ E, a fertile and beautiful island famous for its indigenous pine (*Araucaria excelsa*). Also in the Pacific is the Coral Sea Islands Territory (created by act of the commonwealth on September 30, 1969), scattered over 400,000 square miles (1,000,000 square kilometres) of tropical waters and uninhabited except by observers at the commonwealth meteorological station on Willis Island (established 1921) at latitude 16° S, longitude 150° E.

The Cocos (Keeling) Islands, consisting of two atolls totalling 27 islands, are located in the Indian Ocean in latitude 12°10′ S, longitude 96°55′ E. By their transfer from Britain to the Commonwealth of Australia under the Cocos (Keeling) Act (1955), they ceased to be part of the then British colony of Singapore.

Christmas Island, in latitude 10°30′ S and longitude 105°40′ E, is a source of high-grade rock phosphate. It was transferred to Australia by Britain, under the Christmas Island Act (1958). The minute, uninhabited Ashmore and Cartier Islands, also in the Indian Ocean, are defined by the Ashmore and Cartier Islands Acceptance Act, 1933. Cartier is in latitude 12°32′ S, longitude 123°32′ E, and the Ashmore group of three small exposures lie about 30 miles (48 kilometres) northwest of Cartier.

The claim comprising Australian Antarctic Territory consists of all islands and territories, other than Adélie Land, situated south of the 60th parallel of south latitude and between the 45th and 160th eastern meridians. This territory, over which the British government formerly asserted the crown's sovereign rights, was transferred to Australia under the Australian Antarctic Territory Acceptance Act (1933). Heard Island, in latitude 53°06′ S, longitude 73°30′ E, is perpetually ice-capped; with the nearby McDonald Islands, it was transferred from Britain to Australia in 1947. The Heard and McDonald Islands Act (1953) defines Australia's claim.

A radio network provides channels for administrative and commercial interests and for an extensive exchange of meteorological and other scientific information. Physical contact between the commonwealth and its territories varies from periodic air services from the appropriate Australian capitals to the annual relief of Antarctic stations by vessels strengthened for ice.

Administrative responsibilities

The commonwealth Department of Home Affairs provides administrative services for Norfolk Island, the Cocos (Keeling) Islands, Christmas Island, the Coral Sea Islands, and the Ashmore and Cartier group, except in matters of defense and civil aviation, for which other departments are responsible. The Department of Administrative Services provides policy advice with regard to the phosphate enterprise of Christmas Island, and, in general, the Department of Primary Industry is concerned with fishing rights in the Australian 200-mile zones surrounding the territories, which effectively increase their areas of economic interest very considerably. For the inhabited territories, an administrator or official representative is appointed by the governor general of Australia to assist the government of the territory on behalf of the commonwealth. By act of self-determination the Cocos (Keeling) Islands voted for integration with Australia in 1984. In 1968 and 1975, respectively, the independence of the former external territories of Nauru and Papua (as Papua New Guinea) was attained.

The commonwealth Department of Science and the Environment is responsible for the administration of Australia's Antarctic interests and the enactments defined by the Antarctic Treaty Act (1959), the Australian Antarctic Treaty Acceptance Act (1933), and the acts relating to Heard and the McDonald Islands. The Department of Science and the Environment, through its Antarctic Division, oversees the logistics of the annual relief operations of Australia's scientific bases in Antarctica (for which such departments as that of national development supply certain scientific and technical personnel) and for their general function and administration.

Norfolk Island

PHYSICAL AND HUMAN GEOGRAPHY

Norfolk Island, with a total area of about 13.3 square miles (34.5 square kilometres), is volcanic in origin and rises precipitously from an extensive submarine ridge connecting the North Island of New Zealand with New Caledonia and thence, in general, with Melanesia and New Guinea. It has a mean altitude of 360 feet (110 metres), with two high points, Mt. Pitt and Mt. Bates, just exceeding 1,000 feet (300 metres). To the south, and close inshore, are two small, uninhabited islands—Nepean, of limestone, and Phillip, which is volcanic, eroded, and almost denuded of vegetation by introduced animals. Norfolk Island is 1,041 miles (1,676 kilometres) east-northeast of Sydney and 660 miles (1,063 kilometres) north of Auckland, N.Z., in latitude 29°02′ S and longitude 167°57′ E. The climate is pleasant and equable, with a mean rainfall of 53 inches (1,350 millimetres), and temperatures ranging between 81° F (26.7° C) and 50° F (10.3° C). Although most of the island has been cleared for cropping and pasture, the once dominant Norfolk Island pines (*Araucaria excelsa,* or *heterophylla*) are still a notable feature of the landscape. Kingston, in the south, is the main settlement and administrative centre. Norfolk Island possesses a considerable flora and fauna. To 173 species of native vascular plants have been added some 250 introduced species. Fifty plants are endemic to the island. Apart from 55 indigenous species of birds (14 endemic) the fauna includes one species each of gecko, bat, and turtle; plentiful fishes; and numerous invertebrates. For some decades whaling was a hazardous island industry.

Tourism is the dominant economic activity on Norfolk Island. A large segment of the population derives its income from various aspects of the tourist industry, such as hotels and duty-free goods. Sea and air services are available on a regular basis. The soil is fertile and allows a variety of tropical and semitropical cultivation. There are, however, only a few crops exported from the island, and most foodstuffs must be imported from the mainland. Fish are plentiful, but the lack of a sheltered harbour has prevented any exploitation of this resource.

Imports to Norfolk Island have risen since World War II. The majority of the imports come from Australia, New Zealand, and the Pacific Islands. Goods manufactured and produced on Norfolk Island can be exported to Australia duty free. Passenger service and airfreight are available between Sydney and Norfolk Island while oceangoing shipping services operate at monthly intervals between the island, the mainland, and New Zealand.

Local government

Norfolk Island has a Legislative Assembly and Executive Council, which have ministerial duties and run many of the affairs of the island. The Supreme Court is the island's highest judicial authority and the original jurisdiction for criminal and many civil cases. The Court of Petty Sessions is the lower court for criminal and civil jurisdiction.

Like the other states and territories of the commonwealth, education is free and compulsory for children between six and 15 years of age. Island schools cover the kindergarten years through grade 10. Higher education must be sought at mainland colleges and universities. A limited number of vocational apprenticeships are also available.

HISTORY

Following his discovery of the island, which he found uninhabited, Capt. James Cook, in 1774, wrote enthusiastically on the indigenous "spruce pines," which he believed to be of great potential value as ships' masts, and on the local flax (*Phormium tenax*). Within a few weeks of the settlement of New South Wales by Gov. Arthur Phillip, in February 1788, Norfolk Island was also settled by a small party, including 15 convicts, under the command of Lieut. Philip Gidley King. Its purpose was to forestall any possible French occupation, to propagate the native flax, and to exploit the timber. After 26 years as a British convict colony, with a maximum of 1,100 convicts and free settlers, the island was abandoned. The population was withdrawn, mainly to Van Diemen's Land, renamed Tasmania in 1856.

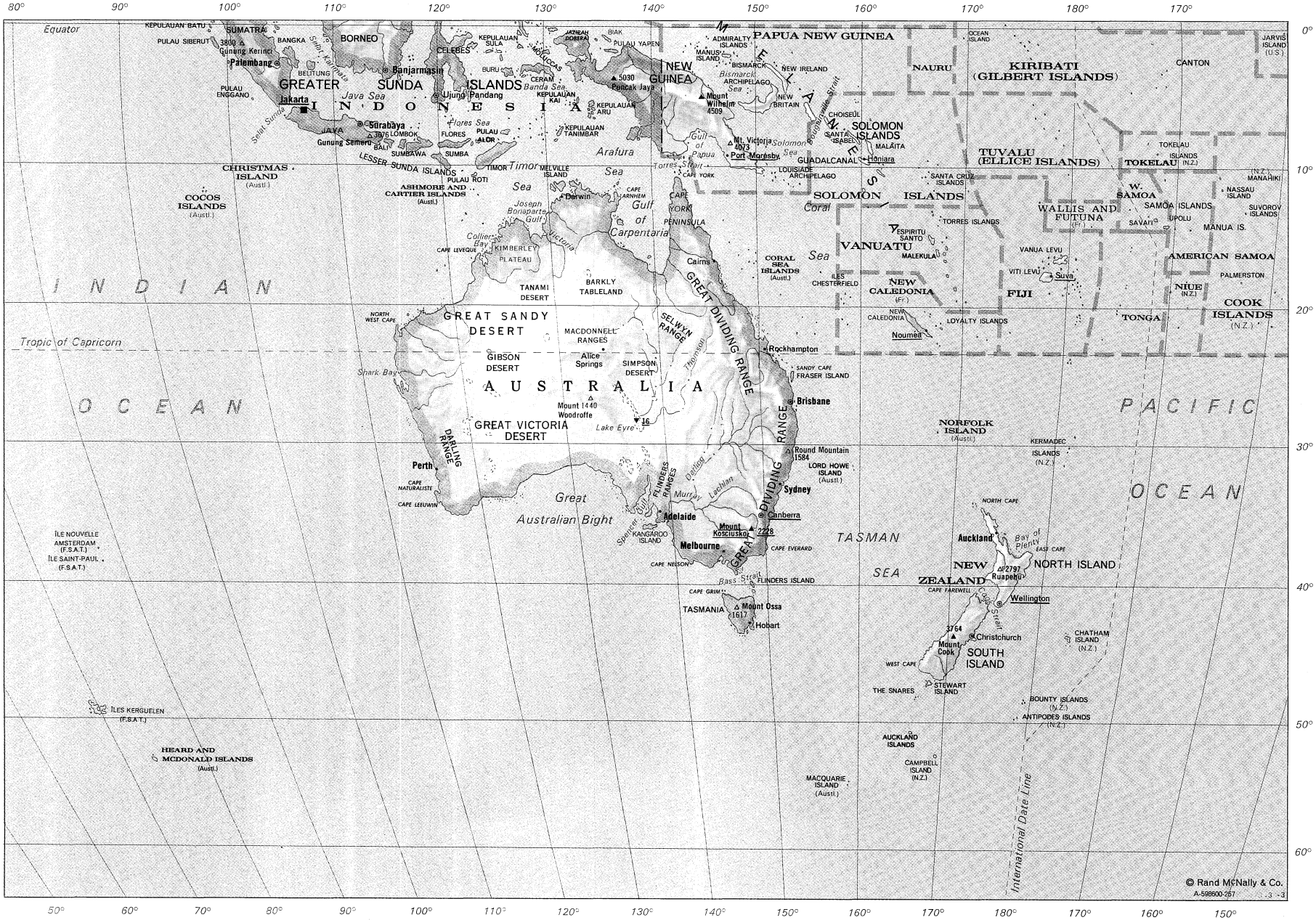
The next phase of Norfolk Island history lasted from June 1825 to June 1856. Reestablished as a penitentiary for the reception of the most desperate criminals from the British convict settlements in Australia, it became a place of merciless discipline and punishment. Fear possessed both prisoners and jailers, and in the three decades of convict occupation and human misery there were few redeeming aspects of settlement or administration. A number of substantial ruins and a few colonial buildings intact in stone remain, giving a poignant and picturesque note to the gentle landscapes.

Although the reasons for again evacuating all convicts to Tasmania resulted from difficulties of supervising the administration as much as that of the convicts, the contemporary need for better accommodation of the Pitcairn islanders contributed reasons for the change. Norfolk Island seemed a most suitable place for resettling the descendants of the mutineers of the "Bounty."

In April 1789, when Capt. William Bligh of HMS "Bounty" and a number of loyal officers had been set adrift, ultimately to reach Timor, the Pitcairners' history began. While most of the mutineers stayed on Tahiti, nine, with a number of Tahitians, both men and women, sailed on, by way of Tonga and Fiji, and eventually reached Pitcairn Island, latitude 25°04′ S, longitude 130°06′ W, a desert island discovered by Philip Carteret in 1767. In the early years violence, bloodshed, and illness ultimately left one European male, Alexander Smith, with a number of Tahitian women and half-caste children surviving in a strangely tranquil and pious community. Smith had changed his name to John Adams. The descendants of the mutineers, with a few adventurous newcomers, comprised 194 persons when, after 67 years and many vicissitudes, they were transported to Norfolk Island, arriving there in the "Morayshire" on June 8, 1856.

Not all of the islanders settled happily on Norfolk; two small parties returned to Pitcairn. Strong Polynesian characteristics and residual traits and customs persisted and are considered to be an influence among some island people to this day. A distinctive society developed, based on neighbourliness, self-help, barter, and respect for authority. Specific ethnic and religious festivals continue to be celebrated. Of the total population, perhaps one-third may claim lineal descent from the Pitcairners. Immigration is now strictly controlled.

Norfolk Island was originally made "a distinct and sep-



AUSTRALIAN EXTERNAL TERRITORIES

Size of symbol indicates relative size of town • ⊙ ■

Elevations in metres

arate settlement" from the mainland colonies on June 24, 1856. Rapidly the islanders established their own systems of land tenure and society generally. In 1897 Britain conferred administrative status on the governor of New South Wales, though the island still remained a separate British colony.

Norfolk Island Act

In 1913, under the Norfolk Island Act (effective 1914), the colony became a territory of the commonwealth, but the precise constitutional relationship with Australia was never the subject of complete agreement. The independent nature of the original Pitcairners has often seemed dominant in island aspirations. An airfield constructed on the island during World War II gave Norfolk a close, reliable link with the outside world and the means to support steady growth of facilities to match its potential attraction for tourists. There was inevitable trading in ideas: Norfolk became aware of modern amenities. Australian enterprise rapidly moved into tax-free Norfolk Island, and many hundreds of new companies were registered there. Australia also discovered that its Norfolk territory required the fuller application of certain conventions of the International Labour Organisation.

With such emphases and with visitors to this beautiful tax-free island reaching 20,000 or more annually (a third of them from New Zealand), a royal commission was appointed in 1975, under Sir John Nimmo, to report on the future status of Norfolk Island. His report (1976), concerned with the territory's constitution, its relationship to Australia, and the most appropriate forms of administration, was exhaustive. It revealed two opposing views: that Norfolk Island should be virtually absorbed into the mainland social and political system as a detached piece of Australia, a new electorate with all the privileges and duties of such; and that it should move at least some distance toward self-government. After long negotiations, a government of the second type for Norfolk Island, becoming operative in August 1979, was established by Australian commonwealth legislation. The Norfolk Island Act, 1979, provided for a nine-member Legislative Assembly elected triennially, an administrator, and an executive council holding ministerial portfolios. The duly elected assembly has the formal right to legislate on the raising of revenue and on other island functions, such as supply, commerce, immigration, and tourism; it may not act in matters of defense, coinage, or certain property transactions. The bill also reserves commonwealth powers of veto and of delaying the implementation of Norfolk Island acts. The recommendations for absorption were set aside, and the course of the territory headed toward internal self-government as a territory under the authority of the commonwealth.

The elections of 1979 were held under a proportional representation system. The legal system includes a Supreme Court, held on the mainland or with visiting mainland judges, and a Court of Petty Sessions under the jurisdiction of local magistrates.

Various Australian government departments are concerned with Norfolk Island as a commonwealth territory. In general, the responsible authority is vested in the Department of Home Affairs. The Department of Primary Industry controls the Australian Fisheries Council, being the advisory body on all aspects of administration and management of Australian fisheries, including those of Norfolk Island. The declaration of Norfolk's 200-mile economic zone, extending its fishing grounds to 124,000 square miles (320,000 square kilometres), is of considerable commercial interest not only to Australia but to its Pacific neighbours.

From the mid-1960s the economy of Norfolk Island changed from one based on subsistence farming, seed production, grazing, and local fishing, with a small tourist industry, to one principally dependent upon and catering to an annual influx of upwards of 20,000 visitors. Island agricultural produce cannot sustain demand, and most food must therefore be imported. There is considerable island revenue from liquor sales (a government monopoly), from sales of postage stamps to collectors around the world, and from customs duty and company fees. Australian currency is legal tender.

Christmas Island

PHYSICAL AND HUMAN GEOGRAPHY

At several points exceeding 1,100 feet (335 metres) in altitude, Christmas Island, with an area of 52 square miles (135 square kilometres), is situated in the Indian Ocean, the summit of an oceanic mountain surrounded by depths of 6,000 feet (1,800 metres). The abyssal Java Trench (24,440 feet, or 7,450 metres, at its deepest) is to the northeast. Abrupt cliffs comprise much of the coastline, but there are sand and coral beaches. The rock is mainly limestone stratified with old volcanic flows, with valuable deposits of lime near the chief port at Flying Fish Cove, below the northern Headridge Hill, and on the Southern Plateau. The island is 815 miles (1,312 kilometres) from Singapore, 220 miles (360 kilometres) from Java Head, and 1,630 miles (2,623 kilometres) from Fremantle, Western Australia. Its climate is salubrious and warm (64°–86° F, or 18°–30° C), with little seasonal variation but with a tropical rainy season between November and April. Rich forest clothes the island, and both flora and fauna are varied and numerous. The latter includes birds, small reptiles, crustaceans, insects, spiders, scorpions, and centipedes. The birdlife, with several endemic species, includes a large number of seabirds nesting on the island. About one-seventh of the flora, numbering 200 species of flowering plants, is endemic. Freshwater is available from springs and wells. Subsistence cropping and fishing are undertaken on a small scale, but virtually all food is imported. Charter flights link Christmas Island with several cities. Radioteletype and the usual overseas radiotelephone services are available. There are several major roads on the island, which has about 800 privately owned vehicles. A 20-kilometre railway transports phosphate from the southern plateau to the port and settlement on Flying Fish Cove. A spur line serves the western part of the island. The economy of the island is based entirely on phosphate mining.

Java Trench

An administrator of Christmas Island is appointed by the Australian governor general, with a staff for secretariat, education, police, radio, and harbour duties. There is also a full-time conservation officer. The judiciary consists of a Supreme Court, District Court, Magistrate's Court, and Children's Court. The Christmas Island Area School provides preschool, primary, and secondary education while trade instruction is available in the Island Technical School.

HISTORY

First sighted in 1615, the island was named on Christmas Day 1643 by Capt. William Mynors of the British East India Company. The British "Challenger" Expedition (1872–76) collected phosphate specimens from Javanese islands, and the expedition naturalist, John Murray, predicted massive deposits of guano on nearby Christmas Island. It was annexed by Great Britain on June 6, 1888. A 99-year lease, granted in 1891 to George Clunies-Ross of Cocos and John Murray, to mine phosphate and cut timber, was transferred six years later to the Christmas Island Phosphate Company Ltd., largely owned by the former lessees. Christmas Island was incorporated in the British crown colony of the Straits Settlements (capital, Singapore) in 1900. In 1942 many residents were evacuated, and the island was subsequently occupied by the Japanese, though phosphate was not mined by them. After the war (1948) the interests of the Christmas Island Phosphate Company were acquired by the governments of Australia and New Zealand. In 1958, when the island became an Australian territory, a new agreement ratified the existing legislature, and it was arranged that the Christmas Island Phosphate Commission (established 1949), on behalf of both governments, should quarry and ship the valuable deposits. Reserves of phosphate of various grades have been estimated at a few hundred million metric tons, but first-grade material is much less plentiful.

The population is mainly Chinese and Malay labourers recruited from Malaysia, Singapore, and the Cocos (Keeling) Islands, working for the commission, with Australian and British management personnel, and the island admin-

Migration of workers

istration. About one-third of the population arrives or departs each year. For long-term Asian residents, legislation has provided opportunities of resettlement in countries as chosen, provided the emigrants conform to immigration requirements. Most have chosen resettlement in Western Australia. Between December 1976 and June 1978, 578 Christmas islanders were permanently resettled under the scheme.

Cocos (Keeling) Islands

PHYSICAL AND HUMAN GEOGRAPHY

The two low coral atolls known together as the Territory of Cocos (Keeling) Islands, lie in the Indian Ocean in latitude 12°10′ S and longitude 96°55′ E, approximately 1,720 miles (2,768 kilometres) northwest of Perth and 2,290 miles (3,685 kilometres) almost due west of Darwin. They comprise 27 small coral islands totalling about 5½ square miles (14 square kilometres) in area. One atoll, North Keeling, is a single island standing about 15 miles (24 kilometres) north of the main lagoon, which is surrounded by the numerous exposures of the South Keeling Islands. Of these, West Island is the largest, measuring 6¼ miles (10 kilometres) long and one-third of a mile (0.5 kilometre) wide. The other principal islands of the South Keelings are South, Home, Direction, and Horsburgh. Coral sand or clinker raises the larger islands to heights of 10 to 20 feet (three to six metres). They shelve steeply to the surrounding sea and to the reef that protects and, from its debris, nourishes them.

Vegetation is extensive on all islands, with valuable coconut palms predominating over scrubby jungle along the outer, seaward margins. On North Keeling and Horsburgh islands there are areas of coarse grass. Large numbers of seabirds—terns, gannets, and petrels—thrive on the islands. Some land birds, white-eyes, Java sparrows, and thrushes, have reached Cocos (Keeling) from Indonesia.

Climate

The climate is equable, the range of temperatures throughout the year varying between 72° and 90° F (22° and 32° C). Rainfall, over five years, averaged 91 inches (2,312 millimetres) a year; relative humidity averages about 75 percent. In 1968 severe damage was caused by a tropical cyclone. No streams or surface water occur, but on the larger islands ample water may be obtained from wells sunk to depths of three to five metres.

The Cocos Islands Council is responsible for a wide range of functions in the Home Island village area. It also advises the administrator on matters concerning the islands and comments upon proposed legislation. The islands operate their own postal service with the revenues used for the benefit of the citizens.

HISTORY

Capt. William Keeling is credited with the discovery of the northern island in 1609, but the group remained uninhabited until 1826 when the Englishman Alexander Hare settled. A year later he was joined by John Clunies-Ross, who brought a number of Malays to assist in harvesting the coconuts for copra. Hare departed in 1831, and Clunies-Ross increased his holding and his labour force and commenced a powerful association which, in changing circumstances, has since been maintained. Descendants of the original workers, who, though often called Cocos Malays, included natives of several Asian countries and some East Africans, are still living in the islands. In the years 1948–51 more than 1,600 of the population, which had grown beyond the capacity of the islands to sustain, were assisted to migrate. The majority were resettled on estates in what is now Sabah, Malaysia. Some went to Singapore or Christmas Island.

Urged by Clunies-Ross, Britain annexed the islands in 1857, when John George Clunies-Ross had succeeded his father. A formal declaration to this effect was made by Capt. Stephen Fremantle of HMS "Juno." Subsequently, supervisory roles were bestowed successively on Ceylon (1878) and the Straits Settlements (1886). In 1886 also, Queen Victoria made her historic grant in perpetuity to George Clunies-Ross of all land above high-water mark, subject to a formal right of repossession by the crown. So began the benign but authoritarian rule, lasting more than 90 years, of the "kings of the Cocos."

The islands, as a site of a vital cable station and as a military base, proved to be of strategic importance and saw enemy action in both world wars. In 1945, during the allied military occupation, an airstrip was constructed on West Island. In 1951 the purchase of land on West Island by the Australian commonwealth was negotiated, the airstrip was reconstructed for civilian use, and until 1967 Cocos (Keeling) was used as a staging point on flights between Australia and Africa.

Australian sovereignty

From the then British colony of Singapore, Australia accepted territorial rights over Cocos (Keeling) on November 23, 1955, effected by an order in council made by Queen Elizabeth II and under reciprocal rights of the United Kingdom and Australia. With the Cocos (Keeling) Islands Act, 1955–75, of Australia, all Cocos Malays born after November 23, 1955, are Australian citizens and entitled to residence on the mainland; former British subjects also had the option of Australian citizenship. Numbers of Cocos islanders have accordingly settled in Australia, mainly in Western Australia. The Cocos Malay group live on Home Island; the remainder of the population are mainland-recruited employees of the territory administration, and their families, living on West Island.

Clunies-Ross interests purchased by the commonwealth

Until 1984 the Territory of Cocos (Keeling) Islands was a non-self-governing territory to which Chapter XI of the United Nations Charter applied. In September 1978, after considerable negotiation, the Australian commonwealth purchased all of the Clunies-Ross commercial plantation and island enterprises; it first leased the property to the Cocos Islands Co-operative Society Limited at a nominal rental and subsequently transferred the ownership of most of the land to the Cocos Islands Council. A high security animal quarantine station has been established to safeguard Australian herds while allowing the entry of genetically superior animals. With plantation palms numbering some 350,000, copra is the sole cash crop and thus the territory's mainstay. The Local Government Ordinance of 1979 established the Cocos (Keeling) Islands Council, which provides local government. In 1984, following an act of self-determination by the Cocos Malay people witnessed by a UN Mission, the Cocos (Keeling) Islands became integrated with Australia.

Coral Sea Islands Territory

Australia's Coral Sea Islands include the Flinders Reefs, Herald Cays, Holmes Reefs, Moore Reefs, Bougainville Reef, Osprey Reef, Willis Group, and others in the northwest of the Coral Sea, and a southeasterly group embracing Frederick Reefs, Saumarez Reefs, and Cato Island. All except Willis Group are uninhabited, but they are occasionally visited by scientists, fishermen, and prospectors for oil or minerals.

British naval vessels—HMS "Cato" (1803), "Frederick" (1812), and "Herald" (1854–60)—discovered and surveyed these low coral islands, all of which are either the tops of volcanic foundations or the dissected remnants of submarine plateaus. Spread over 400,000 square miles (1,000,000 square kilometres) between the Great Barrier Reef and longitude 157°10′ E, and south of latitude 12° S, the islands were declared a Territory of Australia by the Coral Sea Islands Act, 1969, with amendments, 1973.

Coral Sea Islands Act

The territory is administered by the Department of Home Affairs. Several automatic weather stations supplement information from the manned Australian Bureau of Meteorology station (1921) on Willis Group, some 300 miles (500 kilometres) east of Cairns, Queensland.

Ashmore and Cartier Islands

Similarly administered by the Department of Home Affairs are the remote, uninhabited Ashmore and Cartier Islands, the Australian territory nearest to Indonesia. Cartier Island was discovered by Captain Nash, and named after his vessel, in 1799 or 1800. It is about four feet (one metre) above sea level and little more than an acre (0.4 hectare) in extent. On Ashmore Reef there are three small

coral islands—West, Middle, and East—totalling about 230 acres (93 hectares). Turtles are plentiful, and at one time guano was dug, though supplies were soon exhausted. Ashmore Reef, rising from the broad, shallow sea covering the Sahul Shelf, was annexed by Britain in 1878, Cartier in 1909. Australia officially agreed to accept sovereignty in 1933 and an amendment of 1938 annexed the islands to the Northern Territory. There is an automatic weather station on West Island.

Heard and McDonald Islands

Heard Island was occupied continuously from 1947 to 1955 by the Australian National Antarctic Research Expeditions (ANARE), the scientific program then being transferred to stations on the Antarctic mainland. The deserted station on Heard, latitude 53°06′ S, longitude 73°30′ E, is visited occasionally by relief expeditions bound for the more southerly bases and has been used by U.S. scientists and technicians concerned with satellite tracking and other space programs. The island was first sighted in 1833 but was lost and rediscovered once or twice before being named, in 1855, after an American captain. After 1855 many teams exploited the island's elephant seals and penguins for their oil. The island is about 180 square miles (466 square kilometres) in area, and its dormant volcano, Big Ben, rises to 9,005 feet (2,745 metres). The island has been extensively explored by Australian expeditions. The nearby McDonald Islands have not been occupied. Like the lower parts of Heard Island, they are sparsely covered with cushion plants, rough tussock grass, mosses, and lichens, and they are visited by large numbers of elephant seals, leopard seals, and penguins. Several kinds of petrels, albatrosses, and skuas that breed on Heard Island probably also inhabit the McDonalds.

Australian Antarctic Territory

Under the terms of the Antarctic Treaty (1959), until virtually the end of the 20th century, territorial claims and ambitions in the Antarctic were to be held in abeyance. The concept of a United Nations or other international control of the whole continent in perpetuity has been advanced from time to time. Scientific cooperation between the treaty nations has proved successful, and the BIOMASS (Biological Investigation of Marine Antarctic Systems and Stocks) program has engaged the research of the Antarctic Treaty signatory nations. Any real or potential exploitation of Antarctic resources, or of exclusive economic zones, however, has revived consideration of national territorial claims and fishing rights. The focus of world interest on the possible exploitation of hydrocarbon deposits and (though whaling has declined) on the marine wealth of the Antarctic seas in krill trawling and fisheries also has retarded activities pertaining to international control.

Australian Antarctic research

Australia has maintained continuous Antarctic scientific research and an extensive program of physical exploration by sea, land, and air since 1954. Before that Australian explorers had been associated with most of the major British expeditions and played a leading role in some. Apart from the long-established sub-Antarctic base on Macquarie Island (administratively part of the state of Tasmania), in latitude 54°28′ S, longitude 158°57′ E, Australia maintains three continental bases: Mawson station (1954) on MacRobertson Land in latitude 67°36′ S, longitude 62°53′ E; Davis station (1957) on Ingrid Christensen Coast in latitude 68°35′ S, longitude 77°58′ E; and Casey station (1969) in Wilkes Land in latitude 66°17′ S, longitude 110°32′ E. Casey replaced the U.S. base at Wilkes station, built for the International Geophysical Year in 1957 and occupied by Australia during 1959–69.

(J.M.B./Ed.)

BIBLIOGRAPHY

Physical geography: AUSTRALIA. DEPARTMENT OF NATIONAL DEVELOPMENT, *Atlas of Australian Resources* (1st series 1953–60, 2nd series 1962–75, 3rd series 1980–), an official compendium of separate maps on geology, geography (physical and human), and resources (natural and industrial) of Australia, with commentaries; G.W. LEEPER (ed.), *The Australian Environment,* 4th ed. rev. (1970), a brief outline, mainly on land forms, climates, soils, water and irrigation vegetation, crops and pastures, and animal production; R.O. SLATYER and R.A. PERRY (eds.), *Arid Lands of Australia* (1969), symposium papers on many aspects of arid Australia and particularly on land and resources management; V. SERVENTY, *Australia's National Parks* (1969), a beautifully illustrated book on the more important national parks, arranged geographically and dealing with fauna, flora, and geography.

D.A. BROWN, K.S.W. CAMPBELL, and K.A.W. CROOK, *The Geological Evolution of Australia and New Zealand* (1968); D. HILL and A.K. DENMEAD (eds.), "The Geology of Queensland," *J. Geol. Soc. Aust.,* vol. 7 (1960); G.H. PACKHAM (ed.), "The Geology of New South Wales," *J. Geol. Soc. Aust.,* vol. 16, pt. 1 (1969); L.W. PARKIN (ed.), *Handbook of South Australian Geology* (1969); A. SPRY and M.R. BANKS (eds.), "The Geology of Tasmania," *J. Geol. Soc. Aust.,* vol. 9, pt. 2 (1962). The first of these is a general text for students, on regional and historical geology, divided into chapters on systems from the Precambrian to the Quaternary. The others are compilations of known geological data for four of the six Australian States. Further information may be obtained from periodical publications of the Bureau of Mineral Resources (Canberra), the State Geological Surveys, and the Geological Society of Australia.

C.F. LASERON, *The Face of Australia,* 2nd ed. (1957), a readable account of the evolution of Australia's scenery, although unsound in some areas and now outdated; D.N. JEANS (ed.), *Australia: A Geography* (1977), a more recent overview of Australian geography; C.R. TWIDALE, *Geomorphology, with Special Reference to Australia* (1968), a general and fairly elementary text, containing numerous references to Australian features and examples of well-known forms. Perhaps the best technical accounts are contained in the series of texts published by the ANU Press, Canberra: E.C.F. BIRD, *Coasts* (1969); J.L. DAVIES, *Landforms of Cold Climates* (1969), and, with M.A. WILLIAMS, *Landform Evolution in Australasia* (1978); C.D. OLLIER, *Volcanoes* (1969); C.R. TWIDALE, *Structural Landforms* (1971); J.N. JENNINGS, *Karst* (1971); J.A. MABBUTT, *Desert Landforms* (1977); and I. DOUGLAS, *Humid Landforms* (1977).

GEORGE BENTHAM, *Flora Australiensis,* 7 vol. (1863–78); S.F. BLAKE and A.C. ATWOOD, "Geographical Guide to the Floras of the World," pt. 1, *Misc. Publs. U.S. Dep. Agric. 401* (1942); W.E. BLACKWELL and B.J. GRIEVE, *How to Know Western Australian Wildflowers,* 3 pt. (1954–65); N.T. BURBIDGE, "The Phytogeography of the Australian Region," *Aust. J. Bot.,* 8:75–211 (1960); D.W. GOODALL (comp.), "Bibliography of Statistical Plant Sociology," *Excerpta Bot.,* sect. B., 4:253–316 (1962).

A. KEAST, R.L. CROCKER, and C.S. CHRISTIAN (eds.), *Biogeography and Ecology in Australia* (1959); W.D.L. RIDE, *A Guide to the Native Mammals of Australia* (1970); S. BREEDEN and K. BREEDEN, *The Life of the Kangaroo* (1966); N.W. CAYLEY, *What Bird Is That: A Guide to the Birds of Australia,* 6th ed. rev. and enl. by A.H. CHISHOLM *et al.* (1973); E. WORRELL, *Reptiles of Australia,* 2nd ed. (1970); W.R. EASTMAN and A.C. HUNT, *The Parrots of Australia* (1966); T.C. ROUGHLEY, *Fish and Fisheries in Australia,* rev. ed. (1966); I.M. MACKERRAS (ed.), *The Insects of Australia: A Textbook for Students and Research Workers* (1970).

JOHN MCANDREW (ed.), *Geology of Australian Ore Deposits,* 2nd ed. (1965); I.R. MCLEOD (ed.), *Australian Mineral Industry: The Mineral Deposits,* Bureau of Mineral Resources, Geology and Geophysics, Bull. No. 72 (1965); Z. KALIX, L.M. FRASER, and R.I. RAWSON (eds.), *Australian Mineral Industry: Production and Trade, 1842–1964,* Bureau of Mineral Resources, Geology and Geophysics, Bull. No. 81 (1961); H.G. RAGGATT (ed.), *Fuel and Power in Australia* (1969).

Human geography: WILLIAM K. HANCOCK, *Australia* (1930); J.D. PRINGLE, *Australian Accent* (1958, reissued 1978); DONALD HORNE, *The Lucky Country* (1964, reissued 1978); GEOFFREY BLAINEY, *The Tyranny of Distance* (1966, reissued 1975); *The Australian Encyclopaedia,* 4th ed. rev. (1983).

JOHN QUICK and ROBERT R. GARRAN, *The Annotated Constitution of the Australian Commonwealth* (1901, reprinted 1976); PERCY E. JOSKE, *Australian Federal Government,* 3rd ed. (1976). LESLIE F. CRISP, *The Parliamentary Government of the Commonwealth of Australia,* 3rd ed. (1961); JOHN D.B. MILLER, *Australian Government and Politics* (1954, reissued 1971); GEOFFREY SAWER, *Australian Government Today,* 12th ed. (1977); JAMES JUPP, *Party Politics: Australia 1966–1981* (1982).

ERNEST A. BOEHM, *Twentieth Century Economic Development in Australia* (1979); JAMES O.N. PERKINS, *Australia and the World Economy,* 3rd ed. (1979); ROBERT CATLEY and BRUCE MCFARLANE, *Australian Capitalism in Boom and Depression* (1981); L.R. WEBB and R.H. ALLAN (eds.), *Industrial Economics* (1982).

ALAN S. WATT, *The Evolution of Australian Foreign Policy: 1938 to 1965* (1967); *Australia in World Affairs* (quinquennial). GEOFFREY BLAINEY, *Triumph of the Nomads: A History of*

Ancient Australia (1975); HENRY REYNOLDS, *The Other Side of the Frontier* (1981); CHARLES D. ROWLEY, *A Matter of Justice* (1978). *Official Year Book of Australia* and *Australia Handbook* (both annual).

(J.D.Pr.)

History: The best general accounts of traditional Aboriginal life are A.P. ELKIN, *The Australian Aborigines* (1964); R.M. and C.H. BERNDT, *The World of the First Australians,* 3rd ed. (1969); F.D. MCCARTHY, *Australia's Aborigines: Their Life and Culture* (1957); R.M. and C.H. BERNDT (eds.), *Aboriginal Man in Australia* (1965); and R.M. BERNDT (ed.), *Australian Aboriginal Anthropology* (1970). H. SHEILS (ed.), *Australian Aboriginal Studies* (1963), also gives a general coverage focussed on research problems. Most of these volumes include sections on change; but see especially M. REAY (ed.), *Aborigines Now* (1964); I.G. SHARP and C.M. TATZ (eds.), *Aborigines in the Economy* (1966); W.E.H. STANNER, *After the Dreaming* (1969); D.E. HUTCHINSON (ed.), *Aboriginal Progress: A New Era?* (1969); G. BLAINEY, *Triumph of the Nomads: A History of Aboriginal Australia* (1976); and, related to Aboriginal policy and practice, C.D. ROWLEY, *The Destruction of Aboriginal Society* (1970), *Outcasts in White Australia* (1971), and *The Remote Aborigines* (1971). A large series of monographs on various aspects of Aboriginal life has been produced by the Australian Institute of Aboriginal Studies.

Separate and detailed studies of Aboriginal societies are by W.L. WARNER, *A Black Civilization* (1937); C.W.M. HART and A.R. PILLING, *The Tiwi of North Australia* (1960); M.J. MEGGITT, *Desert People: A Study of the Walbiri Aborigines of Central Australia* (1962); L.R. HIATT, *Kinship and Conflict: A Study of an Aboriginal Community in Northern Arnhem Land* (1965); and R.M. and C.H. BERNDT, *Man, Land and Myth in North Australia: The Gunwinggu People* (1970).

Among the classic studies of the pre-anthropological era are those by R.B. SMYTH, *The Aborigines of Victoria,* 2 vol. (1878); J.D. WOODS (ed.), *The Native Tribes of South Australia, Comprising the Narrinyeri* (1879); E.M. CURR, *The Australian Race,* 4 vol. (1886–87); W.B. SPENCER and F.J. GILLEN, *The Native Tribes of Central Australia* (1899, reprinted 1938); A.W. HOWITT, *The Native Tribes of South-East Australia* (1904); W.B. SPENCER, *The Native Tribes of the Northern Territory of Australia* (1914); B. MALINOWSKI, *The Family Among the Australian Aborigines* (1913); and H. BASEDOW, *The Australian Aboriginal* (1925).

The basis of professional anthropological work was established by A.R. RADCLIFFE-BROWN—see his *Social Organization of Australian Tribes* (1931) and *Structure and Function in Primitive Society* (1952); and A.P. ELKIN, *Studies in Australian Totemism* (1933) and *Aboriginal Men of High Degree* (1946); in the contact sphere, A.P. ELKIN's article on "Reaction and Interaction: A Food Gathering People and European Settlement in Australia," *Am. Anthrop.,* 53:164–186 (1951), is significant. In regard to studies on Aboriginal women, see P.M. KABERRY, *Aboriginal Woman, Sacred and Profane* (1939); and C.H. BERNDT, *Women's Changing Ceremonies in Northern Australia* (1950); also F. GALE (ed.), *Women's Role in Aboriginal Society* (1971). On Aboriginal art, see F.D. MCCARTHY, *Australian Aboriginal Decorative Art,* 5th ed. (1958) and *Australian Aboriginal Rock Art* (1958); C.P. MOUNTFORD, *The Tiwi: Their Art, Myth, and Ceremony* (1958), and (ed.), *Art, Myth and Symbolism* (1956); and R.M. BERNDT (ed.), *Australian Aboriginal Art* (1964).

Controversies have stimulated discussion on certain aspects of traditional Aboriginal life. One arose from W.L. WARNER's and A.R. RADCLIFFE-BROWN's interpretation of the so-called "Murngin" kinship system. See also J.A. BARNES, *Inquest on the Murngin* (1967); and R.M. BERNDT in F.L.K. HSU (ed.), *Kinship and Culture* (1971). The concept of the local group vis-à-vis the horde, also arising from Radcliffe-Brown's earlier contentions, has been discussed by R.M. BERNDT in *Oceania,* 30:81–107 (1959–60); W.E.H. STANNER, *Oceania,* 37:1–26 (1965); and L.R. HIATT, *Oceania,* 32:267–286 (1962), and 37:81–92 (1966); as well as by J.B. BIRDSELL, with a commentary by students of Aboriginal affairs, in *Cur. Anthrop.,* 11:115–131, 138–142 (1970). The issues of law and order in the absence of clearly defined political authority, the relevance of kinship in this respect, and the place of tribal elders are the subject of conflicting views. R.L. SHARP in *Systems of Political Control and Bureaucracy in Human Societies,* ed. by V.F. RAY (1958); M.J. MEGGITT, *Indigenous Forms of Government Among the Australian Aborigines* (1962); and L. HIATT, "Social Control in Central Arnhem Land," *South Pacific,* 10:182–192 (1959), have taken one approach against A.P. ELKIN and R.M. BERNDT in *Aboriginal Man in Australia* (1965); and T.G.H. STREHLOW in *Australian Aboriginal Anthropology* (1970). However, the most long-standing controversy, deriving from Durkheim, has centred on totemism and has recently been revived by C. LEVI-STRAUSS in *Le Totémisme aujourd'hui* (1962; Eng. trans., 1963).

C.M.H. CLARK, *A Short History of Australia,* rev. ed. (1969); F.K. CROWLEY (ed.), *A New History of Australia* (1974, reissued 1981); *Australian Dictionary of Biography* (1966–); R. COVELL, *Australia's Music* (1967); H.M. GREEN, *A History of Australian Literature, Pure and Applied,* 2 vol. (1961); R. HUGHES, *The Art of Australia,* rev. ed. (1970).

C.M.H. CLARK, *A History of Australia,* 5 vol. (1962–78); J. GRIFFIN (ed.), *Essays in Economic History of Australia* (1969); T.B. MILLAR, *Australia in Peace and War: Exterior Relations, 1788–1977* (1978); A.G.L. SHAW and H.D. NICOLSON, *Growth and Development in Australia* (1964).

E.M. O'BRIEN, *The Foundation of Australia, 1786–1800,* 2nd ed. (1952); C.A. SHARP, *The Discovery of Australia* (1963); A.G.L. SHAW, *Convicts and the Colonies* (1966); P. BURROUGHS, *Britain and Australia, 1831–1855* (1967); A.C.V. MELBOURNE, *Early Constitutional Development in Australia,* 2nd ed. by R.B. JOYCE (1963); G.H. NADEL, *Australia's Colonial Culture* (1957); S.H. ROBERTS, *The Squatting Age in Australia, 1835–1847* (1964); A.G. SERLE, *The Golden Age: A History of the Colony of Victoria, 1851–1861* (1963), and *Rush to Be Rich: . . . Victoria, 1883–1889* (1971); N.G. BUTLIN, *Investment in Australian Economic Development, 1861–1900* (1964); G.L. BUXTON, *The Riverina, 1861–1891* (1967); R. ELY, *Unto God and Caesar: Religious Issues in the Emerging Commonwealth, 1891–1906* (1976); R.A. GOLLAN, *Radical and Working Class Politics . . . 1850–1910* (1960); J.A. LA NAUZE, *Alfred Deakin,* 2 vol. (1965); J.M. TREGENZA, *Professor of Democracy . . . C.H. Pearson, 1830–1894* (1968); F. ALEXANDER, *Australia Since Federation* (1967); C.H. FORSTER, *Industrial Development in Australia, 1920–1930* (1964); B.D. GRAHAM, *The Formation of the Australian Country Parties* (1966); N. MEANEY, *A History of Australian Defence and Foreign Policy, 1901–23* (1976–); G. SAWER, *Australian Federal Politics and Law,* 2 vol. (1956–63); I.A.H. TURNER, *Industrial Labour and Politics, 1900–1921* (1965); T.B. MILLAR, *Australia's Defence Policies, 1945–65* (1967); R. MURRAY, *The Split: Australian Labor in the Fifties* (1970); K. TENNANT, *Evatt: Politics and Justice* (1970); K. WEST, *Power in the Liberal Party* (1966).

New South Wales: The best source for up-to-date facts and statistics on New South Wales is the *Official Year Book,* issued annually by the government of New South Wales in conjunction with the Commonwealth Bureau of Census and Statistics. Historical, as well as other, information on New South Wales may be found in books on Australia, such as G. GREENWOOD (ed.), *Australia: A Social and Political History* (1955 and 1966); and A.G.L. SHAW, *The Story of Australia* (1955).

Northern Territory: Accounts of investigations of dry-land and irrigation farming, establishment of pastures, and husbandry are published in the COMMONWEALTH SCIENTIFIC AND INDUSTRIAL RESEARCH ORGANIZATION (CSIRO), *Division of Land Research and Regional Survey* (annual). Other references are CSIRO, *Katherine Research Station Report 1946–56* (1959); the DEPARTMENT OF TERRITORIES, *Prospects of Agriculture in the Northern Territory* (1960); B.R. DAVIDSON, *The Northern Myth: A Study of the Physical and Economic Limits to Agricultural and Pastoral Development in Tropical Australia* (1965); J.H. KELLY, *Struggle for the North* (1966); and the annual *Official Year Book of the Commonwealth of Australia and Northern Territory Statistical Summary.*

Queensland: GOVERNMENT STATISTICIAN, *Queensland Yearbook* (annual) and current bulletins; PREMIER'S DEPARTMENT, *The Queensland Scene* (1971), a pictorial account, with brief text; DEPARTMENT OF INDUSTRIAL DEVELOPMENT, *Investment Queensland, Australia* (1969); R.H. GREENWOOD, *Regions of Queensland* (1971); R.W. CILENTO (ed.), *Triumph in the Tropics* (1959), a historical study; STANLEY and KAY BREEDEN, *Tropical Queensland* (1970), on the flora, fauna, and landscape; GEORGE FARWELL, *The Sun Country* (1970).

South Australia: The best general sources of information are the annual *South Australian Year Book* and the *Official Year Book of the Commonwealth of Australia,* both published by the Commonwealth Bureau of Census and Statistics. More detailed statistical information is contained in the *Statistical Register of South Australia* (South Australian Parliamentary Paper, annual). For geographical aspects, see R.J. BEST (ed.), *Introducing South Australia,* pt. 2 (1958); and M. WILLIAMS (ed.), *South Australia from the Air* (1969). A popular survey of the economy is *Business Guide to South Australia,* published by the Premier's Department (1969).

For a comprehensive study of education see *Education in South Australia* (1971), the report of an official enquiry published by the Government of South Australia.

Tasmania: J.L. DAVIES (ed.), *Atlas of Tasmania* (1965), comprehensive topical and topographical map coverage with detailed commentary; HOBART, COMMONWEALTH BUREAU OF CENSUS AND STATISTICS, *Tasmanian Year Book* (published annually since 1967), the most comprehensive and important single reference on Tasmania; WINIFRED M. CURTIS, *The Endemic Flora of Tasmania,* 3 pt. (1967–71), a definitive study, beautifully illustrated; MICHAEL SHARLAND, *Tasmanian Birds,* 3rd ed.

rev. (1958), a field guide; E.R. GUILER, *Marsupials of Tasmania* (1960), an illustrated handbook of diagnostic features; HUNTER VALLEY RESEARCH FOUNDATION, *Tasmania in the Seventies* (1970), an analysis of economic growth and growth potential; P.G. PAK POY AND ASSOCIATES, *Study of the Transport of Goods for Tasmania,* 5 reports (1971), an economic evaluation of Tasmanian transportation commissioned by the state government, including (no. 4) an economic growth study; R.L. WETTENHALL, *A Guide to Tasmanian Government Administration* (1968), a descriptive classification.

Victoria: The MELBOURNE, COMMONWEALTH BUREAU OF CENSUS AND STATISTICS, *Victorian Year Book* (annual), is the outstanding reference on all aspects of Victoria, including articles on a wide range of topics, well illustrated by maps and photographs, and bibliography.

Western Australia: MICHAEL K. MORCOMBE and OSMAR WHITE, *Western Australia in Colour* (1971), has excellent illustrations. Other illustrated books with general texts are OSMAR WHITE, *Under the Iron Rainbow: Northwest Australia Today* (1969); T.A.G. HUNGERFORD and RICHARD WOLDENDORP, *A Million Square: Western Australia* (1969); and IVAN O'RILEY, *Giant in the Sun* (1968). JAMES S. BATTYE, *Western Australia* (1924), was the first substantial historical work, now largely supplanted by FRANK K. CROWLEY, *Australia's Western Third* (1960); and FRANK K. CROWLEY and BRIAN DE GARIS, *A Short History of Western Australia,* 2nd ed. (1969). See also FREDERICK ALEXANDER, FRANK K. CROWLEY, and J.D. LEGGE, *The Origins of the Eastern Goldfields Water Scheme in Western Australia* (1954); and ALEXANDER M. KERR, *Australia's Northwest* (1967) and *The South-west Region of Western Australia,* 2nd ed. (1965), for regional economic studies. CAROLINE GYE, *The Cockney and the Crocodile* (1962); and CYNTHIA NOLAN, *Outback* (1962), are personal accounts dealing with the Aborigines. *The Official Year Book of Western Australia* contains political, social, and economic explanatory and summary data in addition to statistical information.

Australian External Territories: AUSTRALIAN BUREAU OF STATISTICS, *Year Book Australia* (annual); AUSTRALIAN GOVERNMENT PUBLISHING SERVICE, *Commonwealth Government Directory* (annual); R. ROSE, *Australia's Island Territories* (1967); DEPARTMENT OF EXTERNAL AFFAIRS, COMMONWEALTH OF AUSTRALIA, *External Territories* (1970). See also the annual reports of the Australian Commonwealth Departments of Home Affairs, Administrative Services, Science and the Environment, and National Development. Current events are recorded in the Commonwealth Parliamentary Debates (*Hansard*) and the Australian national newspapers.

Norfolk Island: AUSTRALIAN GOVERNMENT PUBLISHING SERVICE, *Norfolk Island* (annual reports); J.J. SPRUSON, *Norfolk Island: Outline of Its History from 1788 to 1884* (1885); A.S.C. ROSS *et al., The Pitcairnese Language* (1964); J.S. TURNER, C.N. SMITHERS, and R.D. HOOGLAND, *The Conservation of Norfolk Island* (1968, for the Australian Conservation Foundation); AUSTRALIAN GOVERNMENT PUBLISHING SERVICE, *Report of the Royal Commission into Matters Relating to Norfolk Island* (1976); PHILIP GRUNDY and ROGER WETTENHALL, "Norfolk Island Versus The Nimmo Report," *Current Affairs Bulletin* (Oct. 1977); STUART INDER, "Islanders or Australians?", *Pacific Islands Monthly* (Feb. 1979); MERVAL HOARE, *Norfolk Island, An Outline of Its History, 1774–1977* (1979); *Norfolk Island, a History Through Illustration, 1774–1974* (1979). Hoare's volumes contain an extensive bibliography of primary and secondary sources.

Cocos (Keeling) Islands: AUSTRALIAN GOVERNMENT PUBLISHING SERVICE, *Territory of Cocos (Keeling) Islands* (annual reports); J.S. HUGHES, *Kings of the Cocos* (1950); CHARLES DARWIN, *The Structure and Distribution of Coral Reefs* (1842); HUGH CLIFFORD, *Heroes of Exile* (1906); N. TARLING, "The Annexation of the Cocos-Keeling Islands," *Historical Studies,* vol. 8 (May 1959); and T.E. SMITH, "The Cocos-Keeling Islands: A Demographic Laboratory," *Population Studies,* vol. 14 (Nov. 1960).

Christmas Island: AUSTRALIAN GOVERNMENT PUBLISHING SERVICE, *Territory of Christmas Island* (annual reports); C.W. ANDREWS, *A Monograph of Christmas Island* (1900); A.C. GIBSON-HILL, "The Early History of Christmas Island," *Australian Territories,* vol. 3 (May 1963); PETER HASTINGS, "A Tale of Two Islands: The Cocos/Christmas Story," *NG & Aust., Pacific & S.E. Asia,* vol. 9 (Sept.–Oct. 1974).

Coral Sea Islands: GOVERNMENT PRINTING OFFICE (Canberra), *External Territories of Australia* (March 1970); AUSTRALIAN BUREAU OF STATISTICS, *Year Book Australia* (annual).

Ashmore and Cartier Islands: AUSTRALIAN BUREAU OF STATISTICS, *Year Book Australia* (annual); ROBERT LANGDON, "Our Far Flung Empire," *NG & Aust., Pacific & S.E. Asia,* vol. 1 (June–July 1966).

Heard Island and the McDonald Islands: P.G. LAW and T. BURSTALL, *Heard Island* (1953); M.C. DOWNES *et al., The Birds of Heard Island* (1959); G.M. BUDD, "The Anare 1963 Expedition to Heard Island," *ANARE Reports,* series A, vol. 1 (1964); P. TEMPLE, *The Sea and the Snow: The South Indian Ocean Expedition to Heard Island* (1966).

Australian Antarctic Territory: D. MAWSON, *The Home of the Blizzard: Being the Story of the Australasian Antarctic Expedition, 1911–1914* (1915); P.G. LAW and J.M. BECHERVAISE, *Anare: Australia's Antarctic Outposts* (1957); R.A. SWAN, *Australia in the Antarctic* (1961); A.G. PRICE, *The Winning of Australian Antarctica* (1962); J.K. DAVIS, *High Latitude* (1962), and *With the "Aurora" in the Antarctic, 1911–1914* (1919); DEPARTMENT OF FOREIGN AFFAIRS, "Antarctica and the Question of the Exploitation of Its Resources," *Australian Foreign Affairs Record,* vol. 48, no 12 (Dec. 1977); DEPARTMENT OF SCIENCE AND THE ENVIRONMENT (Report by the Minister, Hon. J.J. WEBSTER), *Antarctica, an Information Paper* (1977); LENNARD BICKEL, *This Accursed Land* (1977). Current affairs in Antarctica are recorded in the quarterly journals *Aurora, ANARE Club Journal,* and *Antarctic* (New Zealand Antarctic Society). Other works of a more general nature published in Australia include: J.F. LOVERING and J.R.V. PRESCOTT, *Last of Lands, Antarctica* (1979); and J.M. BECHERVAISE, *Men on Ice in Antarctica* (1978), and *Antarctica, the Last Horizon* (1979).

Literatures of Australia and New Zealand

Although derived in part from the English literary heritage and influenced by its traditions, the literatures of Australia and New Zealand developed, from beginnings in the late 18th and early 19th centuries, their own characteristic themes, styles, and idioms. Moreover, they are quite distinct from one another, since they evolved at different rates and were affected by different historical and environmental circumstances; for example, the first settlers in Australia included many convicts transported from Britain, whereas in New Zealand they were largely traders and organized colonizers. Differences in geography, strength of indigenous opposition, pace of settlement, and degree of contact with Great Britain and other countries also contributed to divergences in literary expression. (Ed.)

This article is divided into the following sections:

AUSTRALIAN LITERATURE

Three main periods may be distinguished in the history of Australian literature. In the first (1788–1880) descriptive and documentary writings predominated. The second (1880–1940) was a time in Australia of expansion and consolidation and growing national feeling, and the diversification of Australian society was mirrored in a greater variety in its literature. In the third (from 1940) an increase in literary activity matched and reflected the country's increase in immigration, expansion of industrialization, and growth of large suburban areas. During this period the "cultural cringe," a feeling of many Australian writers that they worked under the shadow of English literature, was largely replaced by a growing confidence in Australia's nationhood and separate cultural identity.

The first hundred years. The life of the convicts, administrators, and soldiers who arrived in New South Wales in 1788 and of the free settlers who followed them was a constant struggle, allowing little time for the arts. Most of those with literary talent employed it in describing their new world. Memoirs and eyewitness accounts of convict life and, later, of conditions in the goldfields abounded and provided inspiration and material for later writers.

Novels of life in the colonies

The first Australian novel is generally agreed to be Henry Savery's *Quintus Servinton* (1830–31), the prototype of the convict novel through which Australians have continued to examine their relationship with Britain and the terms on which they came into existence as a nation. The most famous novel of convict life is Marcus Clarke's *For the Term of His Natural Life* (1874); a less documentary work than Savery's, it angrily indicted the transportation of prisoners and the brutal conditions of penal servitude in Tasmania. An outstanding early novelist was James Tucker, the probable author of *Ralph Rashleigh; or, The Life of an Exile* (written in 1844; published in 1952), who brought together the four cornerstones of Australian preoccupation: Britain, the convicts, the bushrangers, and the Aborigines. Rolf Boldrewood (pseudonym of Thomas Alexander Browne) enjoyed a huge popular following with his adventure novels, of which only *Robbery Under Arms* (1888) achieved classic status. The prolific English novelist Henry Kingsley (the brother of Charles Kingsley) is now chiefly remembered for his two Australian novels, *The Recollections of Geoffry Hamlyn* (1859) and *The Hillyars and the Burtons* (1865), both concerning the successful immigration of a British family; in the former novel the family returns to Britain, while in the latter the family achieves a social status in Australia that it could never have attained back home. Catherine Helen Spence's *Clara Morison* (1854) offered an ironic view of Adelaide in the period, and Mrs. Campbell Praed's *Policy and Passion* (1881) examined the personal life of an Australian politician.

Poetry in the first hundred years developed less steadily than did prose, with poets struggling to adapt the English Romantic tradition of landscape and reflective poetry to Australian subjects. Adam Lindsay Gordon, however, composed his *Bush Ballads and Galloping Rhymes* (1870) in a distinctly Australian idiom.

Nationalism and expansion. The last 20 years of the 19th century saw a growth of nationalism and movement toward federation of the separate states. National pride, the values of country life, and sympathy for the struggles of small landholders were common literary themes, particularly in stories by Henry Lawson and Steele Rudd (pseudonym of Arthur Hoey Davis). Another theme was that of early convict days; four volumes of tales by Price Warung (William Astley) exemplified this concern. Joseph Furphy's large novel *Such Is Life* (1903), describing the rural world of the 1880s, was full of details of station life, conversations of bullock drivers, nationalistic sentiments, and philosophizings about chance and determinism. A tradition of ballad verse that had long flourished was kept alive by Lawson and A.B. "Banjo" Paterson.

Federation (1901) was greeted with patriotic fervour in prose and in verse. In the early 1900s generally, however, nationalistic themes diminished in a growing awareness of the complexity of Australian life, and attention was drawn to the plight of ordinary people and the need for social reform. There was a new emphasis on realism in fiction. Its masterpiece was probably the trilogy of Henry Handel Richardson (Ethel Florence Lindesay Robertson), *The Fortunes of Richard Mahony* (1917–29), a study of fluctuating fortunes among the immigrants who were establishing the new urban Australia in the late 19th century. The last volume, *Ultima Thule,* graphically described conditions in the goldfields and brought its character studies of the temperamentally opposite Richard and Mary to a profoundly moving climax.

New realism

Richardson's skill in evoking the birth of Australia's urban society was matched by writers such as Louis Stone and Edward Dyson with their stories of inner-city life. Katherine Susannah Prichard was to take up their themes in a more radical vein in her novels, notably in *The Roaring Nineties* (1946); Kylie Tennant also made statements on behalf of the underdog in her fiction of the 1930s. Some of the most memorable works of the early part of the century emphasized the heroic nature of the pioneer instinct at the root of Australian society; one greatly popular example was Miles Franklin's fictitious autobiography *My Brilliant Career* (1901). Mrs. Aeneas Gunn's *We of the Never-Never* (1908) examined the relationship of the white inhabitants to the Aborigines, a theme taken up by Prichard in *Coonardoo* (1929) and by Xavier Herbert in *Capricornia* (1938). Vance Palmer, a prolific writer, made a particular contribution to the public discussion of an Australian cultural identity.

The success of poets was slighter but showed diversity of aims and interests. John Shaw Neilson wrote fine, delicate lyrics; Victor Daley produced romantic poems and some sharp satires; C.J. Dennis continued the ballad tradition with popular verse; Christopher Brennan, in his Symbolist poems, owed much to European tradition; and Kenneth Slessor created powerful, dramatic lyrics.

The modern period. During and after World War II, literary magazines proliferated, and the reading public grew. Though factual and descriptive writing remained prominent, from the 1950s onward Australian writers became increasingly speculative and searching. The most influential of mid-20th-century novelists was Patrick White, whose major novels were distinctively Australian but whose treatment had a largeness of vision surpassing nationalistic limitations. The Australia depicted in *The Tree of Man* (1955), *Voss* (1957), *Riders in the Chariot* (1961), *The Solid Mandala* (1966), and *A Fringe of Leaves* (1976), as well as in White's short stories, gave evidence of a critical, poetic imagination. White was awarded the Nobel Prize for Literature in 1973. He was drawn particularly to the underlying formative elements in Australian history.

The novels of Patrick White

The other dominant novelist of the mid-20th century was Christina Stead, with her interest in disrupted lives and financial pressures. Her most widely admired novels are *House of All Nations* (1938) and *The Man Who Loved Children* (1940); but only one, *Seven Poor Men of Sydney* (1934), was set entirely in Australia. Other novelists included Martin Boyd, Randolph Stow, and Thomas Keneally, who was drawn to a variety of historical subjects, including Joan of Arc, the Holocaust, the convict settlements, and World War I. Among novelists of importance in the 1970s and '80s were Thea Astley, Shirley Hazzard, Roger McDonald, and C.J. Koch, whose *Year of Living Dangerously* (1978) helped open up for Australian literature the subject of the Far East.

Aboriginal writing in English was uncommon, but Colin Johnson made a mark with *Wild Cat Falling* (1965). Literature on Australian themes began to be written in European languages other than English, particularly by members of Greek, Italian, and Maltese minorities during the 1960s. Also of interest were the autobiographical short stories of Hal Porter, which revealed the changes in Australian society from the 1930s to the 1960s.

Postwar poetry

A tradition of descriptive poetry remained, but mid-century poets ranged more widely than their predecessors. Robert D. FitzGerald was noted for strenuously argued poems; A.D. Hope for searching, allusive, witty verse; Douglas Stewart for forthright, vigorous works; Judith Wright for sensitive lyrics; and James McAuley for meditative lyrical poetry. Bruce Dawe evinced the Australian voice in his contemporary, journalistic poetry. Among the leading poets of the 1980s were Les Murray, allusive and humane, and Chris Wallace-Crabbe, with his precisely depicted states of mental and physical being. Kath Walker was the first Aboriginal poet who wrote in English.

Australian literature was long deficient in drama; not until the late 1950s and '60s did any plays achieve real success. Best known were Douglas Stewart's *Ned Kelly* (published 1943) and *Fire on the Snow* (performed 1941); Ray Lawler's *Summer of the Seventeenth Doll* (1955); Alan Seymour's *One Day of the Year* (1961); and Patrick White's *Four Plays* (published 1965). Several playwrights came into prominence during the next two decades, among them David Williamson, Jack Hibberd, Alexander Buzo, Jim McNeil, and Peter Kenna. (L.J.K./A.N.R.N.)

NEW ZEALAND LITERATURE

In organized settlement New Zealand was 50 years behind Australia, and for most of the century from 1840 to 1940 a similar gap in national consciousness seemed to prevail. Yet from about 1930 there was a flowering of literature.

Early writings. Up to the 1920s most writers aimed at British readers and often exploited and distorted exotic and indigenous elements of their surroundings. There did emerge some worthwhile chronicles and autobiographical narratives, such as *A First Year in Canterbury Settlement* (1863) by Samuel Butler (author of *Erewhon*), Frederick Maning's *Old New Zealand* (1863), and Lady (Mary Anne) Barker's *Station Life in New Zealand* (1870). Among colonial versifiers the most notable were Alfred Domett and the Scots dialect poet John Barr.

Katherine Mansfield

Katherine Mansfield was one of the great names in short-story writing. She did not expect to find a readership in her own country and emigrated to England to establish her career. But her debt to New Zealand pioneer society was clearly evident in her most admired stories, such as "At the Bay," "The Voyage," and "The Stranger" (in *The Garden Party, and Other Stories,* 1922). Two other writers achieved works of distinction during this period: William Satchell, in *The Land of the Lost* (1902); and Jane Mander, in *The Story of a New Zealand River* (1920).

Outstanding in the 1920s and '30s were the short-story writer Frank Sargeson (pseudonym of Norris Frank Davey) and the poets A.R.D. Fairburn, R.A.K. Mason, Allen Curnow, and Denis Glover; what distinguished their work was that they assumed a readership of fellow countrymen. The yarn spinner had always been characteristic of New World writing, and Sargeson used this to great effect in tales of the Depression years. Two novelists to emerge during this time were John Mulgan, whose *Man Alone* (1939) attained the status of a New Zealand classic, and John A. Lee, with his harrowing *Children of the Poor* (1934). Robin Hyde (pseudonym of Iris Wilkinson) wrote in *The Godwits Fly* (1938) a perceptive study of a Wellington family in the early part of the century.

World War II and after. In the years from 1940 most literary genres flourished, especially the short story, whose best writers included Roderick Finlayson, Maurice Duggan, O.E. Middleton, and Phillip Wilson. The poet James K. Baxter was perhaps the most considerable figure of the postwar period. Among the best known prose writers were Sylvia Ashton-Warner (*Spinster* [1958], *Greenstone* [1966]); Janet Frame (*Faces in the Water* [1961], *A State of Siege* [1966], *The Rainbirds* [1968]); Ian Cross (*The God Boy* [1957], *After Anzac Day* [1961]); Maurice Shadbolt (*Among the Cinders* [1965], *Strangers and Journeys* [1972]); Maurice Gee (*A Special Flower* [1965], *In My Father's Den* [1972]); Dan Davin (*No Remittance* [1959]); and David Ballantyne (*The Cunninghams* [1949]).

Modern prose

The growth of New Zealand writing and a rapid expansion of theme reflected an accelerating national consciousness. The international recognition of Keri Hulme's partly Maori-based novel *The Bone People* (1984) finally disposed of the image of New Zealand as exclusively genteel and concerned only with farming and a sentimental view of Britain. Janet Frame's three-volume autobiography also helped to show a more socially troubled and psychologically stressful side of New Zealand. New poets with a distinctive national voice emerged, among them Fleur Adcock, Lauris Edmond, Michael Jackson, C.K. Stead, and Ian Wedde. Maori writers included Hone Tuwhare, Witi Ihimaera, and Patricia Grace. (M.F.R.S./A.N.R.N.)

BIBLIOGRAPHY

Australian literature: E. MORRIS MILLER, *Australian Literature: A Bibliography to 1938, Extended to 1950,* rev. ed. edited by FREDERICK T. MACARTNEY (1956); H.M. GREEN, *A History of Australian Literature,* rev. by DOROTHY GREEN, 2 vol. (1984); ALEC D. HOPE, *Australian Literature, 1950–1962* (1963); GEOFFREY DUTTON (ed.), *The Literature of Australia,* rev. ed. (1976); BARRY ARGYLE, *An Introduction to the Australian Novel, 1830–1930* (1972); HARRY HESELTINE (ed.), *The Penguin Book of Australian Verse* (1972), and *The Penguin Book of Modern Australian Verse* (1981); PETER FITZPATRICK, *After "The Doll": Australian Drama Since 1955* (1979); CHRIS WALLACE-CRABBE, *The Golden Apples of the Sun: Twentieth Century Australian Poetry* (1980); LEONIE KRAMER (ed.), *The Oxford History of Australian Literature* (1981); LEONIE KRAMER and ADRIAN MITCHELL (eds.), *The Oxford Anthology of Australian Literature* (1985); and WILLIAM H. WILDE, JOY HOOTON, and BARRY ANDREWS, *The Oxford Companion to Australian Literature* (1985).

New Zealand literature: D.M. DAVIN (ed.), *New Zealand Short Series* (1953, reprinted 1976 as *New Zealand Short Stories, First Series,* 1976); C.K. STEAD (ed.), *New Zealand Short Stories, Second Series* (1966); VINCENT O'SULLIVAN (ed.), *New Zealand Short Stories, Third Series* (1975); LYDIA WEVERS (ed.), *New Zealand Short Stories, Fourth Series* (1984); CHARLES BRASCH (ed.), *Landfall Country: Work from Landfall, 1947–61* (1962); J.C. REID and P. CAPE (eds.), *A Book of New Zealand,* rev. ed. (1979); ERIC H. MCCORMICK, *New Zealand Literature: A Survey* (1959); JOAN STEVENS, *The New Zealand Novel, 1860–1965,* 2nd rev. ed. (1966); FLEUR ADCOCK (ed.), *The Oxford Book of Contemporary New Zealand Poetry* (1982); and IAN WEDDE and HARVEY MCQUEEN (eds.), *The Penguin Book of New Zealand Verse* (1985). See also *Landfall* (quarterly), a literary journal.

(L.J.K./M.F.R.S./A.N.R.N.)

Austria

Located in the centre of Europe, Austria is a federal republic with a scenic and largely mountainous terrain of 32,375 square miles (83,850 square kilometres). In the decades following the collapse, in 1918, of the multinational Austro-Hungarian Empire of which it had been the heart, this small landlocked country experienced more than a quarter-century of social and economic turbulence and a Nazi dictatorship. Yet the establishment of permanent neutrality in 1955, associated with the withdrawal of the four-power troops that had occupied the country for a decade, enabled Austria to develop into a stable and socially progressive nation, with a flourishing cultural life that was reminiscent of its earlier days of international musical glory. Its social and economic institutions, too, have been characterized by new forms and a spirit of cooperation, and, although political and social problems remain, they have not erupted with the intensity evidenced in other countries of the Continent.

A great part of Austria's status in the late 20th century can be attributed to its geographical position: it is at the centre of European traffic between east and west along the great Danubian trade route and between north and south through the magnificent Alpine passes. Austria is bordered on the west by Switzerland, and together these countries form what has been characterized as the neutral core of Europe. The tiny principality of Liechtenstein also forms a small enclave on the west. Hungary on the east, Germany and Czechoslovakia on the north, and Italy and Yugoslavia on the south illustrate the variety of political and economic systems within which Austria is embedded (see EUROPEAN HISTORY AND CULTURE and VIENNA).

This article is divided into the following sections:

Physical and human geography

THE LAND

Relief. Mountains and forests give the Austrian landscape its character, although in the northeastern part of the country the Danube winds between the eastern edge of the Alps and the hills of Bohemia and Moravia, in its journey toward the Hungarian Plain. Vienna, the capital of Austria and long one of the great cities of Europe, lies in the area where the Danube emerges from between the mountains into the drier plains.

The landscape of the eastern Alps offers a complex geological and topographical pattern, with the highest elevation—the Grossglockner (12,457 feet [3,797 metres])—rising toward the west. The Austrian Alps may be subdivided into a northern and a southern limestone range, each of which is composed of rugged mountains. These two ranges are separated by a central range that is softer in form and outline and composed of crystalline rocks. North of this massive Alpine spur, which forms the physical backbone of the country, lies a hilly subalpine region, stretching between the northern Alps and the Danube, while to the north of that river lies a richly wooded foothill area. The lowland area east of Vienna may be regarded as a western extension of the great Hungarian Plain.

Drainage. Austria is a land of lakes, many of them a legacy of Ice Age erosion, which scooped out mountain lakes in the central Alpine district, notably around Salzkammergut. The largest lakes—lying partly in the territory of neighbouring countries—are the Bodensee (Lake

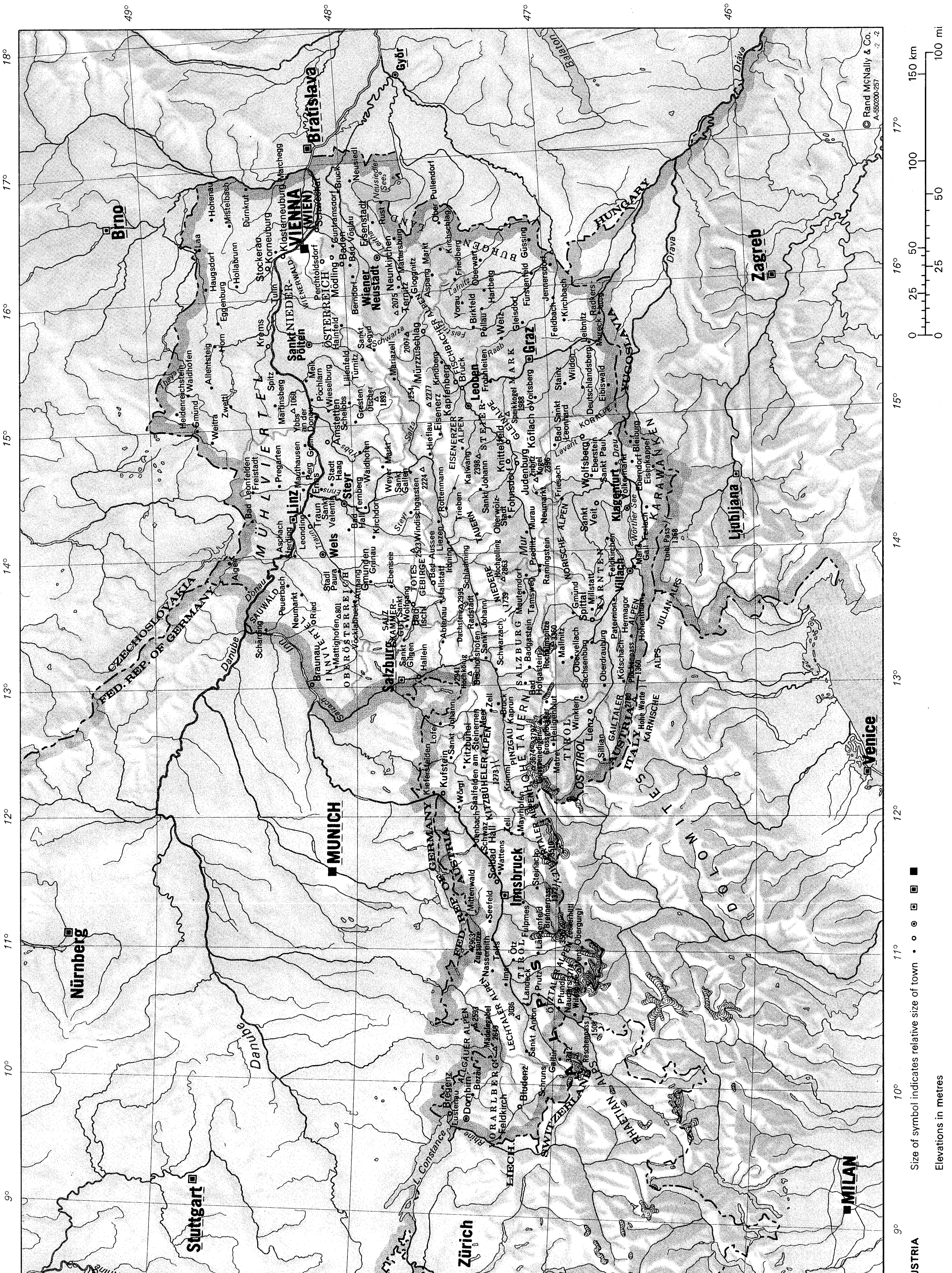
MAP INDEX

Political subdivisions

The name of a political subdivision if not shown on the map is the same as that of its capital city.

Cities and towns

Physical features and points of interest



AUSTRIA

Size of symbol indicates relative size of town

Elevations in metres

of Constance) in the west and the marshy Neusiedlersee (Neusiedler Lake) to the east.

Ninety-six percent of Austrian territory drains to the Danube River system. The main watershed between the Black Sea and the North Sea runs across northern Austria, in some places lying only 22 miles (35 kilometres) from the Danube, while to the west the watershed between the Danube and the river systems emptying into the Atlantic and the Mediterranean coincides with the western political boundary of Austria. In the south the Julian and Karnische Alps, and, farther to the west, the main Alpine range, mark the watershed of the region draining into the Po River of northern Italy.

Climate. The wooded slopes of the Alps and the small portion of the plains of southeastern Europe are characterized by differing climatic zones: the wetter western regions of Austria have an Atlantic climate with a yearly rainfall of about 40 inches (1,000 millimetres), whereas the eastern regions, in particular those under the influence of the drier, more continental type of climate, have less precipitation.

In the lowlands and the hilly eastern regions, the median temperature ranges from 30.4° F (−0.9° C) in January to 68.6° F (20.3° C) in July. In those regions above 10,000 feet, by contrast, the temperature range is between 11.8° F (−11.3° C), with a snow cover of about 10 feet in January, and 35.8° F (2.1° C) in July, with about five feet of snow cover.

The prevailing wind is from the west, and the humidity, therefore, is highest in the west, diminishing toward the east.

Plant and animal life. Two-thirds of the total area of Austria is covered by woods and meadows, and the country is the most densely forested nation in central Europe. Spruce dominates the forests, with larch, beech, and oak also making a significant contribution. In the Alpine and foothill regions coniferous trees predominate, while leaf-bearing deciduous trees are more frequent in the warmer zones.

Wild animals, now protected by conservation laws, include the brown bear, the eagle, and all buzzard species, as well as falcons, owls, cranes, swans, and storks. Game hunting is restricted to certain periods of the year, with deer and rabbits the most frequent quarry. Austrian rivers nurture river and rainbow trout, grayling, hake, pike, perch, and carp.

Traditional regions. Western Austria, comprising the *Bundesländer* (states) of Vorarlberg, Tirol, and Salzburg, is characterized by high Alpine regions with majestic mountains and magnificent scenery. This high Alpine character also extends to the western part of Kärnten (Carinthia), the Salzkammergut of central Austria, to the Alpine blocks of Steiermark (Styria), and to the eastern rim of the Alps.

Alpine scenery

The outer fringes of the Alps dominate the northern portion of Oberösterreich (Upper Austria) with their subalpine characteristics, while the richly wooded Bohemian massif extends across the Czechoslovak border. This part of Austria is furrowed by many valleys that, for centuries, served as passageways leading to the east and southeast of Europe and even, in the case of medieval pilgrims and crusaders, to the Holy Land.

A hilly vineyard district extends into the *Bundesland* of Burgenland. This region, together with the eastern part of the state of Steiermark, lies at the gateway to the Hungarian Plain. Burgenland, Steiermark, Kärnten, and Niederösterreich (Lower Austria) form a region, throughout which agriculture abounds. The city of Vienna and its peripheral suburbs and industrial settlements border both Alpine and lowland regions.

Settlement patterns. The pattern of rural settlement in Austria was shaped centuries ago by the exigencies of the Alpine environment, and new rural building is still influenced by these ancient traditions, especially in the west and in the centre of the country. By contrast, rural housing in the eastern parts of the nation, especially in the lowlands, is dominated more by agricultural needs than by harsh weather conditions.

Austria is not only a mountainous but also a highly urbanized country. In the late 20th century, roughly half of the Austrian population lived in cities and towns with more than 10,000 residents, and about one-fifth of the total Austrian population lived in Vienna. The state capitals have, as a consequence of national economic and social development, grown somewhat in population since the end of World War II. Graz, Austria's second largest city, is the gateway to the Balkans; Innsbruck is the rail and road centre through which the north–south traffic of western Austria passes; Salzburg is one of the best known European centres of musical culture and Baroque architecture; Linz is an important industrial centre; and Klagenfurt lies astride routes that provide access to both Italy and Yugoslavia.

Major Austrian cities

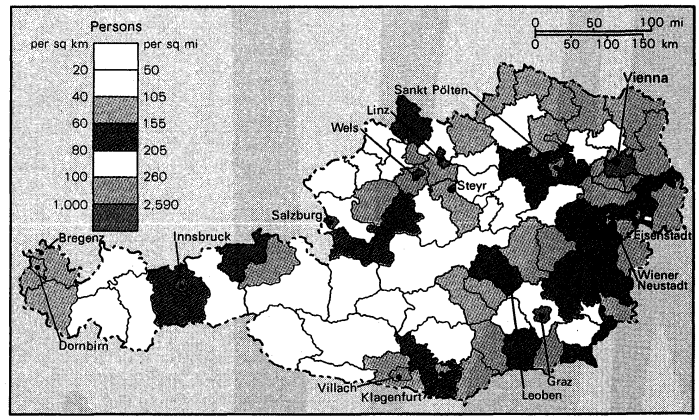
THE PEOPLE

Ethnic and linguistic heritage. Virtually all Austrians are German-speaking. The tiny minority belonging to other nationalities consists of Croats and Magyars living in Burgenland; Slovenes living in the southern part of Kärnten; and Czechs living in Vienna. The population is predominantly Roman Catholic, with Protestant churches, the so-called Old Catholic church, and Judaism constituting the largest remaining affiliations.

The westernmost states, and Vorarlberg in particular, are inhabited by Alemannic groups whose dialect is similar to the Swiss-German dialect, while the dialect of the people in Salzburg and Oberösterreich and the eastern states resembles that which is spoken in Bavaria. People in Steiermark and Kärnten speak a dialect that is clearly distinguishable from that of the people in the west. In general, the German spoken in Austria has a softer, more drawling, and melodious sound than that which is spoken in Germany.

Demography. Throughout the second half of the 20th century, the birth rate and life expectancy in Austria rose steadily; concurrently, the expanding population became increasingly urban as, particularly in the western states, agriculture gave way to industry. Birth rates varied from state to state, being generally highest in the strongly Roman Catholic western states and rural areas and lowest in the urban centres. The almost complete destruction of Vienna's Jewish population (about 200,000 in 1938) during the Nazi era was only partly compensated by an influx from rural areas and by refugees after World War II.

Population shifts



Population density of Austria.

Emigration resulted in some losses in national population during the disastrous post-World War II years, although a stream of refugees from eastern European countries helped to make up some of the deficit. Most of these refugees were soon integrated into Austrian society. As a result of the turmoil in Hungary in 1956, some 180,000 Hungarians crossed the frontiers into Austria, creating a new refugee problem. By the end of the following decade, however, most had migrated to other countries.

THE ECONOMY

The Austrian economy has been shaped by two factors: first, the nation occupies a mountainous area, only half of which is even potentially usable for the production of food; second, and as a result, the success of industrial

production, exports, and trade is basic to the economy. The natural resources available within the country for industrial exploitations are, therefore, of considerable significance.

Resources. Austria is the world's leading producer of natural magnesite, a magnesium carbonate that is used extensively in the chemical and other industries, with Kärnten being the main centre of magnesite production. In the late 20th century, new methods of smelting and processing magnesite fostered a significant increase in production.

Iron-ore deposits are found in the Erzberg in Steiermark—which accounts for the major portion of Austria's total iron production—and in the Hüttenberg region in Kärnten. The large oil refinery at Schwechat provides roughly three-fourths of the nation's consumption of oil and oil products, the balance in crude and refined oils being met by imports. Reserves of oil and natural gas are a steadily diminishing asset. Austria has made several agreements with the Soviet Union to ensure the delivery of increasing quantities of natural gas. Deposits have been explored in the Vienna and Graz basins and in Oberösterreich near Linz and Wels. Coal, which is found only in small quantities and is mostly of the soft variety, is mined chiefly in Upper Austria and in Styria.

Developing oil production

Since World War II, Austria has become one of Europe's foremost producers of hydroelectric power. The issue of nuclear power met with public controversy in the late 20th century.

Agriculture and forestry. Since barely half of Austria's territory can be used for food production, an intense program of agricultural modernization was introduced after World War II. The number of tractors, for example, rose from 1,800 in 1939 to 300,000 40 years later. The rural economy is dominated by small holdings, although some large forest estates still remain. As a result of the modernization program, the yield of the most important agricultural products has risen considerably.

In the lowland regions of the country, the abundance of pasture land and the production of animal feed has promoted cattle breeding. Dairy products are plentiful, resulting in a surplus. With the exception of horses, Austria's livestock has grown in numbers during the modernization period.

Timber from Austria's vast national forests is only partly processed in the country, and untreated lumber makes up a large percentage of the nation's exports. More than half of the forests are part of small holdings; the rest are on large, old estates or are government-owned.

Industry. Industry and trade form the most important sources of national income. Iron and steel production, in particular, increased greatly after World War II: both pig iron and raw steel production, for example, rose more than threefold during the first three decades after mid-century. Despite an abundance of domestic iron ore, scrap iron and iron ore must be imported, due to increased steel production. An oxygen blast furnace, based on the so-called L.D.-process (named after the cities of Linz and Donawitz [now Leoben], where it was developed), reduced the need for imported scrap iron and is now used in many countries under license agreements.

The iron and steel industry

Aluminum became one of the more important Austrian manufacturing industries during the 1960s and '70s. Other manufacturing includes paper production, based on the nation's extensive forest reserves, and the chemicals and plastics industries.

Trade. Austria's main trading partners are its neighbours, Germany and Switzerland, who together account for almost half of Austria's imports and more than one-third of its exports. Trade with the European Economic Community (EEC, or Common Market) as a whole is also important. Significantly lesser proportions of Austria's trade are conducted with countries of the European Free Trade Association (EFTA) and with eastern European countries. Although manufactured goods accounted for about two-thirds of Austria's exports in the late 20th century, timber, iron, and steel continued to be very important export items.

The EEC's important role in Austria's foreign trade has prompted the government to establish a free-trade agreement with the organization (full membership would not be compatible with Austria's policy of neutrality). Since July 1977 trade in industrial goods, with the exclusion of certain products, such as paper and steel, has been duty free.

The tourist trade is Austria's outstanding invisible export. The number of nights spent by foreign tourists in Austria rose almost 20-fold in the first three decades after mid-century. Expenditures by tourists at the end of that period largely met Austria's balance-of-payments deficits.

Finance. Monetary policy is jointly determined by the Ministry of Finance and the Austrian National Bank. The bank operates under a law passed in 1955 and amended in 1969 that states that in its policy decisions the bank must pay "due regard to the economic policy of the government." Half of the shares of the bank are held by the government, and the other half are held by various economic institutions. A majority of the members of the board of governors, including the president and two vice presidents, are appointed by the government. The bank is also charged with the administration of foreign-exchange controls. There is virtually complete freedom for current account transactions and a very liberal regime governing capital account transactions with foreign countries.

Financial services are handled by a wide range of other institutions, including joint-stock commercial banks, private commercial banks, savings banks, provincial mortgage organizations, agricultural credit cooperatives, industrial credit cooperatives, building societies, and specialized banks. Vienna's geographic location has made it an important financial centre for East–West trade.

Administration of the economy. The private sector, comprising roughly three-quarters of Austria's mixed economy, is concentrated in agriculture and food processing, forestry and timber, paper mills, textile and clothing, food and beverages, and retail and wholesale trade. Private enterprise has a relatively low degree of monopolistic concentration, although ties among the guildlike cooperatives (Berufsgenossenschaften und Innungen) are not without influence.

Nationalization of industry

In 1946 the Austrian parliament nationalized a large segment of Austrian industry, which, at that time, was being held under Soviet control as alleged former German property. After the State Treaty of 1955 established Austrian neutrality and brought the end of Soviet authority, these enterprises passed into exclusive Austrian control. The 1946 nationalization covered the three biggest banks and some 70 larger industrial enterprises, chiefly in the fields of iron and steel, aluminum, and machinery. Subsequent reorganization reduced the number of nationalized enterprises to 19. By the fourth quarter of the century, the net value of the output of nationalized concerns amounted to about one-fifth of the total industrial production.

Most of the nationalized industries are organized as joint-stock corporations, and their central direction by the government has undergone several changes, for both political and managerial reasons. Until 1966, for example, a Cabinet member supervised the nationalized industries, but after that date they were handed over to the government-owned Austrian Industrial Administration (Österreichische Industrie–Aktien–Gesellschaft [OIAG]). The Administration's board of directors, although representing the government, is independent as far as managerial decisions are concerned.

The most important source of federal tax revenue is the income tax. In 1973 the sales tax was replaced by a value-added tax on all sales and transactions, with necessities and luxury items assigned special lower and higher rates, respectively.

Transportation. Austria's location in the centre of Europe determines its role in European road, rail, air, and river traffic, whether the flow is from north to south or from east to west. Austria consequently is not only a tourist country but in many aspects is also a freight clearinghouse, servicing the great trade routes running across the Alps and along the Danube River.

Roads. The road system, which sustained heavy damage during World War II, was subsequently adapted to vastly increased traffic requirements. Much of the Austrian road network leads over spectacular Alpine passes.

A federally built east–west highway was in operation between Vienna and Salzburg by the early 1960s. It also connects Salzburg with Munich, Germany. The Trans-European North–South Motorway project will eventually connect Austria with nine other countries. The total number of motor vehicles rose more than 10-fold during the first three decades after mid-century.

Railroads. Forty-one percent of the Austrian railroad network and 381 bridges were destroyed during World War II and had to be repaired as well as modernized in subsequent years. More than half of the federal rail network, carrying two-thirds of the traffic, has been converted to electric traction. Passenger and freight traffic have shown a proportionate increase. The Austrian railroads are state-owned; by a law of 1969, however, the railroad administration—formerly part of the Transportation Ministry—became an independent commercial enterprise.

Importance of the Danube

Water transport. The Danube is the most important river connection between Germany and the Black Sea, and the federally owned Danube Steamship Company plays an important role in both freight and passenger traffic along this waterway. Although Austria is landlocked, its federally owned shipyards build vessels not only for Austria but also for the Soviet Union and for other countries.

Air transport. Austrian Airlines, which began operations in March 1958, has since established service throughout Europe and to several countries abroad. Austria's main airport is at Wien-Schwechat, southeast of Vienna.

ADMINISTRATIVE AND SOCIAL CONDITIONS

Government. *Constitutional framework.* Under the constitution of 1920—with minor changes made in 1929—Austria is a "democratic republic: its power derives from the people." A federal state, Austria consists of nine self-governing states: Burgenland, Kärnten, Niederösterreich, Oberösterreich, Salzburg, Steiermark, Tirol, Vorarlberg, and Wien (Vienna).

In 1934 the Austrian constitution was replaced by an authoritarian regime under chancellors Engelbert Dollfuss and Kurt von Schuschnigg. This, in turn, was eliminated by Hitler after Nazi Germany annexed Austria in 1938. With the liberation of Austria in 1945, the constitution of 1929 was revived and subsequently became the foundation stone of constitutional and political life in the "Second Republic."

The federal president and the Cabinet share the executive authority. The president is elected by popular vote for a term of six years. He acts as head of state, appoints the Cabinet, and calls Parliament into session. He can dissolve Parliament during the four-year legislative period, unless it dissolves itself by law, and can order new elections. He also acts as commander-in-chief of the armed forces.

The president appoints the federal chancellor and, at his suggestion, the other Cabinet members. The Cabinet cannot remain in office if it and its members do not enjoy the confidence of the majority of the National Council.

The Austrian parliament

The Parliament consists of two houses: the National Council (Nationalrat), wielding the primary legislative power; and the Federal Council (Bundesrat), representing the states. The National Council, which had had 165 members since 1920, was expanded to 183 members under a law passed in 1970 that took effect at a national election in 1971. The National Council is elected by all citizens over the age of 19, and every citizen over the age of 26 is eligible to run for office. The distribution of parliamentary seats is based on a system of proportional representation.

The members of the Federal Council represent the states, and the state assemblies, or diets, elect the members by a proportional system based on the population of the state.

The legislative process originates in the National Council. Each bill—except for the budget, which is the sole prerogative of the National Council—must be approved by the Federal Council. The National Council, however, can override a Federal Council veto by a simple majority vote.

Each of the nine states is administered by a government headed by a governor; the governor is elected by the legislative diet, which in turn is elected by general and equal ballot.

The local municipalities each elect a mayor and a city council. Vienna is a unique case: as both municipality and state, its mayor also functions as the governor.

The political process. The first popular election of a president, although provided for by the 1929 amendment to the constitution, did not take place until after the death of the first post-World War II president, Karl Renner (1870–1950), who had been unanimously elected by the National Assembly after the liberation of 1945 and the ensuing unsettled political circumstances. A succession of presidents, belonging to, or nominated by, the Socialist Party have occupied the post since then.

The major parties

In the eight elections held from the immediate post-World War II period to the early 1970s, the stability of the two main Austrian political parties—the People's Party and the Socialist Party—is shown by their respective shares of the vote, which has been in each case, approximately 45 percent of the total votes cast. The Liberal Party has polled an average of only 5 percent of the votes, and the Communist Party has polled an even smaller share. From 1945 to 1966 the two principal parties maintained coalition governments with the chancellorships held by the People's Party. In 1966 the People's Party, with a slim majority of parliamentary seats, formed the government alone, but after the 1970 elections, in which it lost its majority, the Socialists, as the strongest party, formed a minority government. Its first budget was passed with the help of the votes of the Freedom Party. In the parliamentary election that was held in October 1971 the Socialist Party won more than 50 percent of the popular vote, the first party to do so in Austrian history, and it retained this position until the election of 1983, when it won a plurality of the votes and formed a coalition government with the Freedom Party.

The People's Party is the successor of the Christian Social Party, founded in the 1890s. It represents a combination of conservative forces and various social and economic groups forming semi-independent federations within the overall party. The divergencies of economic and social interests of these groups—farmers, businessmen, and blue- and white-collar workers—necessitate policy compromises, which are not always easy to implement.

The Socialist Party, which followed a democratic Marxist program during the 1918–34 period, has adopted a more pragmatic approach since 1945. The party program, in its 1958 and 1978 versions, stressed these new tendencies. Emphasis is placed on social problems, on government influence on an expanding and socially oriented economy, on full employment, and on a rising standard of living, with a more equal distribution of wealth. The Austrian Socialists are no longer an exclusive party of the working class: by the 1970s they had a much wider appeal.

The Liberal Party has undergone several changes in emphasis since it was admitted into politics by the occupying powers in 1949, at which point it was a third party with nationalistic tendencies. The initially prominent influence of former Nazis has diminished.

The Communist Party, always negligible in Austrian politics, reached its high point during the Soviet occupation of the eastern zone of the country, when it attained 5 percent of the popular vote. By 1959 it had lost representation in Parliament, and by 1970 in all of the provincial diets and in most municipal councils.

The participation of the citizens. The Austrian constitution provides for two kinds of popular initiatives: one is the so-called popular demand (*Volksbegehren*), by which 400,000 vote-eligible citizens can petition Parliament for approval of any bill; it can also be initiated, in the case of any bill or proposal, by a majority of the National Council. A total revision of the constitution must be approved by plebiscite.

Justice. The administration of justice is independent of Austrian legislative and administrative authorities. Judges are not subject to any government influence: they are appointed by the Cabinet upon nomination by judicial panels and can neither be dismissed nor transferred without their agreement.

Austria has three high courts: one sitting as the highest body of appeal in civil and criminal matters; a top ad-

ministrative court to which citizens aggrieved by administrative decisions can appeal; and a constitutional court that decides on all constitutional matters, civil rights, and election disputes.

Armed forces. Under the State Treaty of 1955 Austria is permitted armed forces of 50,000 men for the defense of its territory. The armaments permitted Austria specifically exclude nuclear or other major offensive weapons. Only the four signatory powers—the United States, Great Britain, the Soviet Union, and France—can permit such armaments, and they must do so by a unanimous decision.

All male citizens between the ages of 18 and 50 are liable to military service, and conscripts are called up for six months, followed by 60 days of reserve training, with an additional 30 to 90 days of training for reservists in special functions. Austria's armed forces total 38,000; in addition there are some 70,000 reservists called up for short training periods each year. The total mobilizable strength is 155,000.

Education. Austria's educational system is based on obligatory school attendance between the ages of six and 15. A major reform of the school administrative structure, providing for a unitary school system with access to higher education and experimentation with various types of schools, has been in progress since 1970.

Austrian universities

Intermediate schools include those preparing students for university and other higher studies, for teachers' and commercial colleges, and for other specialized institutions.

Austrian universities are among the oldest German-speaking universities of Europe: the University of Vienna, for example, was founded in 1365. Austria has more than 15 universities, including two technical universities, a university specializing in economics and another in agriculture, as well as academies for fine arts, music, drama, and design. Austria also has an extensive adult-education system.

Health and welfare. *Public health.* Public health in Austria is the responsibility of the Ministry of Health and Environment. It supervises 14 subsidiary institutes responsible for the prevention of infectious diseases and inspection of drugs and food. The provincial governments also have public-health centres, and each municipality and rural district must employ a public-health physician.

National health insurance covers expenses of medical and hospital treatment: blue- and white-collar workers and salaried employees are protected in cases of sickness, disability, unemployment, and maternity and in their old age. There are also survivors' pensions. Pension systems for self-employed persons and farmers have also been established.

Cooperative wage–price policy

Wages and cost of living. Government, management, and labour follow a wage–price policy that attempts to avoid social cleavages and strikes through cooperative participation in a joint wage–price commission. The representatives of the various segments of the economy, together with the chambers of commerce, agriculture, and labour and the federation of trade unions, have tried, with the active cooperation of the government, to coordinate wage and price movements. It is within this framework that collective bargaining takes place, and agricultural prices are also negotiated by the wage–price commission without infringement of the market economy.

Social and economic divisions. The three important economic groups in Austria—labour, management, and farmers—have similar structures: each has its own independent organization: the trade unions, the management association, and the farmers' federation. At the same time, laws provide for semi-official "chambers" for each group. This type of guild organization promotes cooperation in the governmental wage–price commission. Despite the divergent interests of the various groups, their cooperation has resulted in relative economic stability, and labour–management relations have remained unmarked by major crises.

CULTURAL LIFE

The cultural milieu. The cultural milieu of Austria has a rich heritage: in architecture and poetry it dates to the Middle Ages, and in medicine and science it can be traced to the 18th and 19th centuries. Similarly, Vienna's art galleries are among the most highly esteemed in Europe because of their wealth of old Dutch masters. Austria's most highly recognized cultural contribution has been in the field of music, and this tradition still dominates its present cultural achievements. Austrian cultural life has exerted attraction beyond its borders. Such great musicians as Beethoven (a native of Bonn), Brahms (from Hamburg), Mahler (from Bohemia) and, to a certain extent, Richard Strauss (from Munich) are considered as much a part of Austria's musical life as the celebrated native Austrians Haydn, Mozart, and Schubert. The Vienna waltzes emanating from the Johann Strauss dynasty, as well as the Vienna operetta, whose presiding genius was Franz Lêhar, round out the rich traditions of Viennese musical life. Austria has also been the birthplace of modern, especially 12-tone, music, centred in the work of Arnold Schoenberg and known as the Second Viennese School.

Musical tradition

Although in literature Austria has often been considered a segment of German culture, writers such as Franz Grillparzer, Johann Nestroy, and Ferdinand Raimund in the postclassic era, and Hugo von Hofmannsthal and Arthur Schnitzler in the 20th century, all developed special Austrian traits. Austrian culture was also strongly influenced by the various cultural elements of the multinational Austro-Hungarian Empire, with its Slavic and Magyar traces. They still exist, making Austria a bridge between various parts of Europe and attracting performing artists from East and West.

Austrian cultural and scientific life had to recover from the severe wounds inflicted by Nazi domination and the loss of intellectuals and artists through emigration and persecution after 1938. A 10-year occupation by four world powers promoted recovery of an independent cultural life only to a small degree. Austria has, nevertheless, managed—particularly since the late 1950s—to regain something of its past cultural posture.

The state of the various arts. Austrian expressionism in literature has gained world renown through the visionary novelist Franz Kafka, a native of Prague who died near Vienna in 1924. As part of a general Kafka revival, one of his haunting novels, *Der Prozess* (*The Trial*), was made into an opera by Austria's best known contemporary composer, Gottfried von Einem.

Georg Trakl, a promising Expressionist who was killed in World War I; Franz Werfel, an Expressionist novelist and playwright; Karl Kraus, a critical lyricist and essayist; and Robert Musil, who wrote along Expressionist lines, have all played a considerable role in contemporary Austria. Hermann Broch, a novelist with symbolic tendencies, is recognized beyond Austria. Heimito von Doderer is considered the most remarkable post-World War II novelist. Modern playwrights who are well known include Alexander Lernet-Holenia and Fritz Hochwälder, who has created historical dramas that have been produced throughout the world.

Austrian theatre and opera

The Austrian theatre and opera recovered quickly after the war. The Vienna Burgtheater, considered to be the best German-speaking theatre in the 19th century, may again be counted among the foremost German stages. It presents the classics, as well as modern English, American, and occasionally Russian and Czech plays. A new trend in the Vienna theatre has developed in so-called cellar-theatres ("off-Broadway" stages), which favour the theatre of the grotesque.

The Vienna Staatsoper (State Opera), which was completely rebuilt after World War II, ranks today with La Scala in Milan and the Hamburg and Munich operas, while the Vienna Philharmonic Orchestra has played in almost all the musical capitals of the world. Austrian theatre life is intense, with large and enthusiastic audiences and a bewildering succession of shows. The theatre has spread to small stages in the Vienna suburbs, while the various state capitals have also become theatre and light-opera centres, developing their own identities. Special performances for students aid an early participation in cultural life.

Modern music was born in Vienna, and Arnold Schoenberg, Alban Berg, and Anton von Webern are considered among its major creative founders.

Clemens Holzmeister, the best known modern Austrian architect, was responsible for the two festival theatres in Salzburg, which combine classical and modern theatre styles. Artur Perotto, of Linz, designed several striking skyscrapers in western Austria, while in Vienna public apartment building has had a strong overall influence on architecture.

In sculpture Fritz Wotruba became internationally recognized. His style has often been compared with that of the modern English sculptor Henry Moore.

Austria's most famous modern painter is Oskar Kokoschka, truly a world figure: he and Alfred Kubin are regarded as two of the foremost creators of modern painting in Austria. In addition, a young group of Surrealists has shown its works in international exhibitions.

Folk art and folk traditions, supported by provincial governments, have survived in western Austria, especially in Tirol.

The oldest Austrian academic research institution is the Academy of Science, whose traditions date to the end of the Middle Ages. More modern scientific foundations, notably the Körner Foundation and the Renner Foundation, support scientific research and other cultural endeavours: their main support comes from government sources. A federal Ministry of Science and Research was established in 1970; it is responsible for university institutions and for the advancement of scientific activities. For statistical data, see the "Britannica World Data" section in the current *Britannica Book of the Year.*

(O.L./K.R.St./Ed.)

History

PREHISTORY AND ROMAN TIMES

In the territories of the Austrian Republic the first traces of human settlement date back to the Early Paleolithic Period. The archaeological material becomes richer and more varied for subsequent periods, giving evidence of several distinct cultures succeeding one another or coexisting. The Austrian site of Hallstatt gave its name to the principal culture of the Early Iron Age (*c.* 800–450 BC). Celtic tribes invaded the eastern Alps around 400 BC and eventually founded the kingdom of Noricum, the first "state" on Austrian territory known by name. In the west, however, the ancient race of Raetians was able to maintain its seat. Then, attracted by the rich iron resources and the strategic importance of the region, the Romans began to assert themselves. After an initially peaceful penetration during the last two centuries BC, Roman troops finally occupied the country *c.* 15 BC, and the lands as far as the Danube became part of the Roman Empire, being allotted to the Roman provinces of Raetia, Noricum, and Pannonia.

Occupation by Roman forces

The Romans opened up the country by an extensive system of roads. Among the Roman towns along the Danube, Carnuntum (near Hainburg) took precedence over Vindobona (Vienna), while Lauriacum (Lorch, near the confluence of the Enns and the Danube) belonged to a later period. Roman municipalities (*municipia*) also grew up at Brigantium (Bregenz), Juvavum (Salzburg), Ovilava (Wels), Virunum (near Klagenfurt), Teurnia (near Spittal an der Drau), and Flavia Solva (near Leibnitz). North of the Danube the Germanic tribes of the Naristi, Marcomanni, and Quadi settled. Their invasions in AD 166–180 arrested the peaceful development of the provinces, and even after their repulse by the emperor Marcus Aurelius the country could not regain its former prosperity. In the 3rd century the Roman frontier defenses began to be hard pressed by invasions from the Alemanni. Finally, in the 5th century, heavy attacks by the Huns and eastern Germans put an end to the Roman provincial defense system on the Danube.

There is archaeological evidence of a Christian cult in this area from the 4th century, and the biography of St. Severinus by Eugippius constitutes a unique literary source for the dramatic events of the second half of the 5th century. At that time several Germanic tribes (Rugii, Goths, Heruli, and later Langobardi) settled on Austrian territory. In 488 part of the harassed Norican population withdrew to Italy.

EARLY MIDDLE AGES

Germanic and Slavic settlement. Following the departure of the Langobardi to Italy (568), further development was determined by the Bavarians in a struggle with the Slavs, who were invading from the east, and by the Alemanni, who settled in what is now Vorarlberg. The Bavarians were under the political influence of the Franks, whereas the Slavs had Avar rulers. At the time of their greatest expansion the Slavs had penetrated as far as Styria, Carinthia, and eastern Tirol. After 624 the western Slavs rose against the Avars under the leadership of the Frankish merchant Samo, whose short-lived rule may also have extended over the territories of the eastern Alps. Around 700 the Bavarian lands again bordered on Avar territory, with the lower course of the Enns forming the approximate frontier. On the death of the Frankish king Dagobert I (639) the Bavarian dukes from the house of Agilolfing became virtually independent.

Christianity had survived only here and there among the remnants of the Roman population, when around 600 and then again around 700 Christian missionaries from the west became active, with the support of the Bavarian dukes. At the end of the 7th century St. Rupert, who came from the Rhine, founded the church of Salzburg. When they were threatened once more by the Avars, the Alpine Slavs (Karantaner) placed themselves (before 750) under the protection of the Bavarians, whose mission was extended to them. At the same time, Bavarian settlers penetrated into the valleys of Carinthia and Styria. Charlemagne, emperor of the neighbouring Franks, however, defeated the Bavarian duke Tassilo III, wiping out the Bavarian dukedom for a century. During the following years (791–796) Charlemagne also led a number of attacks against the Avars and destroyed their dominion. Surviving Avars were made to settle in the eastern part of Lower Austria between the rivers of Fischa and Leitha, where they soon disappeared from history, most probably mixing with the native population.

Frankish domination in the 8th and 9th centuries

As was the usual Frankish practice, border provinces (Marken, or marches) were instituted in the newly won southeastern territories. The Avar March on the Danube and Lower and Upper Pannonia and Karantania were to form a border fortification; but this arrangement soon became less effective because of frequent disagreements among the nobility. To that unrest was added a threat from the Bulgarians and from the rulers of "Great Moravia." Nevertheless, the process of Germanization and Christianization continued, during the course of which the churches of Salzburg and Passau came into conflict with the eastern mission which was led by the Slav apostles Cyril and Methodius. The Frankish kingdom richly endowed the church and nobility with new lands, which came to be settled by Bavarian and Frankish farmers.

In 881 the beginning of incursions by the Magyars led to a first clash near Vienna. By 906, they had destroyed greater Moravia, and, in 907, near Pressburg (Bratislava), the Magyars defeated a large Bavarian army that had tried to win back lost territory. Liutpold of Bavaria and Theotmar, the archbishop of Salzburg, were killed in battle. The Lower Austrian territories as far as the Enns River and Styria as far as the Koralpe fell under Magyar domination. Nevertheless, a certain continuity of German-Slav settlement was maintained, so that, after the victory of the German king Otto I (955) and the further repulse of the Magyars in the 960s, a fresh start could be made.

The Babenberg period. The first mention of a ruler in the regained territories east of the Enns is of Burchard, who probably was count (burgrave) of Regensburg. It appears that he lost his office as a result of his championship of Henry II the Quarrelsome, duke of Bavaria. In 976 his successor, Leopold I of the House of Babenberg, was installed in office. Under Leopold's rule the eastern frontier was extended to the Vienna Woods after a war with the Magyars. Under his successor, Henry I, the country around Vienna itself must have come into German hands. New marches were also created in what was later known as Carniola and Styria. Wars against Hungarians and Moravians took up the reign (1018–55) of Margrave (a count who ruled over a march) Adalbert. Parts of Lower

Austria on both sides of the Danube were lost temporarily; after they were retaken, they became the so-called Neumark (New March), which for some time enjoyed independence—as did the Bohemian march to the north of the Babenberg territories. The position of the Babenbergs was at that time still a modest one; their territorial rights were no greater than those of other leading noble families. Their power within their own official sphere was further diminished by ecclesiastical immunities (Passau in particular, but also Salzburg, Regensburg, and Freising), with numerous monasteries owning large territories as well.

Involvement in the papal-investiture controversy

Austria was repeatedly drawn into the disputes of the investiture controversy in which the Pope and the Holy Roman Emperor fought for control of the church in Germany. In 1075 Margrave Ernest, who had regained the Neumark and the Bohemian March for his family, was killed in the Battle of the Unstrut, fighting on the side of the king (later emperor) Henry IV against the rebellious Saxons. Altmann, bishop of Passau, a leader of church reform and a champion of Pope Gregory VII, influenced the next Babenberg margrave, Leopold II, to abandon the cause of Henry IV. As a result, Henry roused the Bohemian duke Vratislav II against him, and in 1082 Leopold II was defeated near Mailberg and his territories north of the Danube devastated. The Babenbergs, however, managed to survive these setbacks. Meanwhile, the cause of church reform gained ground, with its centres in the newly founded monasteries of Göttweig, Lambach, and, in Styria, Admont.

Under Leopold III (1095–1136) the history of the Babenbergs reached its first culmination point. In the struggle between emperor and pope, Leopold avoided taking sides until a consensus had built up among the German princes that it was Emperor Henry IV who stood in the way of a final settlement. Then Leopold did not hesitate to side with Henry's rebellious son, Henry V (1106). For this he was rewarded with the hand of Henry V's sister Agnes, who had formerly been married to the Hohenstaufen Frederick I of Swabia. The intermarriage with the reigning dynasty not only increased Leopold's reputation but no doubt also brought him additional power. Leopold was even proposed as a candidate to the royal throne, but he declined. It was apparently his intention to concentrate on consolidating his position in Austria. He was the first Austrian margrave to describe himself as the holder of territorial principality (*principatus terrae*), and during his time Austrian common law is mentioned for the first time, another proof of the developing national consciousness.

Leopold's reputation with the clergy was high, and he was eventually canonized (1485). He gave generous endowments to religious communities, establishing the Cistercians at Heiligenkreuz, and he founded, or at least restored, the monastery of Klosterneuburg, which he gave to Augustinian canons. It was in Klosterneuburg as well that he built a residence in which he stayed even after he had acquired Vienna.

Conflict of the Hohenstaufen, Welfs, and Babenbergs

On the death of Leopold III, the Babenbergs were drawn into a conflict between the two leading dynasties of Germany, the Hohenstaufen and the Welfs—on the side of the Hohenstaufen because of their family ties. In 1139 the German king Conrad III bestowed Bavaria, which he had wrested from the Welfs, on his half-brother, Leopold IV. After the latter's untimely death, Henry II Jasomirgott succeeded to the rule of Austria and Bavaria.

Emperor Frederick I Barbarossa tried to put an end to the quarrel between the Welfs and the Hohenstaufen, and in the autumn of 1156 at Regensburg he arranged a compromise. Bavaria was restored to the Welf, Henry the Lion, duke of Saxony, while the Babenbergs were confirmed in their rule of Austria, which was made a duchy, and were given the "three countships," the actual location of which is disputed. Also, the obligations of the dukes of Austria toward the empire were reduced. Their attendence at royal court days was only called for when court was held in Bavaria, and they were compelled to participate only in campaigns of the empire that were directed against Austria's neighbour; that is, Hungary. Henry II Jasomirgott and his wife, Theodora, a Byzantine princess, were granted succession through the female line and the right, in the event of the premature death of their children, to appoint a candidate for the succession. The Babenbergs also were given the right of approving the exercise of jurisdiction by other powers within the new duchy, permitting Henry to exert pressure against such rival internal powers, secular as well as ecclesiastical. The rights of the duke were laid down by imperial charter (*Privilegium Minus*). For centuries, however, Austria continued to contain territorial dominions not subject to the duke. Henry moved his residence to Vienna, where he also founded the monastery of the "Scottish" (actually Irish) monks.

Acquisition of Styria by the Babenbergs

In 1192 the Babenbergs' territory was greatly extended when they won the duchy of Styria. In Styria the margraves of the family of the Otakars of Steyr had gradually asserted themselves—under conditions similar to those of the Babenbergs—over their rivals, the noble families of the Eppensteiner, Formbacher, and Aribonen. The most successful among the Styrian margraves had been Otakar III (reigned 1130–63). Then, in 1180, Emperor Frederick I, in the course of a renewed anti-Welf policy, raised Styria to the status of a duchy and granted it complete independence from Bavaria. A few years later, a treaty of inheritance (Georgenberg; 1186) was concluded between the dukes Leopold V of Austria (reigned 1177–94), a son of Henry Jasomirgott, and Otakar IV of Styria, the ailing last Otakar ruler. When Otakar died in 1192, Leopold succeeded him, and thus the Babenbergs came into the inheritance.

With the exception of a short intermission (1194–98), the reigning Babenberg henceforth ruled both duchies, Austria and Styria. Styria then included parts of the Traungau, which eventually was to become part of Upper Austria, and the province of Pitten, north of the Semmering, afterward assigned to Lower Austria. In logical continuation of the Babenberg policy, Leopold VI the Glorious and his successor, Frederick II the Warlike, the last representative of the dynasty, extended their domains farther south, gaining fiefs in Carniola.

Before he had inherited the duchy of Styria, Leopold V had taken part in the Third Crusade, during which, on the ramparts of Acre, he had become involved in a quarrel with the English king, Richard I the Lion-Heart. Later, on his return journey to England, Richard tried to make his way through Austria in disguise but was recognized near Vienna, taken prisoner, and later handed over to Emperor Henry VI. England had to pay a heavy ransom, a share of which Leopold obtained and invested in the foundation, extension, and fortification of towns as well as in the stamping of a new coin, the so-called Wiener Pfennig. The road connecting Vienna and Styria was improved, and the new town of Wiener Neustadt was established on its course to protect the newly opened route across the Semmering Pass.

On Leopold V's death the Babenberg domains were divided between his sons for four years, until the death of one of them, Frederick I, in 1198. His brother Leopold VI, the most outstanding member of the family, then took over as sole ruler (1198–1230). This was a time of great prosperity for the Babenberg countries. In imperial politics Leopold VI again took sides with the Hohenstaufen, backing Philip of Swabia. In church matters the Duke was a great supporter of the monasteries, founding a Cistercian monastery at Lilienfeld (*c.* 1206). He tried to concentrate patronage rights over ecclesiastical property in his own hands and took rigorous action against the heretics (Cathari and Waldenses). He participated in several crusades in Palestine, Egypt, southern France (against the Albigenses), and Spain (against the Saracens). Leopold VI's efforts to emancipate Austria ecclesiastically by creating a separate Austrian bishopric in Vienna came to naught because of the opposition of the church in Passau and especially Salzburg; nor did his son Frederick II succeed in the same matter. Leopold VI played some role in imperial politics, bringing about the peace Treaty of San Germano between Emperor Frederick II and Pope Gregory IX (1230). He met his death in San Germano, and his body was transported to Lilienfeld to be buried there.

The rule of Leopold VI

A change came about under the last representative of the dynasty, Frederick II the Warlike, Leopold's son. His harsh

internal policy and military excursions against neighbouring lands, together with his opposition to the emperor Frederick II, led in 1237 to the temporary loss of both Austria and Styria. The crisis, however, was overcome, and fresh opportunities were about to open for the Duke when, on June 15, 1246, he was killed in battle against the Hungarians on the Leitha River. With him the male line of the family came to an end.

The political history of Austria from the end of the 10th to the middle of the 13th century is marked by the establishment and consolidation of territories. This process was most advanced in the Babenberg domains but was not confined to them. Dukes Herman (1144–61) and Bernhard (1202–56) of Carinthia achieved a comparable status, and Count Albert of Tirol (died 1253) moved in the same direction. The archbishops of Salzburg strove to eliminate all secular powers and patrons of their see, but in the other territories, secular princes strengthened their rule.

Another milestone of this period was the completion of the colonization of the Austrian territories. New settlements were now established by clearing the woods and advancing to more remote mountain areas. Several old and new settlements grew into market centres and towns and were eventually granted charters. The colonization movement also affected the ratio of German to non-German population. Except for some places in the Alpine regions, the Slavs were gradually assimilated, and the same holds true of the remnants of the Roman population in Salzburg and northern Tirol.

The intellectual life of the period deserves mention, too. The Babenberg court was famous enough to attract some of the leading German poets. At the beginning of the 13th century the Nibelung saga was written down by an unknown Austrian. Historical writing flourished in the monasteries. The era also produced first-rate Romanesque and early Gothic architecture.

LATE MIDDLE AGES

The contest for the Babenberg heritage. Upon the death of Frederick II the Warlike, the Babenberg domains became the political objects of aspiring neighbours. The Emperor and the Pope also tried to intervene. Two female descendants of the Babenbergs, Frederick's niece Gertrude and his sister Margaret, were considered to embody the claims to the heritage. Gertrude married first the Bohemian prince Vladislav and afterward the Margrave Hermann of Baden, who died in 1250. After Hermann's death, Otakar II (Přemysl Otakar II), prince of Bohemia (from 1253 king), married the widowed Margaret. Thereupon Hungarian forces intervened. Under the Treaty of Ofen (1254) Otakar was to rule Austria, while King Béla IV of Hungary received Styria. Troubles in Salzburg, stemming from a conflict between Bohemia and Hungary, inspired a rising among the Styrian nobles. Otakar intervened and in the Treaty of Vienna (1260) took over Styria as well. The state of anarchy that prevailed in Germany during this period proved advantageous to Otakar, who was granted both Austria and Styria in fief from Richard, earl of Cornwall, the titular German king. The grant, however, was only by writ and was invalid according to German law. During the following years Otakar's energetic rule met with growing opposition among the Austrian nobility. He introduced foreigners into important official positions, broke fortresses that had been erected without his consent, and dissolved his childless marriage with Margaret. Otakar had two of the opposition leaders, Otto of Meissau and Seifried of Mahrenberg, executed. The inhabitants of the cities, on the other hand, and the gentry, as well, generally favoured Otakar, who supported the churches and the monasteries. To complete his success, Ulrich of Spanheim, duke of Carinthia, willed Carinthia and Carniola to Otakar in 1269.

Rise of Přemysl Otakar of Bohemia

Reverses came only when Count Rudolf IV of Habsburg was elected German king as Rudolf I on September 29, 1273. Cautiously but nevertheless energetically, Rudolf set about to undermine the powerful position Otakar had created for himself. He challenged the legitimacy of Otakar's acquisitions and finally placed the Bohemian king under the ban of the empire. In 1276 Rudolf and his allies invaded Austria, forcing Otakar to do homage and to renounce his claims to Austria. Two years later, while trying to recover what he had lost, Otakar was defeated by the united forces of Rudolf and the Hungarians and was killed on the battlefield near Dürnkrut (August 26, 1278).

The accession of the Habsburgs. As the German princes had not cared to give Rudolf adequate support against Otakar, he did not feel bound to them and set about to acquire the former Babenberg lands for his own house. In 1281 he made his eldest son, Albert (later Albert I, king of Germany), governor of Austria and Styria; on Christmas, 1282, he invested his two sons, Albert and Rudolf II, with Austria, Styria, and Carniola, which they were to rule jointly and undivided. As the Austrians were not used to being governed by two sovereigns at the same time, the Treaty of Rheinfelden (June 1, 1283) provided that Duke Albert should be the sole ruler. In 1282 Carniola had already been pawned to Meinhard II of Tirol (of the counts of Gorizia), one of the most reliable allies of Rudolf who, in 1286, was also invested with Carinthia.

At first the Habsburg rulers were far from popular in Austria. Albert's energetic and relentless rule roused bad feeling, and the Swabian entourage that had arrived with the new dynasty to occupy key positions was despised by native nobles. There were conflicts with Bavaria, Salzburg, and Hungarian nobles who violated the Austrian frontier. After the death of King Rudolf (1291), all the neighbours and rivals of the Habsburgs and the counts of Gorizia united. Albert, however, succeeded in negotiating a peace with his most dangerous foes, the Hungarians and the Bohemians, and he broke the fortresses of the rebel nobility. Meanwhile, Meinhard II had stifled the uprising in Carinthia.

Rule of Albert of Habsburg

In 1292 Albert had been passed over in the German election, and Adolf of Nassau was called to the throne. When Adolf fell out with the electoral princes, however, they went over to Albert, who had just subdued another rebellion in Austria. After Adolf was defeated and killed near Göllheim (1298), Albert had himself elected a second time. In his Austrian lands Albert's main concern was to provide for an effective administration, in which he was assisted by his privy councillors, most of whom were foreign. Records were set up to codify the prerogatives and returns of the ducal property. Eventually Albert did not spare the church, either. When the Přemysl family died out in 1306, Albert aspired to the Bohemian throne. He had his eldest son, Rudolf III, elected Bohemian king, but Rudolf died the following year. Albert was preparing for a new campaign when he was murdered by his nephew, John, and some accomplices (1308).

On Albert's death the anti-Habsburg movement flared up again in Austria, but his sons, Frederick I the Fair and Leopold I, managed to maintain control. Frederick stood for election as German king (as Frederick III), and for the next years the Habsburg countries had to support the cost of the war with his rival, Louis IV of Bavaria, until 1322, when Frederick was defeated near Mühldorf. Earlier, another more decisive battle had been lost by the Habsburgs to the Swiss at Morgarten in 1315. From that time on, the Habsburg domains in the territory south of the Rhine and the Bodensee (Lake Constance) began to crumble away. Frederick the Fair spent his last years in Austria and was buried in the Carthusian monastery of Mauerbach (1330). He seems to have been the first of the Habsburgs for whom Austria meant home. From his time on, Habsburg rule and Habsburg territories were known as the Austrian domains (*dominium Austriae*), a term that was replaced, in the course of the 14th and 15th centuries, by the new concept of the House of Austria.

After Frederick's death the Habsburgs were for some time ruled out as possible candidates for the German throne; but, under the brothers Albert II and Otto, Habsburg Austria received its first important accession of territory. In 1335 Carinthia and Carniola were acquired after the death of Henry of Gorizia; while, with the help of Luxembourg troops, Henry's daughter Margaret (surnamed Maultasch) managed to retain the Tirol. Albert and his brother Otto had not gotten on too well, but when Albert came to rule on his own, he proved to be of sound judgment and

Acquisition of Carinthia, Carniola, and Tirol

keen on preserving the peace. It was a time of severe catastrophes: bad harvests, floods, and earthquakes, and in 1348–49 the plague, which brought a persecution of the Jews that was suppressed, however, by the Duke. Albert arranged several tours around his domains to establish contacts with the populace and improve jurisdiction. Two campaigns against the Swiss failed to yield any spectacular results, but they helped once more to consolidate the weakened Habsburg position. At his death in 1358, Albert left four sons. Though in 1355 a family ordinance had decreed that all the male members of the family were to rule jointly over the undivided domains, only the eldest among them, Rudolf, was then fit to rule. Throughout his short reign (1358–65), Rudolf IV showed himself extremely energetic and ambitious. He started to rebuild St. Stephen's Cathedral in the Gothic style, and he founded the University of Vienna (1365). With these two projects he imitated and rivalled his father-in-law, the emperor Charles IV, at Prague.

In 1359 Rudolf's forged charter, the Privilegium Majus, by which he claimed immense privileges for Austria and its dynasty, as well as the title of archduke, caused a breach between him and the emperor Charles IV. Charles was not prepared to accept the Privilegium Majus to its full extent (although it later was sanctioned by the Habsburg emperor Frederick III in 1442 and again in 1453). Upon news of the death of Duke Meinhard in 1363, Rudolf prevailed upon the Duke's mother, Margaret, to make over the Tirol to him. On this occasion the Emperor backed the Habsburgs against the Wittelbachs, and the Tirol thus passed to the House of Austria.

The division of the Habsburg lands. Rudolf was succeeded in 1365 by his two brothers, Albert III and Leopold III. After some years of joint rule, however, they quarrelled and in 1379, by the Treaty of Neuberg, partitioned the family lands. Albert, as the elder brother, received the more prosperous countries on the Danube (Upper and Lower Austria). The rest of the widespread domains fell to Leopold (including Styria, Carinthia, Tirol, the old Habsburg countries in the west, and central Istria). The treaty also contained several points on mutual wardship, preemption rights, and common titles, by which some connection between the two lines was to be preserved.

Treaty of Neuberg of 1379

In 1381 the resourceful Duke Leopold took advantage of the weak position of Venice in its war with Genoa and seized Trieste, which had broken away from Venice. His efforts to expand his rule in the west, however, were less successful, though he seemed lucky enough at first. Envisaging a connection between the original Habsburg territories in the west and the new domains in the Tirol, the Habsburgs looked for a foothold in the region west of the Arlberg (modern Vorarlberg). Neuberg on the Rhine was won in 1363 and Feldkirch in 1375. Another important acquisition was the city of Freiburg in the Breisgau. But then Leopold came into conflict with the Swiss, which led to defeat and his death at Sempach in 1386. An army of his brother, Albert III, was likewise defeated near Näfels in 1388, and the Habsburgs suffered heavy territorial losses. Leopold's sons recognized the wardship of Albert, who acquired Bludenz and the Mantafon Valley west of the Arlberg in 1394. In his own domains Albert was forced to check the dynasty of the Schaunbergs (in Upper Austria), who tried to create an independent domain around Wilhering and Eferding. Albert III especially favoured the city of Vienna as his capital, and it was because of his reorganization that the university Rudolf IV had founded there was able to survive.

After Albert's death in 1395, new Habsburg family troubles arose, differences that the treaties of Hollenburg (1395) and Vienna (1396) tried to settle. Under the Vienna treaty, the line of Leopold III split into Styrian and Tirolian branches, resulting in three complexes of Austrian territories—a state of affairs that was to reappear in the 16th century. The individual parts came to be known by the names of Niederösterreich (comprising modern Lower and Upper Austria), Innerösterreich (comprising Styria, Carinthia, Carniola, and the Adriatic possessions), and Oberösterreich (comprising the Tirol and the western domains, known as the Vorlande, or Vorderösterreich).

In 1396 the Austrian estates, or diets, were first assembled to consider the Turkish threat and henceforth were to play an important political role in Austria. In them the nobility usually took the lead, but they also included representatives of the monasteries, the towns, and the marketplaces. In the Tirol, in Vorarlberg, and, at times, in Salzburg, the peasants also sent their representatives to attend the diets. Because of the Habsburg partitions and frequent regencies, the estates were able to gain in importance. They did not obtain the right to pass laws, but they obstinately insisted on the privilege to grant taxes and duties.

The early diets of the nobility, church, and towns

After the short rule of Albert IV (1395–1404) and a troublesome tutelary regime (1404–11), Albert V came into his own, and with him the Danube countries again enjoyed a strong and energetic rule (1411–39). Albert, however, had married the daughter of the emperor Sigismund and was thus drawn into the Hussite religious wars, in the course of which the Austrian lands north of the Danube were ravaged. In the Austrian west, Duke Frederick IV of the Tirolian branch lost the Aargau to the Swiss but was able to assert himself in Tirol against a rebellion of his nobles.

When Sigismund died, Albert inherited his positions. In 1438 he was first elected Hungarian king, with the German (as Albert II) and the Bohemian crowns to follow later. Albert no doubt had many of the qualities of a born ruler, but he died prematurely in 1439 on an unsuccessful campaign against the Turks. Soon thereafter, his widow gave birth to a son and heir, Ladislas Posthumus, to whom Frederick V of Styria, as the senior member of the house, became guardian. Frederick also had Sigismund, the son of Frederick IV of Tirol, under his tutelage.

Thus began the long reign of Frederick V (as Roman emperor he was to become Frederick III). His reign was marked by almost ceaseless strife with the estates, with his neighbours, and with his jealous family. When he tried unsuccessfully to take advantage of a conflict among the Swiss Confederates, the Tirolians made Frederick release Duke Sigismund from tutelage (1446). A few years later, on his return from Rome, where he had been crowned emperor, his enemies at home and abroad in 1452 forced him also to give up Ladislas, who was then the recognized king of Hungary and Bohemia. The boy king's policies were made by Count Ulrich of Cilli. Ulrich was murdered at Belgrade in 1456, however, and a year later King Ladislas died. In Bohemia and Hungary national kings came to power. Frederick now won himself a foothold in the Austrian domains on the Danube and succeeded in acquiring the rich estates and fiefs of Ulrich.

The Burgundian and Spanish marriages. Maximilian I, the son of the emperor Frederick III, was married to the Burgundian heiress, Mary, at Ghent (1477). By that tie to Burgundy the Habsburgs became involved in long struggles with France. After Mary's death (1482), Maximilian, moreover, met with increasing difficulties in the Burgundian countries themselves. In the meantime, another crisis had arisen in the eastern Habsburg domains. Disagreement about the Bohemian succession and a political error of Frederick III, who tried to install the former archbishop of Gran (Esztergom) at Salzburg, led King Matthias I Corvinus of Hungary to march against Austria. Vienna was besieged and finally taken by the Hungarians (1485), as was Wiener Neustadt (1487). The harried Maximilian came into even greater distress in the Low Countries, where the rebellious citizens of Bruges put him under arrest (1488). Sigismund, the Habsburg ruler of the Tirol, who was heavily encumbered by debts, planned to sell his country to the Bavarians. A complete breakdown of the House of Habsburg threatened, but Maximilian was ultimately released. He prevailed upon Sigismund to abdicate in his favour. In 1490 the Habsburgs were able to take over Lower Austria. Maximilian even attacked Hungary but in the Treaty of Pressburg (1491) renounced claims to Hungary, though reserving the succession rights of his family.

Near collapse of House of Habsburg

After the death of his father, Emperor Frederick III, Maximilian came into a heritage that surpassed the endowments of all his predecessors. Furthermore, his son, Philip I the Handsome, who governed the Low Countries, was betrothed to the Spanish infanta, Juana (later Joan

the Mad), and through the unexpected death of male members of the Spanish dynasty this marriage was to raise the Habsburgs to the throne of Spain. In the German Empire as well as in Austria, Maximilian introduced sweeping administrative reforms that were the first steps toward a centralized administration. In 1508 Maximilian assumed the title of elected emperor as he was unable to pass through hostile Venetian territory to go to Rome for his coronation, and henceforth Rome and the pope had no more say in the creation of new emperors.

During Maximilian's last years, Eastern politics again came to the fore. The great crusade he planned against the Turks, however, never materialized. In 1515 Maximilian arranged a double marriage between his family and the Jagiellon line that ruled Bohemia and Hungary, thus reviving earlier Habsburg claims to these countries. Maximilian's energetic reign added greatly to the prestige of the Habsburgs. Thus, his grandson Charles (V) was able to prevail against French intrigue to inherit the imperial crown. Charles's younger brother, Ferdinand I, took over the rule of the Austrian countries but encountered the opposition of the estates, which he cruelly suppressed. In the agreements of Worms (1521) and Brussels (1522) Charles V formally handed over the Austrian lands to his brother. The subsequent years of Ferdinand's reign were troubled by peasant risings in the Tirol and Salzburg and were followed by similar upheavals in Innerösterreich.

Summary of developments in the late Middle Ages

In the late medieval period the Alpine lands were assembled by the Habsburgs into a monarchical union comprising about the extent of the present Austrian state. The process of union was at times intercepted and hindered by the partitions among the dynasty. When the process was finished, however, the territories still preserved their individuality and their own legal codes. During this period the towns developed and prospered, but in the rural settlements a backward tendency had set in. Many settlements were abandoned, especially in Lower Austria. The leading classes lost their interest in rural colonization as they found other and more lucrative sources of income. Mining developed, but trade was impaired by political instability. Until about 1450 the University of Vienna enjoyed some fame in the fields of theology and science. The literary culture of Austria was characterized by remarkable works, among them the rhyming chronicle of the Styrian Otakar aus der Geul, the work of the Carinthian abbot John of Viktring, the poetry of Oswald of Wolkenstein, and the works of the theologian and historian Thomas Ebendorfer. From the middle of the 15th century onward Austria came under the influence of Italian Humanism.

REFORMATION AND COUNTER-REFORMATION

The acquisition of Bohemia. The year 1526 saw the defeat and death of the Jagiellon king of Hungary and Bohemia, Louis II, who fell in the Battle of Mohács against the Turks. In view of the treaties of 1491 and 1515, Ferdinand I and the Vienna court envisaged Hungary and Bohemia plus the adjoining countries falling to the Habsburgs. Thus, the union of Austria, Bohemia, and Hungary became the leading concept of Habsburg politics. After clever diplomatic overtures, Ferdinand was elected king of Bohemia (October 23, 1526). In Hungary, however, there was a split election; János (John) Zápolya, *voivode* (governor) of Transylvania, was chosen by an opposition party, whereupon war broke out between the two candidates.

Ferdinand's troops in Hungary would have been in a stronger position had Zápolya not been assisted by the Turks under Süleyman I. In 1529 the Turks advanced as far as Vienna, which they besieged in vain. Another Turkish offensive came to a halt at Güns in western Hungary in 1532. Ferdinand, on the other hand, failed in his attempt to take Ofen (Hungarian Buda), where the Turks had entrenched themselves. By around the middle of the century the frontiers had become fixed. Hungary happened to be divided into three parts: the west and the north remained with the Habsburgs, the central part came under Turkish rule, and Transylvania and adjoining territory were kept by Zápolya and his successors. This situation was anticipated in the truce of 1547 and became formalized in the Peace of Constantinople (1562).

Ferdinand I's administrative reforms

During a short truce in the fighting against Zápolya and the Turks, Ferdinand started to reorganize Austrian administration. In 1527 he created new central organs: the Privy Council (Geheimer Rat) for foreign affairs and dynastic matters; the Court Council (Hofrat) as the supreme legal authority; the Court Chancery (Hofkanzlei), which served as the central office and only later on was to deal with internal affairs; and the Court Treasury (Hofkammer) for finance and budgeting. As the Court Treasury proved inefficient in the financing of the Turkish war, the Court Council of War (Hofkriegsrat) was established in 1556 to take care of the pay, equipment, and supplies of the troops, acquiring some influence on military operations as well.

The advance of Protestantism. The Protestant movement gained ground in Austria very fast. The nobility especially turned toward the Lutheran creed. For generations eminent families provided the protagonists of Protestantism in the Lower and Inner Austrian territories. The sons of the nobility were often sent to the north German universities to expose them more fully to Protestant influence. From 1521 Protestant pamphlets were produced by Austrian printers. Bans on them, issued from 1523 onward, remained ineffective, however.

Among the peasant population the Anabaptists had a stronger appeal than the Lutherans. As they had no support from the estates and because of their radicalism, however, the Anabaptists were persecuted from the start. In 1528 Balthasar Hubmaier, their leader in the Danube countries and southern Moravia, was burned at the stake in Vienna, and in 1536 another Anabaptist, the Tirolian Jakob Hutter, was put to death in the same way in Innsbruck after he had led many of his followers into Moravia. Ferdinand, for his part, advocated religious reconciliation and looked for means to achieve it; but the dogmatic viewpoints proved irreconcilable. The Peace of Augsburg (1555) finally brought some respite in the religious struggles.

When Charles V abdicated, Ferdinand I became emperor (1558), and thus the leadership of the empire was taken over by the Austrian (German) line of the Habsburgs. Maximilian II, the eldest son, followed his father in Bohemia, Hungary, and the Austrian Danube territories (1564). The next son, Ferdinand, was endowed with Tirol and the Vorlande; Charles, the youngest of the brothers, received the Inner Austrian lands and took up residence in Graz. Maximilian was known for his Protestant leanings but was bound by a promise he had given his father to remain true to the Catholic religion. The Protestants were therefore granted fewer concessions from him than they might have expected.

Catholic reaction and the Counter-Reformation

Meanwhile, Catholic counteractivity began, with the Jesuits particularly prominent in Vienna, Graz, and Innsbruck. A new generation of energetic bishops took part and proved a great asset to the cause. It was also of some importance that the monasteries, though they had been deserted by many of their members and were struggling for existence, had not been secularized. On the Protestant side, it proved impossible to reconcile the various reforming movements. Social differences between them, especially between the nobility and the peasants, also stood in the way of a united Protestant front. The Counter-Reformation scored its first successes in Gorizia and Carniola, where Protestantism had remained insignificant. And in other parts, official religious commissions started to replace the Protestant preachers with Catholic clergymen.

Rudolf II and Matthias. Maximilian's successor, Rudolf II (reigned 1576–1612), had been educated in Spain strictly in the Catholic faith. He had all Protestants dismissed from court service. The conversion of the cities and market centres of Lower Austria to Catholicism was conducted by Melchior Klesl, at that time administrator of the Vienna see but later to become bishop and cardinal. In Upper Austria, where the Protestants had their strongest hold, the situation remained undecided, with the Catholic governor, Hans Jakob Löbl of Greinburg, and the Calvinist Georg Erasmus of Tschernembl leading the opposing religious parties. When Charles's son, Ferdinand II, took over in Styria, he proved to be the most reso-

lute advocate of the Counter-Reformation. It was he who eventually succeeded in uprooting Protestantism, first in Inner Austria and then in the other Habsburg countries, with the exception of Hungary and Silesia.

From local skirmishes along the frontier, a long, drawn-out war with the Turks developed (1592–1606). In 1598 Raab (Hungarian Győr), which served as a bastion of Vienna, was temporarily lost; Gran, Veszprém, and Stuhlweissenburg (Hungarian Székes-fehérvár) passed several times from one side to the other. The introduction of the Counter-Reformation in Hungary, moreover, resulted in a rising of Protestant elements under István Bocskay. But in 1606 at Vienna a peace was concluded between Austria and the Hungarian estates. At Zsitvatorok another peace was negotiated with the Turks, who for the first time recognized Austria and the emperor as an equal partner.

Conflict between Rudolf and Matthias

Political disagreements between Emperor Rudolf, who to an increasing degree showed signs of mental derangement, and the rest of the family led to the "Habsburg Brothers Conflict." Cardinal Klesl in 1607 brought about an agreement among the younger relatives of the Emperor to recognize the Emperor's brother Matthias as the head of the family. As the conflicts with Rudolf persisted, Matthias strove also to come to an understanding with the estates, which were mainly Protestant. The formation of opposing religious leagues in Germany, the Protestant Union and the Catholic League, also added to the general confusion.

Matthias advanced into Bohemia, and, in the Treaty of Lieben (1608), Rudolf conceded to him the rule of Hungary, the Austrian Danube countries, and Moravia, while Matthias had to give up the Tirol and the Vorlande to the Emperor. In 1609 the estates received a confirmation of the concessions Maximilian II had made to them. The cities were guaranteed only in general terms that their old privileges should not be interfered with. At the same time, Rudolf II was forced to grant to Bohemia the so-called Letter of Majesty, which contained far-reaching concessions to the Protestants. After a final defeat of Rudolf in Bohemia in 1611, Matthias was crowned king of Bohemia. Rudolf's death in 1612 finally ended the conflict.

After Matthias had been elected emperor, his principal councillor, Cardinal Klesl, tried in vain to arrange an agreement with the Protestants in Germany. The ensuing years were filled with wars in Transylvania, where Gábor Bethlen came to power. In the Peace of Tyrnau (1615) the Emperor had to recognize Bethlen as prince of Transylvania, and in the same year he extended the truce with the Turks for another 25 years. In the meantime, war had broken out with Venice (1615–17) because of the pirating activities of the Serb refugees (Uskoken) established on the Croatian coast. A settlement was reached in the Peace of Madrid. The situation in Bohemia then reached a critical point, the religious tensions in the country finding a vent in the "Defenestration of Prague" (May 23, 1618), in which two of the Emperor's regents were thrown from the windows of the Hradčany Palace.

The Bohemian rising and the victory of the Counter-Reformation. War became inevitable when Emperor Matthias died in 1619. Not that he had been master of the situation, but his death brought Ferdinand II, the most uncompromising Counter-Reformer, to the head of the House of Habsburg. Ferdinand was hard pressed at first, as Bohemian and Moravian troops invaded Austria. A deputation of the estates of Lower Austria tried to make him renounce Bohemia in a peace treaty and demanded religious concessions for themselves, unsuccessfully, however. The Bohemians were forced to retreat, and imperial troops advanced into their country. The Bohemians deposed Ferdinand from the throne of Bohemia and elected Frederick V in his stead. But two days later Ferdinand II was elected German emperor at Frankfurt (August 28, 1619).

War over the Bohemian crown

War was the only means of resolving the issue. The conflict for the Bohemian crown developed into a European war when Spain, the Bavarian duke Maximilian I, and the Protestant elector of Saxony entered the struggle on the side of the Emperor. The Upper Austrian estates rashly joined Frederick, with the result that their country was occupied by the army of the Catholic League and afterward pledged to Bavaria. At the Battle of the White Mountain, Ferdinand became master of Bohemia, Moravia, and Silesia, while Lusatia was pledged to Saxony. King Frederick fled to the Netherlands. The leaders of the Bohemian rising were executed, and other nobles who had compromised themselves lost their property. Many Protestants left the country. In the new constitution of 1627 Bohemia and its associated lands became a hereditary kingdom. The diets were not dissolved entirely, as the government wanted to make use of their administration, but their influence was restricted to financial matters.

After the death of Matthias, Ferdinand had also inherited the Danubian territories. Tirol, however, retained a special status under a new Habsburg secundogeniture (inheritance by a second branch of the house). Upper Austria, pledged to Bavaria, was disturbed by a great peasant rising. The Protestant peasants were defeated after heavy fighting, and in 1628 the country passed into the hands of the Emperor again.

The Counter-Reformation was vigorously enforced in the Austrian domains. This led to the mass emigration of Protestants, including many members of the nobility. Most went to the Protestant states and to the imperial cities of southern Germany. After the Bohemian victory the war went favourably for the Emperor, and the Peace of Lübeck (1629) seemed to secure the hegemony in Germany for the Habsburgs. But in 1629 Ferdinand's attempt in the Edict of Restitution (Restitutionsedikt) to establish religious unity by force throughout the empire provoked the violent opposition of the Protestants.

The struggle with Sweden and France. July 1630 saw intervention in Germany's religious strife from a different quarter—Sweden. In that month the Protestant Swedish king, Gustavus II Adolphus, landed on the Baltic coast of Pomerania. His purpose was to defend the Protestants against further oppression, to restore the dukes of Mecklenburg, his relatives, who had been driven from their lands by Ferdinand's forces, and perhaps to strengthen Sweden's strategic position in the Baltic. In the ensuing conflict the city of Magdeburg was destroyed by fire after it had been taken by the troops of the Emperor under Gen. Johann Tserclaes, Graf von Tilly (1631). The north German Protestants, who had so far remained undecided, consequently went over to the Swedes. After victories near Breitenfeld and on the Lech, the Swedish troops entered Bavaria.

During the subsequent period of the Thirty Years' War, Ferdinand adopted a rigorous and often unrelenting attitude, though he yielded a little when the Peace of Prague was being negotiated (1635). His successor, Ferdinand III (1637–57), was as loyal to Catholicism as the father had been but showed himself more of a realist. He was not able, however, to prevent the war from again dragging into Habsburg territory, so that in 1645 even Vienna was threatened. The extremist party that had rejected all concessions lost its influence at the Vienna court, and two able diplomats, Maximilian, Graf von Trauttmansdorf, and Isaac Volmar, were entrusted with the representation of a weakened Austria at Münster and Osnabrück, where extended negotiations were conducted until acceptable terms could be settled for Austria. In the Peace of Westphalia (1648) Austria lost its possessions in Alsace, and Lusatia had to be ceded for good to Saxony. The peace in many respects marked the beginning of a new epoch. The Holy Roman Empire from then on was reduced to a loose union of otherwise independent states, and Habsburg politics shifted its emphasis, falling back entirely on the political, military, and financial resources of the hereditary Habsburg lands, now including also Bohemia. The new central organs and the administrative bodies of the territories took on much greater importance than the remaining institutions of the Holy Roman Empire. The Emperor came to rely on a standing army rather than upon troops provided by the German princes.

Economic and cultural consequences of the religious struggle

The heavy drain the religious wars had made on the population of the Austrian territories was compensated for by immigrants from the Catholic parts of the empire and by Croatian refugees from the southeast. The economic position of the peasants on the whole deteriorated. Many members of the nobility, as well as the church, acquired

new property. In mining, boom and depression followed quickly upon each other. The loss of many experienced miners during the Counter-Reformation resulted in difficulties, but the government took several steps toward improving and extending the salt mines. In 1625 it founded the Innerberg Union, under which the Styrian iron industry was reorganized. The Emperor also tried to interfere with the trade organizations of the towns, though without much success. Trade and finance in the Austrian territories was dominated by foreign capital.

The cultural life of the period was also dominated by the religious struggle. In the field of education the schools of the denominational parties rivalled each other. In 1585 a Jesuit university was founded at Graz, while at Salzburg a Benedictine university was established (1623). Austrian humanists produced some outstanding works of poetry and historical writings, and the sovereigns were great patrons of the arts, but on the whole this was an epoch dominated by Italian and Western influences.

AUSTRIA AS A GREAT POWER

After the Thirty Years' War, Austrian politicians were understandably reluctant to enter into another military conflict. In 1654 Ferdinand IV, the eldest son of the Emperor, died. His brother Leopold, who had been destined for a church career, then was considered as heir to the throne and was recognized as such by Austria, Bohemia, and Hungary. In Germany, however, difficulties arose when France declared itself against Leopold. Nevertheless, after the death of Emperor Ferdinand, Leopold was finally elected (1658), after having conceded constitutional limitations that restricted his liberty of action in foreign politics. West German princes under Johann Philipp von Schönborn, archbishop of Mainz, formed the French-oriented League of the Rhine. At the same time, Austria was engaged in the northeast, when it intervened in the war between Sweden and Poland (1658) in order to prevent the collapse of Poland. There were some military successes, but the Treaty of Oliva (1660) brought no territorial gains for Austria, though it stopped the advance of the Swedes in Germany.

During the Thirty Years' War the Turkish front had been quiet, but in the 1660s a new war broke out with the Turks (1663–64) because of a conflict over Transylvania, where a successor had to be appointed for György II Rákóczi, who had been killed fighting against the Turks. The Turks conquered the fortress of Neuhäusel in Slovakia, but the imperial troops succeeded in throwing them back. The Austrian military success was not, however, reflected in the terms of the Treaty of Vasvár: Transylvania was given to Mihály Apafi, a ruler of pro-Turkish sympathies. A minor territorial concession was also made to the Turks. The year after the Turkish peace, Tirol and the Vorlande reverted to Leopold I (1665), and the second period of the Habsburg partition (1564–1665) came to an end.

In Hungary dissatisfaction with the results of the Turkish war spread. Not only the Protestants, who were threatened by the Counter-Reformation, but also many Catholic nobles were alarmed by Habsburg absolutism. A group of Hungarian nobles and the Styrian count Hans Erasmus of Tattenbach entered into a conspiracy. The Austrian government, informed of their activities, had four of the ringleaders executed—an action that led to a rising by rebels known as Kuruzen (Crusaders).

In the meantime, the position of the Habsburgs in the west had again deteriorated. At first Leopold I's leading statesmen, Johann Weikhard, Fürst Auersperg (dismissed in 1669), and the president of the Court Council of War, Wenzel Eusebius, Fürst von Lobkowitz, remained rather passive in view of the expansionist policies of Louis XIV of France. They also stayed outside the Triple Alliance of Holland, England, and Sweden that was concluded in order to ward off the attacks of Louis against the Spanish Netherlands. When Louis actually invaded Holland, the Emperor finally entered the war, but in the ensuing Treaty of Nijmegen (1679) he had to cede Freiburg im Breisgau to France.

The threat to Vienna by the Turks

Another and still more menacing danger appeared in the southeast. After some deliberation, the leader of the Hungarian rebels, Imre Thököli, had asked the Turks for help, whereupon the grand vizier Kara Mustafa organized a large Turkish army and marched it toward Vienna. Habsburg diplomats succeeded in concluding an alliance between Austria and Poland. Meanwhile, imperial troops under Duke Charles of Lorraine tried to hold the enemy but had to retreat. From July 17 to September 12, 1683, Vienna was besieged by the Turks. Deciding against a direct assault, the Turks began to drill tunnels underneath the bastions of the city, when relief columns arrived from Bavaria, Saxony, Franconia, and Poland. King John III of Poland took over the command of the relieving army, which descended upon the Turks and dispersed them. The Emperor concluded a pact with Poland and the Venetian republic—the Holy League. In 1685 Neuhäusel was won back, and in September of 1686 Ofen was captured, despite fierce Turkish resistance.

In 1687 the Hungarian diet recognized the hereditary rights of the male line of the Habsburgs to the Hungarian throne. In 1688 Belgrade was conquered and Transylvania was secured by imperial troops. In the meantime, Louis XIV had begun an offensive against the German Palatinate. This meant that no further troops could be spared for the Turkish war, and in 1690 all recent conquests in the south, including Belgrade, were lost again. A victory of the imperial and the allied German troops under Margrave Louis of Baden near Slankamen (1691) prevented the Turks from advancing farther, but then the Margrave was ordered to the Rhine front. Eventually Prince Eugene of Savoy took over the command and gained a decisive victory over the Turks near Zenta (1697). After another offensive against Bosnia, the Turks finally decided to negotiate a peace. In the Treaty of Carlowitz (1699), Hungary, Transylvania, and large parts of Slavonia fell to the Habsburg emperor. In the meantime, the war in the west had come to an end (Treaty of Rijswijk, 1697), overshadowed already by the question of the Spanish succession.

The War of the Spanish Succession. From 1701 to 1714 Austria was involved in hostilities with France over the issue of the Spanish succession. The childless King Charles II of Spain, a Habsburg, had willed his entire possessions to a Bourbon prince—a grandson of Louis XIV of France. All those who disliked the idea of a French hegemony in Europe consequently united against the French. The Emperor declared war (1701) and was immediately supported by Brandenburg-Prussia and Hanover. In the spring of 1702 England and Holland entered the war in the Grand Alliance against France. Louis XIV, on the other hand, was able to win the electoral princes of Bavaria and Cologne as his allies. At this critical juncture another Hungarian rising, led by Ferenc II Rákóczi, occurred. The rebels were prepared to join forces with the enemies of Austria and for years engaged Austrian troops. The rebels even threatened Vienna, whose suburbs had to be fortified.

Scale of the Spanish war

In the war with France, imperial troops fought on four fronts: in Italy, on the Rhine, in the Spanish Netherlands, and in Spain. Much larger forces were mobilized than had been customary during the 17th century, with the result that the financial drain on the imperial treasury was so heavy that the Emperor had to resort to Dutch and English loans. When Bavaria entered the war on the side of the French, Austria was in further danger, until the Battle of Blenheim (1704), in which a joint English and Austrian army under the Duke of Marlborough and Prince Eugene defeated the French and Bavarian forces.

After a reign of 48 years filled with almost endless troubles, Emperor Leopold died in 1705. He was succeeded by his son, Joseph I (1705–11). In the religious quarrels the new emperor, an ally of Protestant states, showed great restraint and allowed himself to be guided mainly by political motives.

In 1703 the Duke of Savoy, who had left the French to go over to the Habsburgs, found himself in a critical situation; his capital, Turin, had come under French siege. An imperial army under Prince Eugene and reinforced by a Prussian contingent was sent to his aid and succeeded in uniting with the Savoyan forces and relieving Turin after a victorious battle (1706). At the beginning of the next year an agreement was reached under which the French eva-

cuated northern Italy. The same year a smaller imperial army under Wirich, Graf von Daun, conquered Spanish-ruled southern Italy; but an invasion of southern France, which the sea powers had instigated, failed. A quick success, however, fell to the Austrians in a campaign against the Vatican state over a conflict between the Emperor and the Curia concerning mutual feudal rights and because of Pope Clement XI's rather pro-French leanings.

The allies were victorious in the Netherlands, winning the Battle of Oudenaarde and conquering Lille (1708). Paris seemed within easy reach. The Battle of Malplaquet (1709) was another victory for the allies, but they had to pay dearly for it. In the meantime, peace negotiations had foundered. After reverses in Spain and a political change in England, the alliance itself was in danger of falling apart. The situation was further aggravated by the death (1711) of Emperor Joseph I, who left daughters only.

At this juncture, liquidation of the Hungarian rising became possible. Rákóczi, who in 1707 had declared the deposition of the Habsburgs, began to meet with growing opposition among his followers. Imperial troops forced Rákóczi to flee to Poland; and the rebels, who had been promised an amnesty and who were guaranteed religious liberty, made their peace in 1711. From then on the Vienna government tried to be more considerate of Hungary and its aristocracy.

The Peace of Utrecht

The election of Charles VI as emperor was effected without any difficulties. The English left the coalition, and after a military reverse most of the Habsburgs' allies joined the Treaty of Utrecht (1713). In the peace negotiations between Austria and France that were begun at Rastatt, Prince Eugene showed himself an unyielding and successful agent of Habsburg interests. Austria gained the Spanish Netherlands, a territory corresponding approximately to modern Belgium and Luxembourg. These gains were somewhat impaired, however, by the Dutch privilege of stationing garrisons in a number of fortresses. In Italy, Austria received Milan, Mantua, Mirandola, the continental part of the kingdom of Naples, and the isle of Sardinia. The Wittelsbachs of Bavaria regained their country, but the treaty contained an appendix that provided for the eventuality of Bavaria's being exchanged for the Spanish Netherlands. Of its gains, the north Italian territories were of the greatest value to Austria; the possession of Naples and the Netherlands, on the other hand, posed considerable military and political risks.

The problem of the Austrian succession. The extinction of the Spanish line of the Habsburgs and the fact that the emperor Charles VI was the last male member of that house posed serious problems for the Habsburg territories, which at the beginning of the 18th century were held together mainly by the person of the sovereign, notwithstanding the fact that there were some institutions of central administration. A settlement was made in the form of a family ordinance. On April 19, 1713, Charles VI issued a decree according to which the Habsburg lands should remain an integral, undivided whole. In the event of the Habsburgs' becoming extinct in the male line, the daughters of Charles or their descendants and, in default of any descendants of Charles, the daughters of Joseph I and their descendants and, after them, all other female members of the house should be eligible for the succession. As the son that was born to the Emperor in 1716 died after a few months, and only daughters were born to him after that (Maria Theresa, 1717; Maria Anna, 1718; Maria Amalia, 1724), this Pragmatic Sanction (a term used to characterize a pronouncement by a sovereign on a matter of prime importance) became of great significance. Austrian diplomacy in the last decades of Charles's reign was directed toward securing acceptance of the Pragmatic Sanction from all the European powers. It was published in 1720 and by 1722 had been recognized by the estates of all the Habsburg countries. Even the unanimous consent of the Hungarian diet was eventually obtained.

The Pragmatic Sanction of Charles VI

New conflicts with Turkey and the Bourbons. During the War of the Spanish Succession, Turkey had remained neutral toward Austria. But the Turks had attacked the possessions of the Venetians on the Pelopónnisos and the Ionian Isles. Austria tried to intervene and finally declared war. Prince Eugene defeated the Turks near the fortress of Peterwardein and conquered the strong bastion of Temesvár (1716). In the summer campaign of 1717 Belgrade again came into the hands of the imperial troops after a battle had been won against a Turkish relief army. In the Treaty of Passarowitz (1718) a frontier line was agreed upon that corresponded to the de facto situation. The Turks had to cede to the Austrians the Banat, the Turkish part of Syrmia, Walachia Minor as far as the Aluta, northern Serbia, Belgrade, and a strip of land along the frontier in northern Bosnia. A favourable trade agreement was also concluded.

During the Turkish war another crisis emerged. The Spanish minister Giulio Alberoni tried to initiate a policy of expansion in Italy. When Spanish troops landed in Sardinia and Sicily, the Emperor formed an alliance with Great Britain and France, later joined by the Netherlands (the Quadruple Alliance). After the English defeated the Spanish fleet, Madrid recalled its troops from the disputed territories. Austria received the more prosperous Sicily in exchange for Sardinia, which fell to Savoy. Charles then agreed to recognize the Spanish Bourbons. The gains from the Quadruple Alliance plus those of the Treaty of Passarowitz gave the Habsburgs the largest territory they were ever to rule. Their domains were far from unified, however, with the individual provinces showing a wide national, economic, cultural, and constitutional diversity.

Trading interests soon interfered with the alliance with the maritime powers. At first the attempts of the Ostend Company, which was backed by Charles VI, to enter into trade with India were quite successful. Because of the antipathy of the maritime powers, however, it seemed advisable to find an alternative to trade with Dutch and English colonial markets in the vast transatlantic empire of Spain. In 1725 Charles entered into an alliance with Spain, whereupon France, Great Britain, and Prussia formed a rival alliance. But soon after Russia was won over to the Habsburg cause, Prussia changed sides. As the outbreak of a European war seemed imminent, attempts were made at the Congress of Soissons to relax political tensions. Spain abruptly changed its alliances and concluded a treaty (1729) with England and France, the Netherlands joining in later. When Russia also began to waver, Prince Eugene tried to fall back on the traditional alliance with the maritime powers. After prolonged and difficult negotiations, England in 1731 accepted the Pragmatic Sanction, the Emperor in return giving a promise not to marry off his daughter Maria Theresa, the Habsburg heiress, to a prince who was himself heir to important domains. Austria finally dissolved the Ostend Company, having already suspended its charter in 1727. Charles VI then invested a great deal of energy in his endeavours to secure the recognition and the guarantee of the Pragmatic Sanction in the German diet. In this he was opposed by Bavaria and the elector of Saxony, but Austria finally obtained the guarantee of the Pragmatic Sanction at the Regensburg Diet (1732).

The question of the Polish succession

The question of the Polish succession led to a revival of the Austrian conflict with the Bourbon countries. Austria, with Prussia and Russia, favoured Augustus III of Saxony, the son of the deceased king, whereas France backed Stanisław I (Stanisław Leszczyński). On the military intervention of Russia in Poland, the Bourbons attacked Austria. The issue came to be mixed up with the problem of Lorraine, France dreading that on the impending marriage of Maria Theresa to Francis Stephen, duke of Lorraine, the latter's domains would be united with Austria's, so that French plans for the acquisition of Lorraine would be thwarted. France, Sardinia, and Spain simultaneously opened the war against Austria (1733). Prince Eugene, who was now aged, was able only to prevent a major success of the enemy on the Rhine. On the Italian front the Habsburgs fared even worse. The Battle of Parma ended undecided, but the Austrians were finally beaten near Guastalla. The small Austrian force that was stationed in southern Italy was unable to resist the Spanish attack, and Sicily and Naples were occupied by the Spaniards. In 1735 a Russian relieving corps reinforced the Habsburg front on the Rhine, and in northern Italy also there were a few successful operations of some local importance.

Direct contacts between Austria and France eventually led to the preliminary Peace of Vienna (October 3, 1735). Austria lost Naples and Sicily, which fell to a secondary branch of the Bourbons, and had to cede a tract of Lombard territory to Sardinia. As some compensation, Austria received Parma and Piacenza. Francis Stephen of Lorraine was promised Tuscany but had to renounce his hereditary duchy. On these conditions, France agreed to recognize the Pragmatic Sanction. The final peace was then concluded at Vienna in 1738.

Prince Eugene had died during the War of the Polish Succession. It soon proved disastrous that a successor of similar capacity of the prince was not found. During the second Turkish war of Charles VI (1737–39), Austria had joined in the Turkish-Russian conflict but without coordination of military operations. The Austrians, furthermore, underrated the Turkish forces and were themselves reduced by epidemics. The fortress of Niš was taken but was lost again soon thereafter. Peace negotiations conducted at Nemirov were broken off, and the war went on. The Austrians lost another battle at Grocka. Again peace negotiations were launched, in the course of which the larger part of the gains of the Peace of Passarowitz were lost. More disquieting even than the territorial losses was the loss in prestige. The epoch that had been the rise of Austria to a great power thus ended with reverses.

Social, economic, and cultural trends in the Baroque age. The Thirty Years' War and the Turkish wars had resulted in the devastation of large parts of the country and in great losses among the population, which suffered further reduction during the plague years of 1679 and 1713. The territories that had been wrested from the Turks had to be resettled systematically by German and other immigrants. The initiative for resettlement projects came from the official bureaucracy, the settlements being concentrated mainly in the south of Hungary. During the period of religious conflicts many Protestants had been exiled, but in the 18th century transportation to the various underpopulated parts of the empire was often resorted to.

Mercantilist policies after the 1660s

In the industrial and commercial field, mercantilist ideas, encouraged by the government, were prominent from the 1660s. The situation of the peasantry was thoroughly unfavourable. Tentative measures in the reigns of Leopold I and Charles VI to protect the peasants had little effect. Certain "model industries" (mostly textile factories) were established but were only partly successful. The economic policy of the absolutist state also resulted in strong interference with trade organizations. The guilds were suppressed or at least debarred from the new manufactures.

Trade was encouraged but yielded only small gains for the state. Industrial and commercial undertakings were managed in part directly by the state but largely through privileged corporations or private persons. Of some importance were the first (1667) and the second (1719) Oriental trading companies and the Ostend Company (1722). Trade in the Mediterranean was also intensified. Promising colonial ventures in India were discontinued for political reasons, however, in the middle of the 18th century. Under Charles VI new roads came to be planned and built on a large scale.

The state was in permanent want of money. This was a period of perpetual war as well as great economic investments, both entailing an excessive strain on state finances. At first the government resorted to the rich bankers such as Samuel Oppenheimer and his successor Samson Wertheimer for funds. Soon, however, it attempted to establish banking firms that were state controlled. The Banco del Giro, founded in Vienna in 1703, quickly failed, but the Vienna Stadtbanco of 1705 managed to survive; the Universalbancalität of 1715 was liquidated after a short period of operation.

After the victory of the Counter-Reformation, education was almost exclusively in the hands of the Catholic Church. The grammar schools of the religious orders, especially of the Jesuits and the Benedictines, set a very high standard for the most part. In 1677 another university was established at Innsbruck, the theological school of which was to acquire some fame. Historical writing flourished, the most outstanding works being those of two Benedictine brothers, Bernard and Hieronymus Pez; Gottfried Bessel, abbot of Göttweig; and Leopold I's official historiographer, the Jesuit Franz Wagner. The Austrian Jesuits were famous for their scientific and geographic researches, most notably the exploration of China.

Baroque poetry, theatre, and music

Among the achievements of Baroque poetry, mention should be made of Wolf Helmhart of Hohberg, whose works offer interesting insights into the life of the nobility, and of Katharina of Greiffenberg. The theatre of the Baroque was of great appeal, being remarkable for the splendour of its decorations and the ingenuity of stage machinery. The plays produced ranked from the elaborate Italian opera to the blunt humour of the popular play. Music attained an especially high standard, encouraged by three emperors who were composers themselves (Ferdinand III, Leopold I, and Joseph I). Charles VI was also a skillful musician, and he engaged the services of Johann Joseph Fux, who came from eastern Styria and developed into an important composer and teacher.

Austrian Baroque culture is, however, most clearly revealed by the splendours of its architecture. At first the field was dominated by the Italians, but soon native architects stepped forth. Preeminent was Johann Bernhard Fischer von Erlach (first plan of Schönbrunn Palace, Karlskirche in Vienna, Kollegienkirche in Salzburg) and his son Josef Emanuel (Hofbibliothek). They were rivalled by Jakob Prandtauer (Herzogenburg, Melk, and part of Sankt Florian monasteries) and especially by Johann Lucas von Hildebrandt (Schwarzenberg Palace, Belvedere, Peterskirche in Vienna, monastery of Göttweig). Among native sculptors Georg Raphael Donner was the first in rank and quality of work. Fresco painting was represented by Johann Michael Rottmayr from Salzburg, Daniel Gran from Vienna, and Paul Troger from the Tirolian Pustertal.

(E.Zö.)

FROM THE ACCESSION OF MARIA THERESA TO THE CONGRESS OF VIENNA

The war period, 1740–63. In October 1740 the Holy Roman emperor Charles VI, the last male Habsburg ruler, died and was succeeded by his daughter Maria Theresa, the wife of the grand duke of Tuscany, Francis Stephen of Lorraine. Until the election of her husband as emperor in 1745, Maria Theresa was referred to only as queen of Hungary and Bohemia. Her descendants represented the House of Habsburg-Lorraine.

Charles VI had established a reasonably unified order of succession for all the lands under the Habsburg sceptre (the so-called Pragmatic Sanction). By making broad concessions to foreign powers, he had secured their recognition of this act, and he died expecting a smooth succession for Maria Theresa. It was her misfortune that the poor state of Austria's military defenses during the last years of Charles's reign had vitiated his careful diplomatic manoeuvres, and it was of particular importance that four months prior to the Emperor's death another young ruler, Frederick II the Great, had succeeded to the throne of Prussia, which he wanted to raise to great power status. Thus it was that a major German state, which previously had been consistently loyal to the Austrian and imperial cause, became throughout Maria Theresa's entire reign the most determined foe of the Habsburg Empire. The specific issue in the conflict over supremacy in the German orbit was Frederick's intent to wrest the rich province of Silesia from the Habsburgs. The kings of Spain and of Piedmont-Sardinia and the electors of Bavaria and Saxony in full or in part challenged Maria Theresa's claims to succession, despite the fact that they had previously all acknowledged the Austrian heiress' right to rule. Frederick, on the other hand, cared less about the succession than about Silesia.

Conflict with Prussia over Silesia

By October of 1741 Frederick's army, then the best trained in Europe, had occupied Lower Silesia, and Maria Theresa in that month concluded a treaty in which she ceded this major and richest part of the province to Prussia under threats from other quarters. France, the hereditary foe of Austria and main inspiration of the anti-Habsburg coalition, had, in an alliance of Nymphenburg (May 1741), pledged to support Spanish and Bavarian claims on Maria Theresa's inheritance. Charles Albert, the

elector of Bavaria, was promised part of the Alpine Hereditary Lands and Bohemia, and, with French support, was elected emperor as Charles VII. On the day of his coronation, Austrian forces occupied his own capital, Munich.

Maria Theresa was helped by the fact that two major European powers, Great Britain and Russia, did not wish to see the Habsburg Empire dismembered. The British and the Dutch soon gave Austria active support, and Saxony and Sardinia dropped out of the anti-Habsburg coalition to support Maria Theresa, whose chances of consolidating herself on the throne thus became more promising.

But the situation remained serious. Habsburg forces still had to face Spanish troops in northern Italy, and in the spring of 1744 France declared war on Austria. One year later, in May, French troops defeated British and Dutch forces at Fontenoy in the Austrian Netherlands (the major part of present-day Belgium).

Frederick II of Prussia, who had watched developments closely after the definitive peace with Austria (Treaty of Berlin, July 1742), was afraid that the Austro–British–Dutch coalition might eventually be victorious and that Maria Theresa might be able to turn her main forces against him and regain Silesia. Stealing a march on the Habsburgs, he invaded Bohemia in August 1744. This time Austrian resistance was more determined, and though Frederick failed in his designs on Bohemian territory, he succeeded in confirming the victory of his first campaign by a new peace treaty, the Treaty of Dresden, December 1745. Having thus secured his Silesian spoils, he could afford to support the election of Francis Stephen of Lorraine as Holy Roman emperor in succession to Charles VII (died January 1745).

Francis Stephen becomes Holy Roman emperor

The war over the Austrian succession continued. As far as the Habsburg power was concerned, the results of the fighting in Italy, Belgium, and Holland were on the whole indecisive. A defensive alliance with Russia (1746) protected Austria from the danger of a new Prussian attack.

During these latter stages of the conflict, the question of the preservation of the Habsburg Empire was no longer an issue. The problem was merely the price that had to be paid for survival. In the long, drawn-out peace negotiations of Aix-la-Chapelle, Wenzel Anton, Graf von Kaunitz (subsequently Prince von Kaunitz-Rietberg, from 1753 to 1792 Austria's state chancellor) showed for the first time his mettle as a diplomat of consummate skill. Austria had to return some minor principalities in Italy to Spanish Bourbons until such time as the Bourbon lines there became extinct. A small frontier rectification in favour of Piedmont-Sardinia also had to be made, but neither of these adjustments was of great consequence. Far more important and painful was, of course, the loss of the major part of Silesia to Prussia, which opened the way to the rise of a new European great power as the most serious rival of the Habsburgs in Germany. It also meant a considerable drop in the number of Germans living within Habsburg lands, with important consequences in the rise of the national problem in the following century. On the other hand, Maria Theresa, a determined woman of courage, moderation, and charm, had managed to establish good relations with the frequently rebellious Magyars and had secured at least their nominal support at the height of the succession crisis.

The Habsburg Empire was not dismembered. Nor did it become a satellite state under the tutelage of other great powers. The outcome of the War of the Austrian Succession, except for the loss of Silesia, was a genuine defensive victory and proved to the world that Austria represented more than an agglomeration of lands under the same rule, acquired by wars and marriage contracts. In the two centuries since Ferdinand I, regent in the so-called Hereditary Lands, had become king in Bohemia, Hungary, and Croatia, a certain cohesion between the major historic units of the empire had clearly, after all, been established.

As far as the worldwide colonial aspects of the struggle were concerned, the results of the war were indecisive; certainly they were not to the advantage of Austria's traditional ally, Great Britain. A more permanent settlement of the conflict was reached only at the Treaty of Paris between Britain, France, Spain, and Portugal in February 1763—by and large to the disadvantage of the French, by then Austria's nominal ally.

During the brief era of peace—the eight years (1748–56) between the end of the War of the Austrian Succession and the Seven Years' War—Austrian military organization was revamped, and a state-wide system of conscription was introduced. The artillery in particular was greatly improved, and at the end of the period the Empress was ready once more to take up the struggle with Prussia. The new state chancellor, Kaunitz, had set an entirely new diplomatic stage for this difficult enterprise. The diplomatic actions taken at that time are generally referred to as the "reversal of alliances," actually a treaty system intended to isolate Prussia. It was in the interests of both Austria and France, hereditary enemies for two centuries, to become allies, since both were concerned about the rapid rise of an aggressive and unpredictable Prussia in the north. In the event of an Austrian reconquest of Silesia, the Austrian Netherlands were to be ceded to France. Russia, worried about Prussia's future designs in the east in regard to Poland, joined the coalition in January 1757, as did Sweden.

Cooperation with France against Prussia

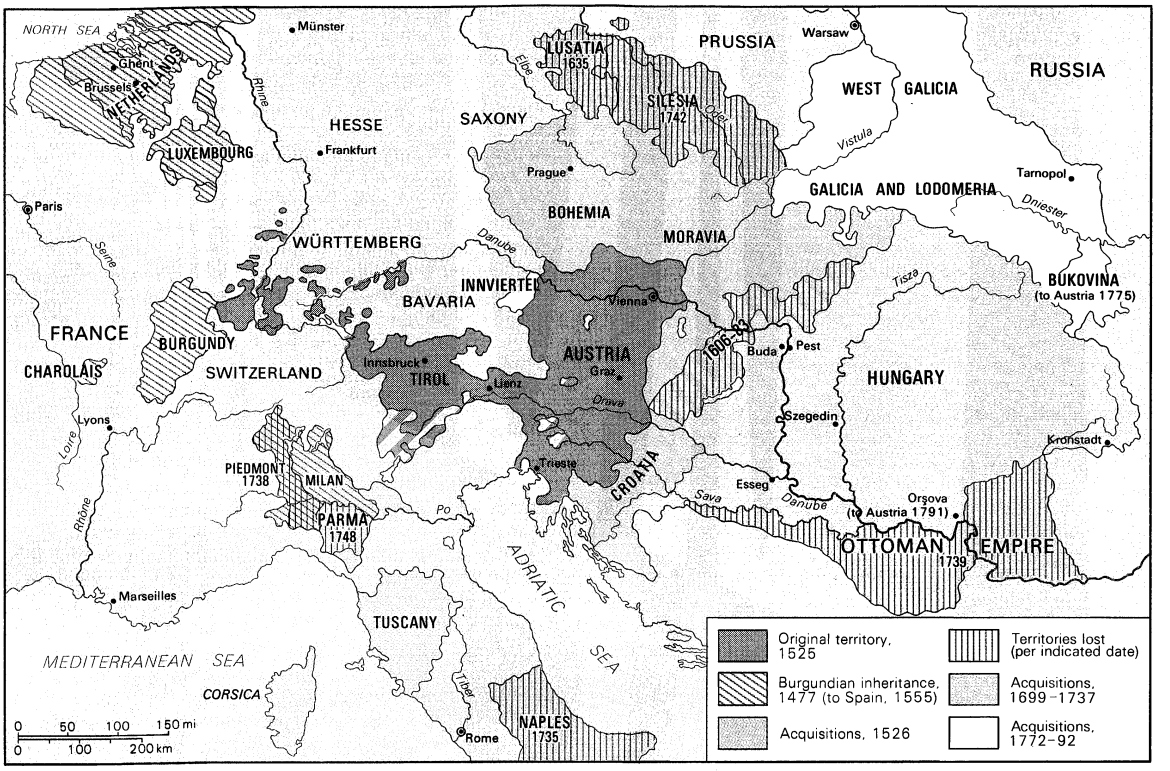
Prussia forestalled the Austrian diplomatic and military preparations by means of a surprise preventive attack in 1756. But it was able to counterbalance the weight of the formidable coalition rallied against it only by a subsidy treaty agreement with Britain, the former traditional ally of the Habsburg Empire. The tremendous superiority of the great coalition in territory and population figures was, however, in part offset by divisions among its members, in whose success Austria had by far the major stake. In a purely military sense, Frederick II had the full advantage of an inner line of defense. His brilliance as a military leader compared to that of the careful but too cautious Austrian commander in chief, Leopold Joseph, Graf von Daun, also counted heavily in his favour. The French army was in full decline, and the forces from the Holy Roman Empire outside of Austria, convoked by the emperor Francis of Lorraine against the Prussian aggressor, failed lamentably.

Though Austrian and Russian victories were sparse compared to a string of Prussian successes, the superior power of the coalition would have inevitably resulted in victory if the new tsar, Peter III, a great admirer of the Prussian King, had not withdrawn in 1762 from the war. Further designs to reverse Russia's position and to side openly with Frederick were foiled by Peter's overthrow a few weeks later. The initiator of the plot, his former consort and now ruling empress, Catherine II, did, however, sign a separate peace with Prussia on the basis of the status quo. Only a few weeks later the French withdrew from the war as well, since the obvious inability of the Austrian troops to regain Silesia killed France's hope for the acquisition of Belgium. In consequence of this, Maria Theresa, in the hardest decision of her reign, was forced to give in. Thus it was that the Treaty of Hubertusberg, concluded with Prussia in February of 1763, merely confirmed the outcome of the two previous Silesian wars. The only minor concession made by Frederick was a pledge to cast the electoral vote of Brandenburg (Prussia) in the next imperial election in favour of Maria Theresa's oldest son, Joseph.

The most decisive result of the Austro-Prussian Seven Years' War was the rise of Prussia to great power status. Austria, though weakened, remained a great power, and compensatory acquisitions for the loss of Silesia were impending. But, taking a long-range view, the Prussian victory represented a decision in the first round of the struggle for supremacy in Germany between the Habsburg Empire and Prussia, a conflict that the Habsburg Empire was to lose decisively within a century.

Foreign policy, 1763–92. In 1772 Austria participated in the first partition of Poland and acquired Galicia. The initiative for the partition came from Frederick II. Catherine II the Great of Russia would have preferred to have an undivided satellite Poland as her neighbour, but Maria Theresa, anxious to prevent a further shift of the balance of power to Austria's disadvantage, participated in the partition, though she regretted the breach of all traditions of international order. In 1774 the Habsburg Empire was given the Bukovina to the southeast of Galicia from the

Austria and the first partition of Poland



Expansion of the Austrian Habsburg domains until 1795.
Adapted from *Westermann Grosser Atlas zur Weltgeschichte*; Georg Westermann Verlag, Braunschweig

Ottoman Empire as reward for Austrian mediation in a Russo-Turkish conflict. These enterprises were due in part to the influence of the young Joseph II, who, after the death of his father, Francis, in 1765, succeeded as shadow emperor of the Holy Roman Empire and became co-regent with his mother in the Habsburg lands.

Joseph's ambitions went much further. In 1778, after the older line of the House of Wittelsbach had expired, he came to an understanding with the head of the younger Palatine line, according to which the Habsburgs would become heirs to the rule of Bavaria. Frederick II quite naturally opposed this agreement, which would much more than offset the loss of Silesia and would have given the Habsburgs renewed predominance in Germany. Joseph refused to yield, even though he was only lamely supported by his mother, and a new confrontation became inevitable. Military operations initiated in Bohemia in the summer of 1778 remained indecisive, however, and in May 1779 the Treaty of Teschen was negotiated between Maria Theresa and Frederick, both then in their old age and reluctant to fight another major war. Thus, Joseph had to put off his ambitious scheme, and the cession of the Innviertel by Bavaria, now incorporated into Upper Austria, had to serve as meagre consolation. Though Joseph's original plan would, in fact, never have been tolerated by the European great powers, in 1785, five years after the death of the great Empress, her intemperate son and successor reverted to his old scheme, this time proposing the exchange of Bavaria against the Austrian Netherlands. Again Frederick thwarted the plan, and Joseph's designs to undo the balance established by the Silesian wars were put to rest.

Joseph's foreign policy lacked success in other respects, too. His eagerness to barter the noncontiguous Austrian Netherlands for Bavaria resulted in part from the fact that he was frustrated in his understandable designs to undo the restrictions imposed on the Austrian rule there by the peace treaties of Utrecht and Rastatt of 1713 and 1714. These restrictions included the closing of the Scheldt River by the Dutch and Dutch rights to garrison the Austrian border fortresses against France at Austrian expense. But Dutch and French objections blocked Joseph's plan.

Failing thus in the west, Joseph hoped, as Russia's ally, to make at least some gains in a new Austrian war against the Turks. This struggle, started in 1788, ended the year after Joseph's death (1791), again in a meagre compromise. Thus the Emperor, whose great gifts included skill neither in diplomacy nor in military strategy, was in his conduct of foreign affairs far more unfortunate than most of his predecessors, who were not in other respects his intellectual peers.

Accession of Leopold II, 1790

His brother and immediate successor, Leopold II, heretofore Grand Duke of Tuscany, certainly was Joseph's intellectual peer, particularly in his ability as diplomat. He inherited from his brother a domestic situation in which Hungary and Belgium were in full revolt. In international relations he had to carry the mortgage of the sterile Turkish War begun by his brother, and he was faced as well by the possibility of a new partition of Poland, which might exclude Austria as beneficiary. But above all he had to deal with the spectre of an ever more serious revolutionary situation in France, which might well involve Austria in a long war of unpredictable outcome.

In domestic matters, Leopold, an enlightened ruler in Tuscany, had to retreat cautiously and regretfully from some of the reform legislation in the Habsburg Empire. In foreign affairs he had to end the Turkish War by the best possible compromise. As to the Polish question, encouraged by the spirit of reform in truncated Poland, he wanted to avoid a new partition altogether, but at the same time he wished to keep relations with Russia and Prussia on an even keel. This included, of course, an understanding in regard to the foremost international problem of his reign—the relationship to revolutionary France. Though naturally opposed to the revolutionary spirit, Leopold's course was remarkably free from considerations arising from close family ties, specifically the fate of his brother-in-law, King Louis XVI, and sister, Queen Marie-Antoinette. Leopold approved of the concept of a French constitutional monarchy, but to be prepared for all contingencies, in July 1790, he secured an understanding with Frederick William II, the new king of Prussia, concerning common interests of both countries in the east. At this time relations with France were aggravated further by the French seizure of the estates of great nobles in the Alsace, since they could be considered subjects of the Holy Roman Empire as well. The establishment of the headquarters of anti-revolutionary French political refugees in

the German Rhineland created additional tension. In a second meeting with Frederick William II at Pillnitz in Saxony in August 1791, the two rulers publicly expressed their concern with the French situation. But joint intervention would depend on British and Russian support. In February 1792 the Emperor considered it opportune to go one step further and concluded a defensive alliance with Prussia. Whether he would eventually have taken the final step and gone to war with France is conjectural; on March 1, 1792, he died unexpectedly. Less than two months later Austria did find itself at war, and while the declaration to that effect was made by France, the new emperor, Francis II, Leopold's son, a young man of modest abilities, had done little to prevent it.

The struggle with France, 1792–1815. In the ensuing era of almost continuous warfare that lasted for nearly a quarter of a century, the Habsburg Empire was involved more heavily than any other continental European power. This was partly because of its geographic location in the centre of Europe, partly because of Francis' position as Holy Roman emperor until 1806. The gradual transition from enlightened regimes of Joseph II and Leopold II to an increasingly conservative spirit under his rule was of additional significance. At the beginning of the war period Austria fought the spirit of the French Revolution. Supported by Spain, Portugal, and, above all, Prussia, it was engaged in the War of the First Coalition from beginning to end (1792–97). Yet Prussia dropped out of the war in 1795 to prepare for the third partition of Poland, after Prussia and Russia had jointly excluded Austria from the second partition. The Austrians fought against the French revolutionary armies in Belgium, Holland, and the Rhinelands with indifferent success. While the emperor Francis' far abler younger brother Archduke Charles checked the French in southern Germany, the young Bonaparte routed the Austrians in Italy, crossed the Alps, and invaded Carinthia and Styria. In April 1797 Austria had to agree to the armistice and Preliminary Peace of Leoben and in October of the same year to the permanent one of Campo Formio. The Habsburg Empire was let off relatively lightly. While the Austrian Netherlands (an indefensible province anyway) and Lombardy had to be ceded, territories of the Republic of Venice to the east of the Adige were gained. This forced participation in the destruction of the ancient Venetian republic put Austria in no better light than Prussia and Russia in the second partition of Poland. On behalf of the Holy Roman Empire, Francis also had to cede the left bank of the Rhine to France and accept a scheme by which the German ecclesiastical princes were deprived of their secular powers (Congress of Rastatt, 1797–99). This action further weakened Francis as Holy Roman emperor, although his position as ruler of the Austrian lands was strengthened somewhat by the third and final partition of Poland in 1795, in which Austria gained West Galicia as far as the Bug River.

Gains and losses at the Peace of Campo Formio, 1797

In the War of the Second Coalition (1798–1802), three major European powers—Prussia, England, and Austria—supported by Portugal, Naples, and the Ottoman Empire, represented the anti-French alliance. The theatres of war were again southern Germany and Italy, and, this time, Switzerland. When, in 1800, the Russian armies withdrew, the Austrian position became precarious. French advances in northern Italy and southern Germany were followed by the Armistice of Steyr in December 1800 and by the Treaty of Lunéville in February 1801. According to the terms of the latter, the provisions of the Treaty of Campo Formio of 1797 and the cession of the left bank of the Rhine to France were confirmed. The system of French satellite republics in Holland and Italy had to be recognized. Thus the main Habsburg territories were left intact, but Austria's position in Germany was further weakened.

Austria's chances in the War of the Third Coalition (1805–07) at first seemed promising. There was genuine military cooperation with Russia, and the coalition was able to count on the naval and financial support of Britain. Prussia, however, delayed a decision concerning its participation. The Austrians were defeated in Germany and the critical situation thus created forced their retreat from Italy as well. In November 1805 French troops occupied Vienna for the first time. On December 2 the Austrians and Russians were badly defeated by Napoleon at Austerlitz in Moravia, and the harsh Treaty of Pressburg was imposed on Austria hardly three weeks later. According to its principal provisions the Venetian territories gained in 1797 had to be ceded together with Tirol, Vorarlberg, Brixen, and the Trentino. The confirmation of Austria's incorporation of Salzburg and Berchtesgaden (1803) could not serve as adequate compensation for these severe losses. The Habsburg Empire had ceased to be a great power, and the Holy Roman Empire, in the face of French advances toward the east, was doomed.

Battle of Austerlitz, 1805

It had not been difficult to foresee these developments when Bonaparte proclaimed himself emperor of the French in 1804. To maintain Austria's position as best he could, and anxious to preserve the imperial title for his house, Emperor Francis proclaimed himself emperor of Austria on August 14, 1804. This purely declaratory manifesto, never submitted for the consent of the Estates of his lands as the Pragmatic Sanction had been a century before, pertained to all the Habsburg realms and might well therefore have been challenged, above all by Hungary. Yet this was not a major issue in the turmoil of the times. A month after Napoleon established the Confederation of the Rhine, Francis accordingly abdicated the empty title of Holy Roman emperor (August 1806). Invested only with an imperial title devoid of a great historical tradition, he ruled henceforward as Emperor Francis I of Austria.

The period from 1805 to 1809, when Austria was compelled to join the French continental blockade system, was one that saw the rise of feelings that were truly national, even though they were largely confined to the Austro-German orbit. A new, energetic, and idealistic foreign minister—Johann Philipp, Graf von Stadion—attempted to prepare the diplomatic ground for a renewed struggle against French expansion, and the archduke Charles organized a national militia. Encouraged by continued British resistance and by the guerrilla warfare of the Spanish people against the French occupation, Austria, on April 9, 1809, declared war on France. In contrast to a prostrate Prussia and to the German princes who catered to the whims of the foreign conqueror, the Habsburg Empire—bravely and vainly—tried to redress the balance in Europe. By mid-May 1809 Vienna was in French hands, though this time only after an artillery bombardment. The Austrians fought on, and a week later, at Aspern, on the left bank of the Danube opposite Vienna, Archduke Charles administered the first defeat to a French army commanded personally by Napoleon. He failed, however, to take advantage of this victory, and six weeks later the Austrians were decisively defeated at Wagram. Peace was dictated by the victors at the imperial castle of Schönbrunn in October 1809. This time Salzburg, a part of Upper Austria, and northern Tirol became parts of Bavaria; southern Tirol became part of the satellite kingdom of Italy, and West Galicia part of a new puppet Grand Duchy of Warsaw. The western parts of the Austrian southern Slav territories were ceded to France, and a heavy indemnity had to be paid.

A by-product of the defeat was the crushing of the heroic rising of the Tirolean peasants and the execution of their brave leader, Andreas Hofer. A far-reaching economic consequence of the war period was the state bankruptcy of 1811, which reduced the Austrian currency to one-fifth of its previous value.

Now a remarkable change in Austrian foreign policy took place. Shortly before the conclusion of the Treaty of Schönbrunn, Count (later Prince) Clemens Metternich, formerly ambassador to Napoleon's court, was appointed minister of foreign affairs in place of Stadion, whose anti-French policy had failed. Metternich conducted foreign policy for almost 40 years, until the outbreak of the Revolution of 1848. His first task was to shift from an anti-French policy, based partly on ideological grounds, to one of expedient cooperation until the day when new opportunities to restore Austria to its former rank as a great power would present themselves. He pursued this policy with great skill. One of his first moves was to persuade his imperial master to agree to Napoleon's request for

Metternich becomes minister of foreign affairs

the hand of Francis' eldest daughter, Marie-Louise (1810). Humiliating as acceptance of this demand by an upstart ruler was to an ancient dynasty, Metternich considered it conducive to Austrian interests.

In the War of 1812, Austria—France's new ally by marriage—was forced to put up an auxiliary corps under Karl Philipp, Prince zu Schwarzenberg, but Metternich's instructions were that any significant encounter with the Russian forces should be avoided. After the annihilation of the great French army in Russia and the continuation of the war in Germany by Russia and Prussia with the support of England and Sweden, the Habsburg Empire managed to stay neutral, thus raising the price for its subsequent intervention. It was clear that such intervention could only be on the side of the anti-French alliance, but Austria first offered Napoleon its "armed mediation," provided Napoleon would give up his overlordship of the German territories on the right bank of the Rhine and dissolve the Grand Duchy of Warsaw. The demand for the return of all Austrian territories ceded to France in the Treaty of Schönbrunn was raised. Napoleon, as expected, rejected the offer, and in August 1813 Austria became a partner in the War of Liberation, although the national issues previously raised by Stadion were played down by the conservative Metternich.

By its belated entry into the grand coalition, Austria had secured particularly favourable terms. Even though the Habsburg forces were smaller in number than those of Russia and Prussia, Prince zu Schwarzenberg was appointed commander in chief of the allied armies. Although a strategist of no better than average ability, he proved to be a skillful moderator among the ambitious generals of the coalition. His much abler chief of staff, Joseph, Graf Radetzky, must be credited with drawing up the plans for the great Battle of Leipzig (October 16–19), which broke the back of the French operations in Germany. Napoleon rejected an offer on the part of the allies that would have conceded the Rhine and Alp frontiers including Belgium and Holland to France. It is probable that Metternich at this time would have liked to keep a greatly weakened Napoleon in power as the best bulwark against the danger of a new French revolution.

On New Year's Eve, 1813, the war was carried into France. Had it not been for the tremendous superiority of the allies and the exhaustion of the French people, Schwarzenberg's indifferent strategy would hardly have secured victory. As it was, the allies were able to take Paris by the end of March 1814, force Napoleon's abdication, and secure the restoration of the Bourbons. The Treaty of Paris of May 30, 1814, gave France the relatively favourable frontiers of 1792. The reorganization of Europe was to be arranged at a congress that was to meet in Vienna in September 1814 under Metternich's chairmanship.

Austria and the Vienna Settlement, 1815

In the negotiations, which lasted until June 1815, Austria sided in general with Bourbon France and with England. Austria regained most of the territories that it had lost in the Treaty of Schönbrunn. Its former rule in the Austrian Netherlands and in southwest Germany was not restored, since Metternich saw little advantage in the control of noncontiguous areas. The Habsburg Empire was compensated in Italy: not only were Lombardy and the Venetia recovered; Tuscany, Modena, and—for the lifetime of Marie-Louise—Parma and Piacenza, also, were established as Habsburg semi-independent appendages in Italy. Napoleon's consort, now separated from him, was to rule in these two last-mentioned principalities. It is questionable whether the restoration of the Holy Roman Empire might not have served Austria's interests better than the establishment of the German Confederation (June 1815) under Austria's presidency.

In line with Emperor Francis' and Metternich's wishes, the confederation gave Austria, though only jointly with Prussia, far-reaching control over German affairs. This control was aimed in particular at preventing the establishment of representative constitutional governments and liberal institutions in any of the 50 German states. Reestablishment of the Holy Roman Empire would have struck a strong chord in the heart of the Austro-German peoples but this was precisely what Metternich, fearful of any expression of the popular will, did not want. Following the same line of thought, he believed that the dismemberment of Italy would prevent the future unification of the country under liberal national banners.

Metternich's skill restored Austria to great power status and even resulted in his diplomatic leadership as the so-called coachman of Europe. But the two major issues that were to severely weaken the position of the Habsburg Empire within the next half-century—the German and the Italian problems—were accentuated rather than solved by the settlement of 1815.

Reforms and their reversal, 1740–1815. The enlightened era in the Habsburg Empire comprises the reigns of Maria Theresa (1740–80) and those of her two brilliant sons, Joseph II (as co-regent from 1765 to 1780 and in his own right from 1780 to 1790) and Leopold II (1790–92). Maria Theresa, much better adjusted to reality than her immediate successor, drew chiefly on pragmatic experience. She wanted to secure the well-being of her subjects and to affirm the control of the state as far as compatible with traditions and customs. Joseph II intended to move faster and more in line with abstract, enlightened principles. He had little patience with references to tradition and was in some respects less prejudiced than his mother. Leopold II was, in his basic views, closer to his brother than to his mother, whom he resembled in his judicial temperament. While he would presumably have liked to expand Joseph's reforms, it was the main task of his all-too-short reign to quell the revolts in Hungary and the Austrian Netherlands and to settle the unrest in the Hereditary Lands and Bohemia. Thus he was forced to retreat in essence from the reforms of Joseph II to those of Maria Theresa. Under his son, Francis (1792–1835), the trend gradually became one of extreme conservatism and outright reaction, in part because of the preferences of this mediocre ruler but largely as a result of the counter-revolutionary spirit of the times.

Administrative reforms

The basic idea of the administrative reforms was to transform the estates system into a partially bureaucratic administration based on civil service rules, although Maria Theresa considered it expedient to preserve at least the external, but by then rather hollow, shell of the estates structure. Thus, on the provincial level, the speakers of the crownland Estates presided over most of the agenda of the provincial administration. On the top level of administration for the whole empire, the state chancellery of old was divided into a court chancellery entrusted with domestic agenda and a state chancellery proper for foreign affairs. An advisory state council was composed in part of great nobles and in part of high bureaucratic officials. A directory on public affairs for the Bohemian and Hereditary Lands was subsequently converted into the United Austrian and Bohemian Court Chancellery. A new commerce directory for the whole empire was established.

Maria Theresa introduced restrictions on the largely arbitrary patrimonial jurisdiction of the lords on their estates and transferred venue in matters of capital offenses to fewer and better courts. She steadfastly opposed the abolition of torture, though not on account of any basic cruelty of her character. She simply could not imagine how sufficient evidence could be gathered without forced confession. It was not until 1776 that she was finally persuaded to abolish torture. In this respect, Austria had lagged behind France, Prussia, and even Russia.

Joseph, who had much more understanding of judicial questions than did his mother, did much to improve civil and criminal procedure. He provided free legal counsel for peasants in litigation with lords. Although criminal justice administered on the basis of the strictly utilitarian philosophy of the Emperor was still extremely harsh, Joseph did much to improve civil law. The superb code of civil law of 1811, initiated in substance under Joseph, came to fruition only under Emperor Francis.

Economic policy

As to public finance, both the nobles and the church lost various privileges in regard to exemption from taxation. Craft guild restrictions, particularly those concerning the admission of apprentices, were loosened. Much was done to help the peasants in the newly gained trans-Carpathian territories (Galicia and Bukovina) by direct and indirect

government assistance, the abolition of customs duties for exports to other crownlands, and by the granting of various privileges for new settlers. The basically mercantilist economic policy of Charles VI's reign was somewhat changed in line with the influence of physiocratic and so-called populationist theories. Henceforth, skilled human labour, and not precious metal, was gradually to become the yardstick of national wealth. This led, on the one hand, to restrictions on emigration, and, on the other, to improvements in vocational training. Severe import restrictions on so-called luxury goods, which frequently had killed demand rather than encouraged the rise of competitive domestic industries, were somewhat modified. Joseph II went further. In regard to industrial and commercial enterprise, he was, much in contrast to his overall political philosophy, a strong supporter of private initiative.

Maria Theresa went a long way toward alleviating the lot of the unfree peasants. For the first time in Austrian history, the service obligations of the peasants were strictly defined. Above all, the ambiguous double role of the lord as landowner and official of local government was put under government supervision.

Joseph expanded the reforms of Maria Theresa. He further reduced the governmental functions of the lord and by legislation of 1781 abolished serfdom altogether outside of Hungary. Similar legislation was introduced in Hungary between 1784 and 1786. The mistaken belief of the Hungarian peasants that abolition of serfdom made them free owners of the land was, however, a major cause of the revolutionary situation at the time of the Emperor's death. The full conversion of the peasant's personal services outside of Hungary into obligations to be paid off in cash was somewhat precipitate, for money, in an agricultural economy still largely based on barter, was lacking. Leopold II reversed Joseph's peasant legislation so that things stood much as they had been at the time of the death of Maria Theresa. There was little further change until 1848.

Educational reforms of Maria Theresa

The greatest single achievement of Maria Theresa's reign was her educational reform at the elementary and intermediate levels. Three types of schools were introduced: (1) elementary schools in all but the smallest villages, in which reading, writing, and arithmetic were taught, with attendance being compulsory; (2) district schools in every administrative district, in which history, more advanced study of the vernacular language, geometry, drawing, and some vocational training were offered; and (3) so-called normal schools, established in the crownland capitals. Meant to be terminal schools for the urban middle class, they served also as teacher-training institutions. As for higher education, reforms under Maria Theresa were handicapped by her mistrust of—as she perceived them—radical enlightened ideas. She was, however, persuaded to transfer censorship from Jesuit control to a new and somewhat more liberal state agency. Some new chairs in the natural sciences and law were established at the University of Vienna, but progress in this respect was limited, and non-Catholics were altogether barred from graduating. Maria Theresa also founded some outstanding special schools. One of them was meant to train young nobles for public service, another, officers for the armed forces.

Joseph's reign proved to be disappointing as far as education was concerned. Censorship, it is true, was eliminated—at least as far as criticism of the government was concerned. But the Emperor's utilitarian objectives precluded practically any purpose in higher education other than the training of civil servants. The publication of literature that could not serve as a textbook for disciplines of immediate practical use was not encouraged. Under Emperor Francis, censorship was reintroduced in full force.

Radical church reforms are usually associated with Joseph's reign, but here also Joseph followed in the footsteps of his mother. Both rulers were devoutly religious, but both believed in firm state control of ecclesiastical matters outside of the strictly religious sphere. Following populationist doctrines, the Empress ordered restrictions of religious holidays and the prohibition of ecclesiastic vows prior to the 24th birthday. She insisted that clerics were subject to the jurisdiction of the state in nonecclesiastical matters. The acquisition of land by the church was to be controlled by the government. She took action against the Society of Jesus (Jesuits), but only in 1774, after the Pope had ordered its suppression.

Joseph's most radical measures were the issue of the Edict of Tolerance of 1781 and his monastic reforms. The edict and the legislation attached to it gave Protestants near equality and gave Jews the right to enter various trades, as well as permission to study at universities. In this respect the difference between the Emperor and his mother was fundamental. While Maria Theresa viewed Protestants as heretics and Jews as the embodiment of the Antichrist, Joseph fully respected other Christian denominations and entertained secret plans to establish an Austrian state church independent of Rome.

As to the monasteries, Joseph held that institutions not engaged in useful work for the community, above all agriculture, care of the sick, and education, should be dissolved. Consequently, about a third of the Austrian monasteries ceased to exist, their former members being ordered to learn skills adapted to secular life. The property of the dissolved institutions was used to pay for the upkeep of parishes and to finance the establishment of new parishes.

Control of church discipline and church property were further tightened by Joseph; seminaries for the training of the clergy were secularized. He even tried, without success, to simplify radically the Catholic liturgy. Many of his religious policies were discontinued in the reaction that followed, but the Edict of Tolerance and the monastic reform were maintained.

Joseph was unsuccessful in his efforts to achieve empire-wide administrative centralization. In particular, his attempt to enforce the use of German as the language of administration in Hungary was a failure and certainly accelerated the growth of Magyar nationalism and anti-German feeling there. On his accession, Leopold II was forced once more to recognize Hungary as a separate unit of the Habsburg lands.

THE AGE OF METTERNICH, 1815–48

Austria's leading diplomatic position after the conclusion of the second peace treaty of Paris in November 1815 was never anchored primarily in the power potential of the Habsburg Empire but in the skill of Metternich and his adviser, Friedrich von Gentz, in establishing a common conservative platform acceptable to east and west. Its ideological foundation was meant to be the Holy Alliance of the European powers of September 1815, as proposed by Tsar Alexander I and as opposed by the British government. It was meant to organize Europe on the basis of Christian authoritarian principles, implying the possibility of intervention by foreign powers in revolutionary or even merely liberal movements abroad.

The Concert of Europe

In the ensuing period of the Concert of Europe (an attempt to establish a directorate of the great powers over European affairs), which saw limited cooperation among the great powers at the conferences of Aix-la-Chapelle (1818), Troppau (1820), Laibach (1821), and Verona (1822), France was commissioned to put down a revolutionary rising in Spain and Austria was to suppress a revolt against the brutal oppression by the Bourbon regime in the Kingdom of the Two Sicilies. These actions caused Britain to withdraw from the concert of the five great powers. The shattering blow to Metternich's concept of the anti-revolutionary unity of the European powers, however, was Russia's support of the Greek independence movement between 1821 and 1830. Therewith his system in international relations had for all practical purposes broken down and could not be restored but merely patched up.

Two immediate problems for Metternich were the maintenance of Austria's hold in Italy and the struggle with Prussia for supremacy in Germany. As to Italy, the secret Carbonari Movement and the liberal Young Italy promoted by Giuseppe Mazzini were in full swing throughout the whole Restoration period. The French July Revolution of 1830, another defeat of the Metternich system, encouraged open revolt in the Austrian appendages of Parma and Piacenza and in the Papal states. Austrian troops had to restore the old order by force of arms.

The handling of the Italian question, while it did not really undermine Austria's power, gave Austria the image of a tyrannical oppressor of a freedom-loving, highly cultured people. That the Austrian administration in Lombardy-Venetia was neither corrupt nor by contemporary standards particularly inefficient counted little in its favour.

Even more serious was the German problem, which involved issues of real power for the whole Habsburg Empire. Such issues had great emotional impact on the Austro-Germans in general and in particular on the liberal intelligentsia. In their rejection of liberal, potentially revolutionary trends as dangerous to their rule, the German princes saw eye to eye, but the Emperor of Austria and the King of Prussia were deeply divided as regards the issue of supremacy in Germany.

The Wartburg Festival (1817) of German academic youth celebrating the tercentenary of the beginning of the Reformation had worried the princes deeply. The assassination of the playwright August von Kotzebue as an alleged Russian spy by a radical German student gave the conservative powers the opportunity for which they had been waiting. Under the leadership of Metternich and with the active cooperation of the increasingly reactionary Tsar Alexander, the Carlsbad Decrees were passed in August 1819. These put the German and Austrian universities under strict government control. Student associations were forbidden, censorship was strengthened. An investigatory commission was set up in Mainz, and teachers, writers, and students suspected of liberal views were blacklisted throughout Germany and Austria. In 1824 the German Confederal Assembly in Frankfurt renewed these provisions at the instigation of Metternich for an indefinite period. New oppressive measures on an even larger scale were again introduced at Metternich's behest by the German Confederation as answer to republican demonstrations at Hambach (Palatinate) in 1832 and at Frankfurt in 1833. The restrictions were in essence still in force at the time of the outbreak of the Revolution of 1848.

The Carlsbad Decrees, 1819

In the conflict between the two leading powers in the confederation, Prussia took the lead first in the economic field. In 1819 Prussia and its largely noncontiguous domains were merged into one customs association. Treaties of adherence by small neighbour states followed. By 1829 most German states had joined the association. There were some major exceptions—most notably Austria. The whole plan was indeed directed primarily against Austria. Counter movements such as a south German customs union and a more limited central German one collapsed. When, on January 1, 1834, the German customs union (the Zollverein) was completed, Frankfurt, Baden, and the Hansa cities stayed outside with Austria. Prussia had however won an important political as well as economic victory. It made the German north and west (the Rhineland) the centres of gravity of further industrial development.

At this time Metternich's power was not only externally but also internally in decline. In 1824 Franz Anton, Graf von Kolowrat, a great noble of Czech origin, was appointed minister of state. Somewhat more enlightened than Metternich, particularly in regard to the Austrian Slavs, he carried great weight in matters of domestic administration. The situation changed further when in 1835 the emperor Francis died and was succeeded by his eldest son, the feeble-minded crown prince Ferdinand. Metternich believed that even a feeble-minded prince could not be bypassed, but the actual power of the government was transferred to a state conference consisting of two archdukes (both of limited intelligence) and Kolowrat and Metternich as permanent members.

Metternich's decline

Metternich's last success occurred in 1846 when a revolt, initiated by Polish nobles in Galicia, was suppressed after the peasants, indirectly supported by the government, had turned against their oppressive masters. Because the revolutionary ferment had been fed from the city republic of Cracow, Russia and Prussia agreed that Austria should incorporate the ancient Polish coronation city into Galicia.

REVOLUTION AND COUNTERREVOLUTION, 1848–59

The revolutions of 1848–49. Metternich's victory was short-lived. Early in March 1848 a revolution began in Austria as soon as news about the success of the revolution in France had spread. In Vienna, crowds of people led by students and young academicians asked for liberalization of the regime. A clash with the military led to bloodshed. The riots, in which workers participated, thereupon spread further. The demand for the aged Metternich's resignation was met by the crown at once, for it was expected that after this symbol of reaction was overthrown, order would quickly be restored. But encouraged by this first and unexpected quick success, the revolution spread across the empire.

Three main aspects of the Austrian Revolution may be distinguished: social, democratic liberal, and national or multi-national.

As to the social revolution, the Habsburg Empire was less industrialized than any major European state west of Russia. Thus, while workers participated courageously in the uprisings in Vienna, the impact of their intervention was relatively small and had little effect on governmental policies. There existed a far stronger movement for the full emancipation of the peasants, which meant giving them clear title to the lands without any further obligation of service to the lords. The newly elected constituent assembly in Vienna, in September 1849, passed legislation to this effect. This legislation was sanctioned by the Emperor, but the actual enforcement of this sweeping reform, in particular in regard to the financial indemnity to be paid by the peasants, was left to the postrevolutionary regime. Still, the influence of peasant emancipation on the revolution was a powerful one. The largest and basically most conservative social class appeared to be satisfied, and by the autumn of 1848 it was no longer a major factor in the revolution.

Concerning the liberal democratic revolution, a moderately conservative new cabinet had by April 1848 drawn up a new constitution usually referred to as the Pillersdorf Constitution, after the name of the minister of the interior. This document was, on the whole, fairly democratic. At about the same time, a liberal government under Count Lajos Batthyány was formed in Hungary, in which some of the great leaders of the nation and the dominant figure of the more radical reform policy, Lajos Kossuth, participated. These men drew up a new constitution for Hungary. It provided far-reaching home rule for the country and, except for two provisions, was agreed to by the Emperor. The unresolved provisions were a demand for a separate budget and a more important demand that Hungarian troops not be called to action without the approval of the Hungarian government.

Developments in Hungary

In the meantime František Palacký, the great Czech historian and recognized intellectual leader of his people, had in a public letter rejected the participation of Czech representatives in the first freely elected German national assembly, which was to convene in May 1848 in Frankfurt. While the lands of the Bohemian crown were legally a portion of the territory of the Habsburg Empire that belonged to the German Confederation, Palacký held that the Czechs, who strongly affirmed the existence of a multinational Austria as a bulwark against Russian advances to the west, could never be part of Germany.

German Austrian liberals and enlightened conservatives participated gladly in the National Assembly in Frankfurt. One of the few enlightened, though not liberal, members of the imperial house, Archduke John, was elected temporary regent of the new Germany, while a moderate Austrian liberal, Anton Ritter von Schmerling, was appointed prime minister in September 1848. Inasmuch as the Frankfurt assembly had no executive power, these positions meant little in practical terms. The assembly immediately began to work on the draft of a constitution, and here the Austrian question represented one of the foremost divisive issues. The Grossdeutsch (Large German) faction wanted to include the Austrian Germans but was opposed to any overall association with the non-German majority of the Habsburg Empire. The Kleindeutsch (Small German) faction thought the only way to avoid this difficulty was to exclude Austria altogether. The solution adopted in the German federal constitution in May 1849 affirmed the incorporation of the German Austrian lands as long as they

had a constitution and administration separated from that of the non-Germanic parts of the Habsburg Empire. But the practical effect of these provisions was nil, because by the time the constitution was adopted the counter-revolution was in full swing. Hardly a month later the Austrian deputies at Frankfurt were recalled. The Prussian government followed suit and about a month later the remainder of the parliament, which had been transferred to Stuttgart, was dissolved by simple police action. In essence this meant the end of the revolution in Germany, but not the end of the Austro-Prussian conflict as it had developed during the revolution.

The Prussian government wanted to exclude Austria in a roundabout way from the confederation. A union scheme was therefore proposed by Prussia and put before a new, no longer revolutionary, assembly at Erfurt. Austria did not participate in these proceedings. According to the Erfurt plan, supported for a time by Saxony and Hanover, the German Confederation should be divided into two sections, an inner German section under the leadership of Prussia and a second, purely nominal, section consisting only of Austria in alliance with the inner section. The Austrian government, under its energetic prime minister, Felix, prince zu Schwarzenberg, rejected this plan and forced the states that supported Prussia to withdraw from the scheme. Finally, the old confederation was again called into session at Frankfurt. Schwarzenberg would in fact have liked to create a central European union including the whole of the Habsburg Empire. But this goal, which would have made Austria supreme in the whole of central Europe, was unobtainable.

Provisions of the Erfurt plan

In deciding to take his stand on a relatively minor issue, the question of the right of Prussia or Austria to intervene in a revolutionary situation in Electoral Hesse, Schwarzenberg forced Prussia to withdraw and hence to recognize the restoration of the Confederation of 1815 under Austria's presidency (the Punctation of Olmütz, November 1850). This was the last, and a merely diplomatic, victory of Austria over Prussia. It was achieved only because Tsar Nicholas I backed the Habsburg Empire, which seemed to him the more reliable conservative power. The Austrian triumph proved to be short-lived. It merely strengthened the Prussian resolve to undo the humiliation as soon as an opportunity offered itself.

Such an opportunity was not long in coming, since the lack of cohesion within the Habsburg Empire had become painfully obvious during the course of the revolution. While riots in Prague, Lwów, and Cracow could be quelled, events in Hungary took a collision course. The new government there, though eagerly introducing educational and social reforms, was at the same time set on a policy of rapid Magyarization. The result was risings among non-Magyar Hungarian national groups—Croats, Serbs, Slovaks, and Romanians. The conflict with the government in Vienna stiffened when the diplomatic Batthyány was replaced by the far more radical Kossuth, in September 1848.

A further aspect to be considered was Austria's involvement in a war with Piedmont-Sardinia, whose king, Charles Albert, intended to drive the Austrians out of northern Italy. His army invaded Lombardy in March 1848. By August, however, the Sardinians were decisively defeated in the Battle of Custoza. An armistice was concluded in late summer of 1848, but the war, because of violation of its provisions, was resumed by Sardinia in March 1849. After the new attack was repelled, peace was restored on the basis of the status quo in August 1849. Although the Sardinians were no match for the Austrian army, the Italian campaigns made it impossible for Austria to deal with the Hungarian revolution effectively.

The Slav problem

In the meantime, the Slavic question had become a major issue in the western part of the empire as well. In June 1848 a Slav congress was convoked in Prague, in which Czechs, Croats, Poles, Ruthenians, Serbs, Slovaks, and Slovenes were represented. Before the congress could terminate its deliberations, street riots offered the government a pretext to terminate the proceedings. Even though a final program of action could not be agreed upon, the results of the Prague Congress were of great importance in the history of Pan-Slavism. Beyond this, various plans for federalization and local autonomy within the empire were discussed.

By September 1848 the Hungarian situation had deteriorated further. Troops led by the ban (viceroy) of Croatia, Josip von Jelačić, entered the country under orders to restore royal authority. The action created general indignation in Magyar Hungary, and the imperial commander in Budapest was lynched during a riot. Thereupon the Hungarian Reichstag was dissolved by royal decree.

The situation in Hungary encouraged a new revolutionary rising in Vienna, in early October 1848. It was the last major attempt in the German-speaking orbit to regain the revolutionary initiative. The imperial minister of war, Theodor, Graf Latour, was lynched by a mob, and an imperial army under the command of the arch-reactionary Prince Alfred Windischgrätz occupied Vienna, meting out harsh retribution to the revolutionary leaders.

A momentous change took place on December 2, 1848. The feeble-minded Emperor Ferdinand, dubbed by official historiography "the Benign," abdicated in favour of his 18-year-old nephew, Francis Joseph, from whom a more energetic stand against the revolution could be expected. The first prime minister of the young Emperor was Schwarzenberg, a gifted conservative and an utterly ruthless and shrewd statesman, whose advice Francis Joseph accepted more indiscriminately than he did that of any other of his ministers.

Schwarzenberg was even less inclined to compromise in the internal affairs of the Habsburg Empire than in those concerning the German Confederation. Thus, an open break with Hungary occurred in mid-December 1848, and Windischgrätz's army marched into the country on a full wartime footing. Under this mediocre commander, the Austrians did poorly against the effective tactics of the insurgents. The Hungarian Reichstag, encouraged by this turn of events, in April 1849 declared that the Habsburgs had forfeited their right to rule. Hungary was accordingly proclaimed a republic under the presidency of Kossuth.

One factor in the Hungarian decision was the changed situation in Austria. By early March of 1849 the constitutional committee of the Reichstag, after prolonged and thorough discussion on a high intellectual level, had adopted the draft of an Austrian constitution (the Kremsier draft, after the Moravian town in which the Reichstag was reconvened). The draft provided for a constitutional monarchy, giving the sovereign only a suspensive veto against parliamentary legislation. Administrative organization along national lines was provided on the district level, yet crownland administration, the old historic boundaries, and the essence of the centralistic structure of the empire remained intact. A weakness of the draft was that it did not deal with the Hungarian problem. But, all things considered, the bill represented a reasonable compromise between the federalist and centralistic concepts, the former promoted mostly by Slavs, the latter primarily by German liberals.

The Kremsier constitution

The Kremsier constitution was the first reform plan drawn up by genuine representatives of the Austrian peoples, and it was precisely for this reason that the cunning Schwarzenberg decided to foil it. He had the Reichstag dissolved by police and warrants issued for the arrest of some of the reformers. At the same time he had the minister of the interior, Stadion, issue by decree a new, strictly unitary and centralistic constitution for the empire as a whole, including Hungary. The Stadion constitution was more conservative than the Kremsier draft, but it still subscribed to representative government. Enactment was, however, held in abeyance. Thus not only the Austrian peoples felt deceived but Stadion and other members of the Cabinet as well, since the constitution was never put into force and was even formally rescinded in December 1851. In the long history of Austrian government, the dissolution of the Reichstag of Kremsier was one of the most fatal governmental mistakes.

In March 1849 Piedmont-Sardinia had resumed military operations against Austria. This fact and the tense political situation in Germany caused Hungary's military situation to deteriorate rapidly. In April 1848 Windischgrätz was

replaced as commander, and after an interim the cruel and ruthless Julius, Freiherr von Haynau, was appointed as his successor. By mid-May 1849 the Hungarian revolutionary army had reconquered Budapest, presenting the young Francis Joseph with the painful alternative of risking his throne or asking the Tsar for help. Nicholas I, afraid of the possibility of the spread of the revolution to Russian Poland, complied with the Austrian Emperor's request. Russian armies entered Hungary and cooperated with Austrian forces. Budapest was reconquered by mid-July, and on August 13, 1849, the Hungarian troops capitulated. Kossuth and some of his followers managed to flee to Turkey. Even Tsar Nicholas recommended mercy for the gallant Hungarian military commanders. The answer of the Schwarzenberg-Haynau regime was the execution of 14 Hungarian generals—those who had surrendered to the Russians rather than to the Austrians—by hanging. About a hundred other executions, including that of Kossuth's moderate predecessor, Count Lajos Batthyány, followed. These actions as well as many long prison sentences and property confiscations imposed on minor rebels had only a limited deterrent effect, but they revolted public opinion across Europe. Hungary was dismembered and Transylvania, Croatia, and other areas were organized as separate crownlands under strictly authoritarian rule. The prostrate country was divided into five military districts and put under the administration of the stern Archduke Albrecht, whose intervention in the March 1848 Revolution in Vienna had made him hated by all liberals.

Defeat of the revolt in Hungary

All things considered, the revolution across the empire had not accomplished very much. Absolutism was seemingly more firmly entrenched than before, and the political clock had been put back beyond the regime of Maria Theresa. And yet a regime so badly shaken as Austria's could not hope to rule unchallenged in the future. The unresolved social, constitutional, and national issues became more intense, and new changes were soon in the offing.

The neoabsolutist era, 1849–60. The neoabsolutist era in Austria from the breakdown of the revolution in 1849 to 1860 must be judged from the point of view of the domestic policies of the regime—which were controversial but not entirely negative—and of a foreign policy that proved to be of disastrous consequence.

Positive domestic achievements were the establishment of a unified customs territory for the whole empire, distinct progress in industrialization, the promulgation of a code for trades and crafts, and some rudimentary beginnings of social legislation. Enactment of the full emancipation of the peasants, initiated by the revolutionary legislation, worked well. One-third of the cost was to be carried by the former owners of the land, one-third by the peasants themselves, and one-third by the government. This was, on the whole, a solution more equitable than that adopted in Prussia, and in Russia in the 1860s. Some improvements in standards were made in the universities and in the curricula of the Gymnasia. Overall credit for the reform policies belonged largely to Alexander Bach, the successor of Schwarzenberg, who had died in April 1852, lamented by few but the young Emperor.

Emancipation of the peasants

The regime's policies on other matters were more typically reactionary. Freedom of the press and jury and public trials were abandoned; corporal punishment by police orders was reintroduced. Informers flourished: the observation of the liberal reformer Adolf Fischhof that the regime rested on the support of a standing army of soldiers, a kneeling army of worshippers, and a crawling army of informers was exaggerated but not entirely unfounded. One of the most unwholesome developments was the conclusion in 1855 of a concordat with the Holy See that gave the church more power than it had possessed since the middle of the 18th century. Jurisdiction in marriage questions was handed over to the church, as well as control of censorship and elementary education. Church control extended indirectly to secondary education, also, because the priests, who were entrusted with compulsory religious education, had the right to see to it that instruction in any other field, be it physics or history, did not conflict with their teachings. In the second half of the 19th century, a regime of this type could possibly have stayed in power if it had been successful in the conduct of foreign affairs. But this was not to be the case.

Exclusion from Germany and Italy. In the protracted Crimean War crisis (1853–56) the Habsburg Empire took a stand against the Russian occupation of the Danube principalities (Moldavia and Walachia). By threatening military intervention on the side of the Western allies, Austria forced their evacuation by Russian troops and from 1854 occupied the territories itself for the duration of the war. In December 1854, Austria joined the Western alliance but refrained from a declaration of war on Russia and from active military participation. The continuing threat of such action forced Russia to accept the humiliating Treaty of Paris in March 1856.

The consequences for the Habsburg monarchy were grave. Russia, Austria's ally for well over a century, moved into the camp of its enemies. The tsars, Nicholas I and his son Alexander II, never forgave the alleged ingratitude for the Russian intervention in Hungary in 1849. The Western powers, on the other hand, were highly dissatisfied that the Habsburg Empire had refused to take the last decisive step and become a combatant in the war. In retaliation, and in payment of political debts, they supported the cause of Italian unification, skillfully engineered by the Sardinian prime minister, Count Cavour.

The end of Austria's alliance with Russia

Though Karl Ferdinand, Graf von Buol-Schauenstein, Metternich's successor as minister of foreign affairs, was blamed for allowing Austria to fall between two chairs, the fact is that the Habsburg Empire simply had been caught in an insoluble dilemma. By the nature of its own absolutist regime, it was drawn to tsarist despotism; but in justified fear of encirclement and of the effect of Russian-inspired Pan-Slavism on Austria's Slav population, the empire had to look for the support of the West with its alien political values. This was the basic contradiction in the policy of the Habsburg power.

Napoleon III, largely to distract attention from his own semidictatorial regime and to assuage the Liberals, deemed it advisable to support the cause of Italian independence. In the secret agreement of Plombières of July 1858, he pledged French military support for the liberation of Lombardy and Venetia. This was technically a defensive pact, but in face of a political situation in which three major European powers—France, Great Britain, and Russia—sided with the Italian cause, the Austrians would have been well advised to ignore any Piedmontese provocations. But Buol blundered Austria into an ultimatum by which in April 1859 it demanded Piedmont-Sardinia's demobilization within three days. This was all the justification Cavour and Napoleon III needed to start the Austro-French Piedmontese War. French and Italian military leadership and preparations were far from satisfactory, and conditions on

Adapted from *Westermann Grosser Atlas zur Weltgeschichte*; Georg Westermann Verlag, Braunschweig

Austrian Empire, 1815–59.

the Austrian side were even worse. Under an inferior commander, Franz, Graf von Gyulai, the Austrians hesitated to take the offensive. In the Battle of Magenta on June 4, they were defeated and Milan had to be evacuated. Gyulai was discharged and the inexperienced emperor Francis Joseph took over the supreme command. On June 24, 1859, the Austrians were decisively beaten in the Battle of Solferino, and two weeks later Napoleon III and Francis Joseph, who was deeply shocked by the carnage of the encounter, concluded the preliminary Peace of Villafranca, followed in November 1859 by the permanent Peace of Zürich. According to its terms Austria had to cede Lombardy—except for Mantua and Peschiera—to Napoleon III, who in turn would give it to Piedmont-Sardinia. The rulers in the Austrian appendages, Modena and Tuscany, who had been driven out by their subjects in the course of the war, were to be restored. These terms represented a kind of compromise. Napoleon III was not anxious to continue a war that might have brought about Prussian intervention on the side of Austria.

The Schleswig-Holstein crisis

The underlying conflict between Austria and Prussia over supremacy in Germany erupted into a crisis as a result of the involvement of the two powers in Danish affairs. When in 1863 the King of Denmark decreed a union of Schleswig and Holstein under one constitution and finally incorporated Schleswig into Denmark, he thereby violated international agreements and the charter of the German Confederation, to which Holstein belonged. Protests by Austria and Prussia on behalf of the Confederation were of no avail, nor were they meant to be. Consequently the two major German powers and Denmark went to war in 1864.

The Habsburg Empire's interest in this conflict was not as peripheral as it seemed, for the government was afraid that if action was left to Prussia alone the war would end with unilateral Prussian aggrandizement. Secondly, and perhaps equally important, the issue of Schleswig-Holstein was of major concern to German nationalists. After a brief war, Denmark was forced, by the Peace of Vienna of October 1864, to cede Schleswig, Holstein, and the small duchy of Lauenburg to the Austrian and Prussian sovereigns. But it became increasingly clear that Bismarck did not want to establish the principalities as states within the German Confederation under a sovereign prince. He worked instead in an underhand way for annexation by Prussia. A temporary arrangement was, however, reached at the Convention of Gastein (August 1865), according to which Holstein was put under Austrian military administration and Schleswig under Prussian. Lauenburg was sold outright to Prussia.

The real reason behind this temporary arrangement was the desire of both sides to gain time to make military preparations for the showdown that Bismarck considered inevitable and the Austrian government likely. Further efforts to settle the status of the duchies on a permanent basis by compromise failed, chiefly because this issue served Bismarck with a useful pretext for war. In further preparation Bismarck in April 1866 concluded an alliance treaty with Italy, which promised Italy Prussia's support for the conquest of the Austrian-held Venetian province. The treaty was to lapse if Italy did not open hostilities against Austria within three months. Thus a time limit was set for the beginning of the war.

The pretext for the openings of hostilities was a minor one. When Austria convoked the Estates of Holstein in early June 1866, Prussia declared the Convention of Gastein violated and occupied the duchy by military force. Thereupon Austria asked for mobilization of the confederal armies against the aggressor, Prussia. All member states agreed. Prussia in turn declared the charter of the confederation void and invaded Saxony, Hanover, and Electoral Hesse. The great war for supremacy in Germany, sometimes also referred to as the war between brethren or the Seven Weeks' War, began on June 16, 1866. Four days later Italy joined in the hostilities.

Francis Joseph's initial and grave error was his selection of the commander in the northern theatre of war. Gen. Ludwig August, Ritter von Benedek, a subcommander with many years of experience in the Italian theatre, was against his will put in charge of the main forces against Prussia. Archduke Albrecht, cousin of the Emperor, was given the far less risky command in the south against Italy, so that a member of the dynasty would not be compromised in case of defeat.

The Prussian forces, in accordance with the superior strategy of the chief of staff, Helmuth von Moltke, started a three-pronged attack on Bohemia. Austrian army units were defeated in most of the preliminary engagements in Bohemia. Benedek wanted to withdraw to Moravia, but when the Emperor appealed by implication to his sense of honour, he accepted battle on July 3 at Sadowa (Königgrätz) against three converging Prussian armies. The Austrians, their brave resistance notwithstanding, were defeated and withdrew in disarray. Before the Prussian offensive could be carried to the gates of Vienna, a temporary truce was arranged at Nikolsburg, followed by the preliminary peace that was concluded there on July 26. Bismarck, concerned with the possibility of immediate French intervention and with long-range plans for future friendly relations with Austria, wanted to avoid unnecessary humiliation of the Habsburg power. The terms that were confirmed by the permanent Treaty of Prague of August 23 were therefore moderate. Austria would have to recognize the dissolution of the German Confederation and the reorganization of Germany without its participation. Austria's rights in Schleswig-Holstein were transferred to Prussia and it had to pay a relatively minor indemnity. Yet no territorial cessions were demanded, and as a point of honour it was allowed to secure the preservation of the territorial integrity of its most faithful ally, Saxony.

Defeat by Prussia at Sadowa

Before evaluating the peace, the outcome of the war against Italy must be considered. By 1866 the Emperor and most Austrian statesmen had belatedly realized that Austrian rule in Venetia in the face of the opposition of the overwhelming majority of the population there made little sense. Nonetheless, a secret semi-official Italian offer to buy the province was rejected as dishonourable. A conditional offer to cede Venetia outright to Italy came too late in view of the Prusso-Italian alliance. Instead Emperor Francis Joseph concluded on June 12, 1866, a strange agreement with Napoleon III. According to its main provisions in regard to Italy, Austria would cede Venetia, win or lose, to Napoleon, who in turn would give the province to the new Italian kingdom. This meant in effect that the Habsburg power would fight a regular war for the sake of prestige and medieval chivalry, although the outcome was settled before the fighting started. The Italians in turn refused to accept a gift from Napoleon and preferred to fight. The Austrians defeated the Italians at Custoza in Lombardy as well as in the naval Battle of Lissa (Vis) off the Dalmatian coast. But this was of no consequence because of Italy's partnership with Prussia and the latter's victory at Sadowa. By the Peace of Vienna with Italy on October 3, 1866, the cession of Venice, this time to Italy directly, was confirmed.

Significance of the defeat

Of all the wars in the long history of the Habsburg monarchy prior to World War I, the brief Seven Weeks' War of 1866 had probably the most far-reaching effect. Its outcome banished Austria from the rank of the genuine first-rate powers. Beleaguered German nationalism within the empire increasingly assumed a belligerent tone, quite different from the moderate German-directed centralism that had been dominant for many generations. This spirit was challenged not only by a proud Magyarism but also by the combined force of the empire's Slavs—about half the population. This in turn made the Habsburg power increasingly vulnerable to Russian pressure from the east and within 13 years forced it into a German alliance, not as an equal but as a junior partner. Equally important were the internal changes brought by military defeat in two wars within seven years.

THE TRANSITION TO CONSTITUTIONAL GOVERNMENT, 1860–66

In March 1860 Francis Joseph ordered that the Reichsrat, a kind of empire-wide, purely advisory council of state, should be enlarged by the addition of 38 members proposed by the diets and selected by the crown itself. Only

in matters of public finance was this body given a share in legislation, yet it was entrusted with the formidable task of advising the emperor concerning the promulgation of a new imperial constitution. No agreement, however, could be reached. The moderately liberal centralists, represented largely by the Germans, demanded a strong empire-wide legislature and restriction of the agenda of the old Estates diets. They were faced by conservative federalists, largely Magyar, Czech, and Polish nobles, who wanted to strengthen the diets' position. Magyar influence was again on the rise. The Emperor himself sided naturally enough with the conservative forces, which were federalist mainly on the strength of historic and not ethnic claims. The result was the October Diploma of 1860, a constitution proclaimed by decree. It enlarged the diet to 100 representatives and broadened the legislative functions of the Reichsrat in matters of finance, commerce, and industry. Foreign and military agenda remained the exclusive domain of the emperor. Hungary's constitutional status within the empire was restored as it had existed prior to the revolution of 1848, but the concessions agreed to in March 1848 were not recognized. The federalists, particularly the Magyars, objected because their demands had been met only halfway. The centralists rejected the constitution just as strongly, as fully at variance with their claims. In effect, the whole unfortunate legislation had to be abandoned, and on the advice of the new German centralist cabinet, a new one, technically a revision of the October Diploma, was decreed (the so-called February Patent of February 26, 1861).

Constitutional changes

The October Diploma was really no constitution at all in the representative sense; the February Patent was at least an inadequate one. It provided for a bicameral system, an empire-wide house of representatives composed of dietal delegates, and a house of lords, consisting partly of hereditary members and partly of men of special distinction appointed for life. Furthermore, a parliamentary body for the Habsburg lands exclusive of Hungary was established. Opposition of the national groups rendered the constitution unworkable, and in 1865 it had to be suspended. Absolutism was, for all practical purposes, restored under a new prime minister, Richard, Graf Belcredi. The crown could hardly expect that this renewed elimination of constitutional government would be permanent, and the outcome of the war of 1866 made this assumption a certainty. In the meantime, negotiations with the Magyars, who were under the leadership of the highly respected moderate liberal Ferenc Deák and Count Gyula Andrássy, continued. (R.A.Ka.)

AUSTRIA-HUNGARY, 1867–1918

The liberal ascendancy. *The Ausgleich of 1867.* The economic consequences resulting from the defeat in the war of 1866 (Seven Weeks' War) made it imperative that the constitutional reorganization of the Habsburg monarchy, under discussion since 1859, be brought to an early and successful conclusion. Personnel changes facilitated the solution of the Hungarian crisis. Friedrich Ferdinand, Freiherr (later Graf) von Beust, who had been prime minister of Saxony, took charge of Habsburg affairs, first as foreign minister (from October 1866) and then as chancellor (from February 1867). By abandoning the claim that Hungary be simply an Austrian province, he induced Francis Joseph I to recognize the negotiations with the Hungarian politicians (Deák and Andrássy) as a purely dynastic affair, excluding non-Hungarians from the discussion. On February 17, 1867, Francis Joseph I restored the Hungarian constitution; a ministry responsible to the Hungarian parliament was formed under Count Gyula Andrássy; and in May 1867 Law XII was approved by parliament, legalizing what became known as the *Ausgleich* ("compromise"). This agreement was a compromise between the Hungarian nation and the dynasty, not between Hungary and the rest of the empire; and it is symptomatic of the Hungarian attitude that Hungarians referred to Francis Joseph and his successor as their king and never called him emperor.

Beust becomes chancellor

In addition to regulating the constitutional relations between the king and the nation, Law XII accepted the unity of the Habsburg lands for purposes of conducting certain economic and foreign affairs in common. The compromise was thus the logical result of an attempt to blend traditional constitutional rights with the demands of modern administration. In December 1867, the *engerer* Reichsrat, the section of parliament representing the non-Hungarian lands of the Habsburg monarchy, approved the compromise. Though, after 1867, the Habsburg monarchy was popularly referred to as the Dual Monarchy, the constitutional framework actually was tripartite, comprising the common agencies for economics and foreign affairs, the agencies of the kingdom of Hungary, and the agencies of the rest of the Habsburg lands—commonly but incorrectly called "Austria." (The official title for these provinces remained "the kingdoms and lands represented in the Reichsrat" until 1915, when the term "Austria" was officially adopted for them.)

Under the Compromise of 1867, both parts of the Habsburg monarchy were constitutionally autonomous, each having its government and its parliament composed of an appointed upper and an elected lower house. The "common monarchy" consisted of the emperor and his court, the minister for foreign affairs, and the minister of war. There was no common prime minister and no common cabinet. The common affairs were to be considered at the "delegations," annual meetings composed of representatives from the two parliaments. For economic and financial cooperation, there was to be a customs union and a sharing of accounts, which was to be revised every ten years. (This decennial discussion of financial quotas became one of the main sources of conflict between the Hungarian and Austrian governments.) There would be no common citizenship, but such matters as weights, measures, coinage, and postal service were to be uniform in both areas.

Although there was no common prime minister or cabinet, there soon developed the so-called *gemeinsamer* Ministerrat, a kind of crown council in which the common ministers of foreign affairs and war and the prime ministers of both governments met under the presidency of the monarch. The common ministers were responsible to the crown only, but they reported annually to the delegations representing both parliaments.

The December constitution

The Compromise of 1867 for all practical purposes set up a personal union between the lands of the Hungarian crown and the western lands of the Habsburgs. The Hungarian success inspired similar movements for the restoration of states' rights in Bohemia and Galicia. But the monarch who only reluctantly had given in to Hungarian demands was unwilling to discontinue the centralist policy in the rest of his empire. Public opinion and parliament in Austria were dominated by German bourgeois liberals who opposed federalization of Austria. As a prize for their cooperation in compromising with the Hungarians, the German liberals were allowed to amend the February constitution of 1861; the Fundamental Laws, which were subsequently adopted in December 1867 and became known as the December constitution, lasted until 1918. They granted equality before the law and freedom of press, speech, and assembly and protected the interests of the various nationalities, stating that

> all nationalities in the state enjoy equal rights, and each one has an Inalienable right to the preservation and cultivation of its nationality and language. The equal rights of all languages in local use are guaranteed by the state in schools, administration, and public life.

The authority of parliament was also recognized. Such provisions, however, were more a promise than a reality. Although parliament, for instance, did theoretically have the power to deal with all varieties of matters, it was, in any case, not a fully representative parliament (suffrage was restricted, and it was tied to property provisions until 1907); and the king was authorized to govern without parliament in the event that the assembly should prove unable to work. Austrian affairs from 1867 to 1918 were, in fact, determined more by bureaucratic measures than by political initiative; Josephine traditions (see above *Foreign policy, 1763–92*) rather than capitalist interests characterized the Austrian liberals.

Domestic affairs. After the December constitution had been sanctioned, Emperor Francis Joseph appointed a new Cabinet, named the "bourgeois ministry" by the press, because most of its members came from the German middle class (though the prime minister belonged to the Austrian high aristocracy). In 1868 and 1869, this ministry was able to enact several liberal reforms, undoing parts of the Concordat of 1855. Civil marriage was restored; compulsory secular education was established; and interconfessional relations were regulated, in spite of a strong protest from the Catholic Church. In 1870, the Austrian government used the promulgation of the dogma of papal infallibility as pretext for the total abrogation of the concordat.

The progressive legislation of the bourgeois Cabinet stood in sharp contrast to its inability to cope with the demands of the non-German nationalities. In 1868 the Czechs and the Poles issued declarations demanding a constitutional status analogous to that of the Hungarians. The government in Vienna did give the Poles in Galicia a considerable amount of self-government, which was later used to Polonize the Ruthenian minority. In 1871 a ministry for Galician affairs was set up, and the Poles remained the staunchest supporters of the Austrian government well into World War I.

Czech demands for autonomy

The bourgeois ministry was split into a liberal-centralist and a conservative-federalist faction; its members could not reach an agreement among themselves on policies to be adopted. The liberal members of the Cabinet opposed Czech demands; the conservatives showed themselves willing to consider them. Francis Joseph, indignant because of the anti-clerical policy of the liberals, dismissed the prime minister, Fürst Carlos Auersperg, in 1868, replacing him with the conservative Eduard, Graf von Taaffe, his boyhood friend. A period of indecision nevertheless persisted. The Emperor wavered between the liberals, whose anti-clericalism and parliamentarianism he disliked but with whom he sympathized in their centralist, German-oriented policy, and the conservatives, who had his favour in political legislation but aroused his fears by their demands for federalization. Neither Taaffe nor his successors, Leopold Hasner (from December 1868) and Potocki (from April 1870), could solve the Czech problem. The Franco-German War of 1870–71 temporarily diverted public attention from the Czech demands, though public opinion was divided strictly along lines of nationality: Austro-Germans celebrated the victories of the Prussian army, whereas the Slavs were decidedly pro-French. The Austrian government remained neutral because conflicting international interests had blocked the Austro-French negotiations that had culminated in a meeting of Francis Joseph and Napoleon III at Salzburg in 1867. The victory of Chancellor Otto von Bismarck and the establishment of the Second German Empire under the leadership of the Prussian king gave finality to the military decision of 1866. Austria was definitely excluded from the German scene, and a reorientation of dynastic interests seemed a logical consequence. Francis Joseph decided to explore the possibilities of satisfying the Czechs with some measure of federalism. On February 5, 1871, he appointed as prime minister Karl Siegmund, Graf von Hohenwart, a staunch clericalist. The driving mind in Hohenwart's Cabinet, however, was the minister of commerce, Albert Schäffle, an economist whose socialism may not have appealed to the Emperor but whose federalism did.

Failure to satisfy Czech demands

As a first step toward conciliation with the Czechs, the Hohenwart Cabinet dissolved parliament and the provincial diets. When the Bohemian elections improved the federalist position, Hohenwart proceeded to deal directly with the Czechs, copying in certain measure the method used to conclude the compromise with Hungary. Secret talks with Czech leaders František Ladislav Rieger and František Palacký led to the issuance of an imperial rescript by Francis Joseph on September 12, 1871, promising the Czechs recognition of their ancient rights and showing his willingness to take the coronation oath. The Czechs answered this rescript on October 10, 1871, by submitting a constitutional program of 18 articles, called the Fundamental Articles. According to this program, Bohemian affairs should be regulated along the principles of the Hungarian compromise, raising Bohemia to a status equal to Hungary. With this, Hohenwart, who had been up against violent German opposition from the very first day of his appointment, aroused Hungarian resistance, too. Andrássy, fearing that the Czech program could incite minority groups in Hungary, succeeded in convincing Francis Joseph that the stability of the Habsburg monarchy was endangered by the Czech program. On October 27, 1871, Hohenwart was dismissed and Francis Joseph returned the government to the hands of the German liberals.

The new prime minister, Prince Adolf Auersperg, entrusted the key ministries of his Cabinet to university professors and lawyers. The "ministry of doctors," as it was nicknamed by the people, concentrated on legal and administrative reform and endeavoured to strengthen the German control in parliament. After the dismissal of Hohenwart, the Czechs turned to passive resistance, withdrawing from the Bohemian diet and again abstaining from attendance at the parliament in Vienna. This attitude gave the government the chance to weaken the federalist position by introducing a bill for electoral reform. Instead of the existing modus, whereby the diets selected the deputies that were sent to parliament, the new bill set up electoral districts, each of which was to elect one deputy directly to the Reichsrat. The new system, however, preserved the old division of the electorate into *curiae* (socioeconomic classes), making parliament in this way a representation of German bourgeois interests.

The economic crisis of 1873

By a strange coincidence, the political victory of German capitalism took place at the very moment of a severe economic crisis. In April 1873, the opening of the Vienna International Exhibition had been thought of as a manifestation of the material progress and economic achievements of the Habsburg monarchy. The so-called *Gründerjahre,* or years of expansive commercial enterprise during the late 1860s and early 1870s, however, were characterized not only by railroad and industrial expansion and the growth of the capital cities of Vienna and Budapest but also by reckless speculation. Warning signs of an imminent crisis were disregarded, and in May, soon after the opening of the exhibition, the stock market collapsed.

The ensuing depression forced the government to abandon liberal bourgeois principles. The state took over the railroads and instituted public-works projects in an attempt to alleviate popular distress. A far-reaching consequence of the stock-market crash of 1873 was the permeation of anti-Semitism into Austrian politics. Jews were accused of being responsible for the speculative stock-market activities, even though official investigations proved that many elements of the population, including some ministers and high aristocracy, had participated in the *Gründungsfieber,* or "speculative fever," and the attendant scandals. The government survived the crisis, however, and German liberal political rule continued for five more years. German liberalism passed into eclipse not because of economic or domestic crisis but as a consequence of its opposition to foreign expansion.

International relations: the Balkan orientation. After his appointment as foreign minister, on November 14, 1871, Count Gyula Andrássy conducted the foreign affairs of Austria-Hungary with the intention of preserving the status quo. Discarding the anti-Bismarck bias of his predecessor, Beust, he sought the friendship of the German *Reich* with the intention of strengthening his position in the unavoidable confrontation with Russia over Balkan problems. The Three Emperors' League (*Dreikaiserbund*) of 1873, by which Francis Joseph and the German and Russian emperors agreed to work together for peace, gave expression to this policy. It also represented Andrássy's intention to strengthen Austria's position in a possible confrontation with Russia over Balkan problems, because the league made a change of the status quo in the Balkans dependent on German consent.

The Three Emperors' League of 1873

The continuing decline of Ottoman power encouraged the Balkan nations in their opposition to Turkish rule, and in 1875 there were revolts and upheavals. Andrássy failed to induce the Turkish government to adopt a reform program, and by 1876 Russian intervention seemed to be imminent. Russia offered to join with Austria-Hungary in

partitioning the Balkans between them, but Andrássy was convinced that Austria-Hungary was a "saturated state" unable to cope with more nationalities and lands, and thus he temporarily resisted the offer. He was aware, however, that Russia could not be restrained altogether; and thus, through Bismarck's mediation, there were concluded two secret agreements, at Reichstadt (Zákupy) in July 1876 and at Budapest in January 1877, whereby Russia gave up its plans for a "great partition" and settled for the territory of Bessarabia and, in return, acquiesced in Austria-Hungary's acquiring Bosnia and Hercegovina. Austria-Hungary and Russia further agreed, however, to refrain from intervention for the time being, and it was only when great-power mediation proved unable to settle the conflict between Serbia and Turkey that Russia declared war on Turkey, in April of 1877, after having once more secured Austro-Hungarian neutrality. In February 1878, with the war won, the Russians did not content themselves with Bessarabia, but, in the Treaty of San Stefano, violated Austria-Hungary's Balkan interests by creating a Great Bulgaria. Having Great Britain as an ally in his opposition to the Russian advance in southeastern Europe and Bismarck as an "honest broker," Andrássy managed at the Congress of Berlin in July 1878 to force Russia into retreating from its excessive demands. Bulgaria was broken up again, Serbian independence was guaranteed, Russia retained Bessarabia, and Austria-Hungary was allowed to occupy Bosnia and Hercegovina. Military occupation of the two provinces turned out to be more than the expected mere formality. It took 150,000 Habsburg troops and several weeks of fighting before the lands were under Habsburg authority. Since no agreement could be reached on whether the newly acquired lands should aggrandize the Hungarian or the Austrian part of the monarchy, an ingenious solution placed them under the jurisdiction of the common Habsburg ministry of finance.

Decisions of the Congress of Berlin, 1878

National conflict and reform. *Taaffe's ministry.* The German liberals had opposed the Balkan policy of Andrássy; and, out of fear that the Slav element in the monarchy would be strengthened by the addition of new Slav population, they voted against the occupation of Bosnia and Hercegovina, in this way withdrawing support from the government. When Prime Minister Auersperg resigned, the era of German liberal predominance came to an end. In 1879, the same year in which the Dual Alliance with the German *Reich* bound the Habsburg monarchy to Germany's foreign policy, the appointment of Taaffe as prime minister signified a reorientation in domestic affairs. From 1879 onward, the German element in the Habsburg monarchy was on the defensive, fighting stubborn and senseless rearguard actions against the Slav drive for political and national equality.

For the elections of 1879, a coalition had formed, consisting of clericals, German aristocrats, and Slavs, which gave itself the name of the Iron Ring. Taaffe had first tried to form a cabinet above parties. It was to include even the liberal Karl, Edler von Stremayr, who had presided over a caretaker government after Auersperg's resignation. The situation in parliament had decisively changed when the Czechs were persuaded by Taaffe, in 1879, to give up their boycott. In April 1880, as a first step, language ordinances were issued that made Czech and German equal languages in the "outer [public] services" in Bohemia and Moravia. In 1882 the University of Prague was divided, giving to the Czechs a national university. In the same year, an electoral reform reduced the tax requirement for the right to vote from ten to five florins, thus enfranchising the more prosperous Czech peasants and weakening the hold of the German middle class. The Taaffe government is also remembered for social-reform legislation; the laws of 1884 fixed the maximum working day (at 11 hours), outlawed the employment of children under 12, required a Sunday rest day for workers, and set up compulsory insurance against accidents and sickness.

Reforms of the 1880s

Despite the conservative character of the government, political life in the Habsburg monarchy underwent a decisive change during the Taaffe period. In the 1880s, the traditional party lineup decomposed, and new alignments and parties formed that were essentially radical and aggressive. The Slav orientation of the Taaffe Cabinet did not satisfy the Czechs but rather encouraged a mood of belligerence; because the moderate Old Czechs failed to live up to radical demands, the nationalistic Young Czechs were able to gain support from the electorate. Similarly, in German Austria and, especially, in Vienna, the moderate liberals were increasingly challenged by extremist groups—notably, German nationalists. In 1882 their "Linz program" proposed the restoration of German dominance in Austrian affairs by detaching Galicia, Bukovina, and Dalmatia from the monarchy, reducing relations with Hungary to a purely personal union under the monarch, and establishing a customs union and other close ties with the German *Reich*. This Pan-Germanic program found its chief protagonist in Georg, Ritter von Schönerer, a deputy to the Reichsrat, who also introduced, for the first time, a note of anti-Semitism into German nationalism. Although his version of extreme chauvinism and racialism never attracted more than a small number of followers, in a modified and moderate way Pan-Germanism and anti-Semitism became the ideological support of the bureaucracy and officer corps; though these elements did not favour union with Germany, they did feel that the Habsburg monarchy had the task of bringing German culture to the "inferior" non-German nationalities. The period also witnessed the founding of parties for the masses. While Schönerer and Pan-Germanism appealed to the educated classes, Karl Lueger transformed the Christian Socialism of Karl, Freiherr von Voegelsang, into a political organization which appealed to small shopkeepers, artisans, tradesmen, and lower bourgeois circles of Vienna and the surrounding countryside. The 1880s finally saw the transition of the workers' movement from the welfare and adult education societies into a political party. Although workers' movements had been weakened in Austria by personal rivalries and government persecution, Victor Adler in 1889, at a conference in Hainfeld, managed to unite the competing Marxist groups into the Social Democratic Party. Political life in Austria from 1890 well into the 1920s was dominated by these three movements originating in the 1880s: Pan-Germanism, Christian Socialism, and Democratic Socialism.

Reaction of the German nationalists

Taaffe continued to probe for compromises between nationalities that were becoming increasingly radical in their demands. In 1890 he tried to negotiate an agreement between the Old Czechs and the German liberals whereby Bohemia would have been divided for administrative and judicial purposes along lines of nationality, but he was balked by the more chauvinistic Young Czechs and German nationalists, and his efforts led to riots in Prague in 1893. When Emil Steinbach joined Taaffe's Cabinet as minister of finance in 1891, he encouraged Taaffe and the Emperor to try electoral reform as an instrument of breaking nationalist opposition. It was hoped that, by extending the franchise, nationalistic antagonism could be allayed and the growing unrest among urban workers could be placated. On October 10, 1893, a suffrage bill was introduced, giving the vote to virtually every literate adult male (though preserving the traditional system of voting in *curiae*). The conservative groups of all nationalities joined forces against this bill, and, under pressure from the Hungarian government, Taaffe had to resign on November 11, 1893. Though failing in political matters, the Cabinet had been successful in introducing some economic and social reforms: between 1888 and 1892 a system of cooperative banks for farmers was organized; the taxation system was revised; Austrian currency was stabilized by a return to the gold standard; and the florin was replaced by the crown, which remained the Austrian currency until 1924.

Taaffe's later attempts to lessen national tensions

Badeni's ministry. The franchise question continued to dominate Austrian domestic affairs and became closely welded to the nationality conflicts. Alfred, Fürst zu Windischgrätz, sought to win the support of parliament by forming a cabinet in which the clerical conservatives, the Poles, and the German liberals were represented. They were united, however, only in opposition to universal suffrage. Each minister defended his national cause, and the ministry was torn by ceaseless conflict. The end came in June 1895, when the government fulfilled an old promise

and introduced Slovene classes into the grammar school at Cilli (Celje), in Styria. Because the school had been exclusively German, this was regarded as a grave blow to the German cause, and the German liberals resigned, forcing Windischgrätz himself to resign.

Deeply embittered by the conduct of the German liberals, Francis Joseph on October 2 entrusted the task of solving the country's problems to a Polish aristocrat, Kazimierz Felix, Graf Badeni, known as a "strong man" for the high-handed way in which he had acted as governor of Galicia. Little noticed at the time, the appointment of Badeni symbolized the breakdown of German control over the Habsburg monarchy. For the first time in Habsburg history, the Germans controlled none of the key positions of government. Not only the Prime Minister (Badeni) but also the Finance Minister (Leo, Ritter von Biliński) and the Foreign Minister (Gołuchowski, who had succeeded Count Gusztáv Kálnoky von Köröspatak the year before) came from the Polish part of the empire. Badeni managed to induce parliament to accept a compromise franchise bill that introduced qualified universal male suffrage but preserved the system of class voting (a fifth *curia* was even added).

Significance of Badeni's appointment

The shortcomings of the new system enraged the parties representing the masses of the population. In the 1870s and 1880s, decisive economic changes with far-reaching social consequences had occurred in the Habsburg lands. Though remaining primarily agrarian, the Habsburg lands had undergone an industrialization that had resulted in an unprecedented growth of urban centres. Vienna, which had about 430,000 inhabitants in 1851, found itself a metropolis of 1,800,000 at the turn of the century; and this phenomenon was paralleled in other areas, especially in Bohemia, which had become the industrial centre of the western part of the Habsburg lands. The socioeconomic development naturally began to affect politics. From 1890 on, the advance of the Social Democrats and Christian Socialists caused considerable tension in Vienna. In October 1894, the Social Democrats were able to organize their first impressive, orderly mass demonstration in the capital, and the communal elections of 1895 had made the Christian Socialists the strongest party in Vienna, ending the long liberal rule. When the Emperor refused to confirm Karl Lueger, the popular leader of the Christian Socialists, as mayor of Vienna, there were demonstrations and protests. Not until Lueger was elected mayor the fifth time did Francis Joseph agree to confirm him, in April 1897. Furthermore, a few weeks earlier, the elections held on the basis of the new suffrage had strengthened the radical elements in the Reichsrat; the Young Czechs, for instance, had completely overwhelmed the conservative Old Czechs.

Growing strength of socialist parties in the 1890s

Counting on support from the Slav and conservative parties in parliament, Badeni dared to take up the Bohemian language question again. In April 1897, he issued a famous language ordinance that introduced Czech as a language equal to German even in the "inner service"—that is, for communications within government departments. This decision would have meant that civil servants in Bohemia and Moravia would have to be able to speak and write Czech as well as German. Since the Germans refused to learn Czech, this would have put them in a definite disadvantage in Bohemia's administration. The publication of the ordinance thus provoked violent German reactions: university professors signed resolutions of protest; mass meetings incited the public; and German deputies in the Reichsrat began to obstruct all legislative activities. The protest reached its climax in November 1897, when parliamentary sessions turned into bedlam, and popular protests against Badeni led to street demonstrations. The mass protest was not restricted to Vienna; it was even worse in some German towns in Bohemia; in Graz, clashes between soldiers and the masses ended in the death of one demonstrator. For a moment it seemed as if 1848 was about to return. To pacify the public, Francis Joseph gave in; on November 28, 1897, he dismissed Badeni, and asked Paul, Freiherr Gautsch von Frankenthurn, a former minister of education, to form a government out of the German parties of parliament. Gautsch's attempts to appease the Germans ran into obstruction from the Czechs. The stage of violence was shifted from Vienna to Prague and from the Reichsrat to the Bohemian diet. In March 1898 Gautsch was replaced by the former governor of Bohemia, Franz Anton, Fürst von Thun und Hohenstein, who in turn failed within a year. neither Manfred, Graf Clary und Aldringen, who formally revoked the Badeni language ordinance, nor his successor, Wittek, who headed a short-lived cabinet of a few weeks, managed to solve the nationality problem.

Koerber's ministry. Finally, on January 18, 1900, Francis Joseph asked Ernst von Koerber, a former minister of the interior, to form a new Cabinet. Koerber was the first and only commoner to be appointed prime minister by Francis Joseph; as a leading bureaucrat, he formed his ministry from the ranks of other bureaucrats, concentrating, in the subsequent years, on the administration of public affairs and economic programs rather than trying to deal with political problems. First by imperial decree and then, after some political bargaining, by consent of parliament, Koerber carried through a program of economic expansion, social legislation, and administrative reform, liberating the press from government and police control. By devious politicking, Koerber managed to keep government activities free from national strife, but he could not prevent national emotions from becoming more and more extremist. The national conflict came to be fought over educational matters, and in the final years of Koerber's government the desire for national universities aroused the sentiment of the Italians, Slovenes, and Ruthenians, turning the traditional Czech-German conflict into a multinational one. In December 1904, Koerber's various manoeuvrings faltered, and he was driven from office by a combination of parties.

Nationalism and electoral reform. The political climate in Austria was further complicated by the worsening of relations between the Emperor and the Hungarian government. Magyar separatists had agitated for the separation of the Habsburg army, and when Francis Joseph used an address to the troops at Chlopy in 1903 for an unequivocal reaffirmation of the common and unified character of his army, a controversy developed that had repercussions in the Austrian half of the Dual Monarchy. The plan to use universal suffrage to break the opposition in Hungary furthered the cause of political democracy in Austria. The demand for universal and equal suffrage had increased since the Russian revolution in the winter of 1905; after Koerber's Cabinet had run aground over a minor financial matter toward the end of 1904, Gautsch was chosen by the Emperor to introduce universal franchise in Austria. Though the first bill, introduced to parliament by Gautsch in February 1906, ran into the opposition of the middle-class and conservative parties that still controlled parliament, the realization of this program could no longer be blocked. Imperial interest and popular pressure—the Social Democrats had organized mass rallies to support the bill—combined to overcome parliamentarian opposition. After Gautsch had resigned in March 1906 and Prince Conrad von Hohenlohe-Schillingsfürst had failed to master the situation, Max Wladimir, Freiherr von Beck (prime minister from June 1906), managed to carry the bill through parliament. In January 1907 Francis Joseph sanctioned the law giving the vote to every male of 24 or over and abolished the *curiae.* Membership of the Reichsrat was increased from 425 to 516; the returns of the election of 1907 made the Germans now inescapably a minority, with 233 members, though certainly the strongest national group. The Czechs could count on 107 seats, the Poles 82, the Ruthenians 33, the Slovenes 24, the Italians 19, the Serbo-Croats 13, and the Romanians 5.

Universal suffrage

Universal suffrage brought the expected decline of the chauvinistic parties. The Young Czechs as well as the Pan-Germans were reduced to small factions without parliamentarian influence, whereas the Christian Socialists and the Social Democrats returned as the two strongest parties out of more than 30 represented in parliament; the Socialist delegation in the Austrian parliament was, in fact, larger than in any other country. The Austrian constitution, however, did not force the emperor to form

his government according to the composition of the parliament. Neither the Social Democrats nor the Christian Socialists were able to acquire any significant influence on the shaping of Austrian government affairs.

Beck remained in office and satisfied the Christian Socialists with some concessions but for the most part based his policy on the support of the conservative parties. In 1905 the diet of Moravia had succeeded in finding a compromise between German and Czech national demands, and it was hoped a similar compromise could be achieved for Bohemia. But, within a short time, national conflicts got the upper hand again, and parliamentary debate and public opinion were once more excited by national strife. In 1908, however, international complications diverted attention from domestic affairs.

Foreign policy, 1878–1908. *The alliances.* The occupation of Bosnia and Hercegovina in 1878 had reasserted Habsburg interests in Balkan affairs. Facing the possibility of conflict with Russia in this area, Austria-Hungary looked for an ally, with the result that in 1879 Austria-Hungary and the German Reich had joined in the Dual Alliance, by which the two sovereigns promised each other mutual support in the case of Russian aggression. The signing of the Dual Alliance was Andrássy's last act as foreign minister, for he resigned shortly afterward, but the alliance survived as the main element in the international position of the Habsburg monarchy until the very last day of the empire. Under Andrássy's successors—Heinrich, Freiherr von Haymerle, and Kálnoky—Habsburg foreign policy continued its conservative course.

In 1881 an alliance with Serbia, which after the Congress of Berlin turned to Austria-Hungary for protection, made this Balkan state a satellite of the Habsburg monarchy.

The Three Emperors' Alliance of 1881

The Three Emperors' Alliance (Russia, Germany, Austria-Hungary) of the same year brought Russian recognition of Habsburg predominance in the western part of the Balkan peninsula. The signatories of this alliance promised to consult one another on any changes in the status quo in the Ottoman Empire, and, while Russia was given assurances that its position regarding the Straits and Bulgaria would be recognized, Austria-Hungary received from Russia the promise that there would be no objection to a possible annexation of Bosnia-Hercegovina in the future.

The Three Emperors' Alliance was an important element in the structure of alliances that Bismarck set up to stabilize the European continent. Having decided to rely on Austria-Hungary as the fundamental partner in international affairs, Bismarck had to endeavour to neutralize all the areas in which the Habsburg monarchy might possibly be drawn into a conflict. It was essential to avoid being involved in a controversy at an inopportune moment and in a region of little interest to Germany. Bismarck therefore attempted to lessen the possibility of a conflict between Austria-Hungary and Russia by making them partners in the Three Emperors' Alliance. And when, in 1882, Italy approached Germany to find a partner in its anti-French policy, Bismarck used the opportunity to neutralize another European trouble spot. He informed the Italian foreign minister that the road to Berlin led through

The Triple Alliance

Vienna, with the result that the Triple Alliance (Italy, Germany, Austria-Hungary) was signed in May 1882. It was primarily a defensive treaty against a French attack on Italy or Germany. It further stated that, in the event of any signatory coming to war with another power, the partners of the alliance would remain neutral. The treaty did not settle the problems still existing between the Habsburg monarchy and the Italian kingdom, but for Bismarck it sufficed that they were neutralized.

In 1883 Bismarck acted once more to reduce the danger of war in "Europe's backyard" by arranging a defensive agreement between Austria-Hungary and Romania. The Triple Alliance and the Romanian Alliance not only strengthened the international status quo but also gave security to the internal order of the Habsburg monarchy by weakening the irredentist movements in Transylvania and the Italian parts of Austria-Hungary.

The deterioration of German–French relations in the following years convinced Bismarck of the indispensability of the Triple Alliance, and he made every effort to force Vienna to renew the alliance in 1887. By threatening to withdraw protection against Russian aggression, Bismarck forced Kálnoky to consent to his demands, but there can be no doubt that Austria-Hungary was clearly impeded in its national interests by having to adapt its foreign policy to the German and Italian demand for the isolation of France. Although Kálnoky succeeded during the negotiations in avoiding any new obligation in western Europe, he was less successful in defending more immediate Austrian interests. He managed to evade the Italian request for the support of an active Italian colonial policy, but he was unable to keep Italy out of involvement in Balkan affairs. It might be that in view of his own conservative and defensive policy he saw an advantage in having Italy as a third partner in the maintenance of the status quo against possible Russian expansion. At any rate, it was on Kálnoky's initiative that the original Italian demand for a declaration in favour of the status quo along the Ottoman coasts and the Adriatic and Aegean seas was extended to the interior of the Balkan peninsula.

Kálnoky's Balkan policy

On top of this Kálnoky granted the Italians the right to ask for compensation in case of any change in the territorial status quo without defining this term. In a certain way, all the differences and clashes between Austrian and Italian Balkan policy in the first decade of the 20th century can be traced to the introduction of this clause (later formulated in article VII of the treaty) at the renewal of 1887. In the same year, Bismarck built around the Triple Alliance a system of alliances and agreements which amounted to complete isolation of France and obliged the major European powers to guarantee the status quo along the borders of the Ottoman Empire. The First and Second Mediterranean Agreements of 1887 joined Great Britain to the powers (Austria and Italy) interested in blocking Russia from the Straits, and enabled Kálnoky to abandon direct agreements with Russia. The Three Emperors' Alliance of 1881 was allowed to expire, and Austria-Hungary was thus left without any formal understanding with Russia. Count Agenor Gołuchowski, who followed Kálnoky as foreign minister in 1895, decided, however, that direct relations with St. Petersburg should be renewed. In April 1897, Francis Joseph and Gołuchowski visited St. Petersburg. The agreements signed as a result of this initiative aimed at excluding Italy from Balkan affairs and sought to entrust preservation of the Balkan order to the bilateral cooperation of the two eastern monarchies rather than to a multilateral alliance system. The final years of the 19th century were marked by a change from static continental policy to a more dynamic world policy, and the ensuing mobility in international relations reduced the value of the Triple Alliance.

The Bosnian crisis. The Austro-Russian agreements of 1897 came to bear when, in 1903, a major revolt occurred in Macedonia. Following a meeting of Tsar Nicholas II and Francis Joseph in October 1903, their foreign ministers drafted a reform program for the Ottoman Empire. A mutual neutrality agreement was added the following year, leaving Austria-Hungary a free hand in the event of a conflict with Italy and enabling Russia to turn and face Japan. Explicitly excluded from the agreement with Russia were Balkan conflicts. When King Alexander of Serbia was assassinated in a military revolt in 1903 and the Obrenović dynasty was replaced by the Karageorgević, Serbian relations with the Habsburg monarchy deteriorated. The Serbs adopted an expansionist policy of unifying all southern Slavs in the Serbian kingdom, and, in order to block a Serbian advance, the Habsburg monarchy applied economic pressure. In 1906 all livestock imports from Serbia into the Habsburg monarchy were prohibited. This conflict, the so-called Pig War, did not crush Serbia but rather pushed it into the Russian camp.

When, in 1906, Count Gołuchowski was replaced as foreign minister by the former ambassador to St. Petersburg, Alois, Freiherr Lexa von Aehrenthal, a turning point in Austrian foreign policy was signalled. Aehrenthal made a belated effort to free Austria-Hungary from its submission to German interests and to engage in a dynamic Balkan policy. A first step in this direction was his proposal for the construction of a railroad through the Sandjak of

Novi Pazar. The combined Russian and Serbian opposition forced Aehrenthal to abandon the project temporarily and made it clear that any advance in the Balkans would probably result in war with Serbia and perhaps with Russia as well. The danger of such a conflict arose within a short time. In July 1908, following a revolution in Turkey, the Young Turk movement announced the reform of the Turkish constitution. Afraid that this constitutional change could undermine the Habsburg position in Bosnia and Hercegovina, two provinces that nominally were still under Ottoman suzerainty, Aehrenthal decided to use the opportunity to fortify the Austro-Hungarian position in the Balkan peninsula. In September 1908, he met with the Russian foreign minister, Count A.P. Izvolsky, and secured, so he thought, Russian approval of the proposed annexation in return for Austria's support in having the Straits opened to Russian warships. On October 6, 1908, the annexation was announced, immediately bringing a violent reaction from Serbia. When Izvolsky found that his plans for the Straits were opposed by Great Britain and France, he retracted his tentative support of Austria and supported the Serbian position. The situation became serious, and for a while war seemed imminent. Franz, Freiherr Conrad von Hötzendorf, the chief of the general staff of the Habsburg monarchy, who had long advocated preventive war, pushed for an aggressive move, but Aehrenthal had apparently never planned more than going to the brink of war. In March 1909, a German ultimatum forced the Russians to withdraw their support from Serbia, and, since the Turkish government had agreed to the annexation of the two provinces in return for a monetary compensation, Serbia also had to come to terms with the Habsburg monarchy. The Bosnian crisis was settled, but the Serbians felt their national pride deeply wounded and continued to stir unrest in the southern Slav provinces of the Habsburg monarchy.

Annexation of Bosnia and Hercegovina

The last years of peace. *Conflicts of nationality.* The annexation crisis had repercussions among the other Slav nationalities in the monarchy. For several years, Czechs had been attracted by the Pan-Slav movement, and in July 1908 a Pan-Slav congress was held in Prague. During the diplomatic crisis of the following winter, the Czechs unabashedly took the side of the Serbs, and, on the day of the 60th anniversary of Francis Joseph's accession to the throne, martial law had to be declared in Prague. National strife broke out once more all over the monarchy, and parliamentary activities were all but blocked by filibustering and the riotous activities of the deputies. Prime Minister Beck had resigned in November 1908; and his successor, Richard, Freiherr von Bienerth, after having accomplished little with a cabinet of civil servants, tried to appease the nationalities by including *Landsmannminister* (national representatives) into his Cabinet (February 1909).

Obstruction in parliament continued. The Germans, in control of the government and the central administration, continued to assign to the monarchy the role of an outpost of German culture; the Slavs, however, increasingly wanted to make Austria the home of Slav national aspirations. The Czech agrarian leader František Udržal stated in parliament: "We wish to save the Austrian parliament but we wish to save it for the Slavs of Austria who form two-thirds of the population." A population census taken in 1910 confirmed the Slav claim: out of the 28,324,940 inhabitants of the western half of the Austro-Hungarian Empire, only 35.58 percent regarded themselves as German; 17.77 percent Poles; 12.58 percent Ruthenians; 23.02 percent Czechs and Slovaks; 4.48 percent Slovenes; 2.8 percent Serbs and Croats; and 2.75 percent Italians. The Slav predominance was weakened by the attitude of the Poles, who remained loyal to the central government, thus allowing the national conflict to assume the character of a primarily Czech–German quarrel. Even the Social Democratic Party could not overcome nationalist antagonism. In 1899, at the party congress at Brünn, the Social Democrats had presented a national reform program based on democratic federalism, granting the right of national decisions to territorial units formed on a basis of nationality. Karl Renner and Otto Bauer, who later became leaders of German–Austrian Socialism, drafted various programs for the solution of the nationality problem in books published between 1900 and 1910. But these efforts could not prevent the Socialists from splitting along national lines, too, and in 1910 the Czech Socialists declared themselves independent of the Austro-German Socialist Party.

National composition of the empire in 1910

Party rivalries. Such national differences weakened the Socialist position in the elections of 1911. Over 50 parties had competed in the campaign, and, since the German nationalist parties had allied in the Deutscher Nationalverband, they managed to return to parliament as the strongest single party, gaining 104 seats out of 516. The Christian Socialists, weakened by personal rivalry, suffered heavy losses, winning only 76 seats. The German Social Democrats received 44 seats and the Czech Socialists 24. The Czech parties were badly divided, those representing the Czech middle class gaining 64 seats. Bienerth found himself unable to form a workable ministry, and he was replaced by Gautsch, who tried to reconcile the Germans and the Czechs. For a while negotiations seemed quite successful, but extremist incidents deadlocked the talks, and the Gautsch Cabinet was replaced by a new ministry headed by Karl, Graf von Stürgkh (November 1911). Unable to deal with the nationality problem in a parliamentarian fashion, Stürgkh repeatedly suspended the Reichsrat. It was characteristic of the general political climate in Europe that Stürgkh had to concentrate his legislative program on the improvement of Austrian armament, for international crises overshadowed the nationality conflict.

Stürgkh's suspensions of the Reichsrat

Conflict with Serbia. Ever since the Bosnian crisis of 1909, Austrian diplomats had been convinced that war with Serbia was bound to come. Aehrenthal had fallen sick soon after the annexation of Bosnia and Hercegovina, and after a long illness he died (February 1912), at a moment when an Italian–Turkish conflict over Tripoli had provoked anti-Turkish sentiment in the Balkan states. Leopold, Graf Berchtold, who directed Austro-Hungarian foreign policy from 1912 on, did not possess the qualities required in such a critical period. Aehrenthal had been able to silence the warmongering activities of Conrad, who continued to advocate preventive war against Italy and Serbia, but Berchtold yielded to the aggressive policies of the military and the younger members of his ministry. During the Balkan Wars (1912–13), fought by the Balkan states over the remnants of the Ottoman Empire, Austria-Hungary twice tried to force Serbia to retract from positions gained by threatening it with an ultimatum. In February and October 1913, military action against Serbia was contemplated, but in both instances neither Italy nor Germany was willing to guarantee support. Austria-Hungary had to acquiesce in the territorial changes in the Balkan peninsula, changes that eliminated the Turks from Europe. By supporting Bulgaria against Serbia, Austria-Hungary alienated Romania, which country had shown resentment against the Habsburg monarchy because of the treatment of non-Magyar nationalities in Hungary. Romania thus joined Italy and Serbia in support of irredentist movements inside the Habsburg monarchy. By 1914, leading government circles in Vienna were convinced that offensive action against the foreign protagonists of irredentist claims was essential to the integrity of the empire.

The assassination of Francis Ferdinand

In June 1914, Archduke Francis Ferdinand, the heir of Francis Joseph, participated in army manoeuvres in the provinces of Bosnia-Hercegovina, disregarding warnings that his visit would arouse considerable hostility. When Francis Ferdinand and his wife were assassinated by a Bosnian nationalist at Sarajevo on June 28, 1914, the Austro-Hungarian foreign office decided to use the opportunity for a final reckoning with the Serbian danger. The support of Germany was sought and received, and the Austro-Hungarian foreign office proceeded to draft an ultimatum putting the responsibility for the assassination on the Serbian government and demanding full satisfaction. The attitude of the foreign office was shared by Conrad and by Stürgkh but was opposed by the Hungarian prime minister, Count István Tisza, who wanted an assurance that a military move against Serbia would not result in territorial acquisitions and thus increase the Serb element in the monarchy. His demand satisfied, Tisza joined the advocates of war. In ministerial meetings on

Peoples of Austria-Hungary in 1914.
Adapted from W. Shepherd, *Historical Atlas;* Barnes & Noble Books, New York

July 15 and 19, a deliberately provocative ultimatum was drafted in words that supposedly excluded the possibility of acceptance by Serbia. The ultimatum was handed to the Serbian government on July 23. The Serbian answer, handed in on time on July 25, was declared as being insufficient, though Serbia had agreed to practically all Austro-Hungarian demands with the exception of two that, in effect, entailed constitutional changes in the Serbian government. These were that certain unnamed Serbian officials be dismissed at the whim of Austria-Hungary and that Austro-Hungarian officials participate, on Serbian soil, in the suppression of organizations hostile to Austria-Hungary and in the judicial proceedings against their members. In its reply, the Serbian government pointed out that such demands were unprecedented in relations between sovereign states but nevertheless agreed to submit the matter to the Permanent Court of Arbitration or to the arbitration of the great powers. On receiving this reply, the Austro-Hungarian ambassador immediately left Belgrade, severing diplomatic relations between the two countries. Berchtold and his government were clearly determined to make war on Serbia, regardless of the fact that such action might result in war between the great powers. While the European governments frantically tried to offer compromise solutions, Austria decided on a *fait accompli.* On July 28, 1914, Berchtold asked Emperor Francis Joseph to sign the declaration of war, informing him that

> it cannot be excluded that the [Triple] Entente powers [Russia, France, Great Britain] might make another move to bring about a peaceful settlement of the conflict unless a declaration of war establishes a fait accompli [*eine klare Situation geschaffen*].

In the meantime the German government had taken control of the situation and, placing German strategic and national plans over Austro-Hungarian interests, had changed the Balkan conflict into a Continental war.

World War I. *Subordination to Germany.* The German declaration of war against France and Russia subordinated the Austro-Serbian conflict to the German aim of settling its own rivalries with France and Russia. According to the terms of the military agreement between Germany and Austria-Hungary, the Austro-Hungarian army had to abandon plans to conquer Serbia and instead protect the German invasion of France against Russian intervention. The setbacks that the Austrian army suffered in 1914 and 1915 can be attributed, to a large extent, to the fact that Austria-Hungary became a military satellite of Germany from the very first day of the war, though it cannot be denied that the Austrian high command proved to be quite incompetent. Conrad had clamoured for preventive war since 1906, but, when he received his chance, in July 1914, it turned out that the Austrian army had no plans for an expeditious offensive. Similarly, after Italy entered the war on the side of the Allies in May 1915, Conrad was unprepared. The fact that only after the Germans had taken command could the Russian front be stabilized and that Serbia and Romania were not defeated until 1915 and 1916, respectively, did little to enhance the prestige of the Austrian government.

Internal disintegration. In July 1914, parliament had been out of session, and Stürgkh had refused to convene it. This and the military censorship that was established immediately after the outbreak of the war concealed the discontent of the non-German population. While German public opinion in Austria had welcomed the war enthusiastically, and while some Polish leaders supported the war out of their anti-Russian resentment, the Czech population openly showed its animosity. The Czech leader Tomáš Garrigue Masaryk, who had acted as one of the most prominent spokesmen for the Czech cause, emigrated to western Europe. Karel Kramář, who had supported the Pan-Slav idea, was put on trial for high treason and declared guilty on the basis of shaky evidence. German nationalism was riding high, but in reality the German Austrians had little influence left. In military matters, they were practically reduced to executing German orders; in economic affairs, the Hungarians, who controlled the food supply, had the decisive influence. Count Tisza, who had opposed the war in July 1914, became the strong man of the empire. On his advice, Berchtold was dismissed in January 1915, and the foreign office was once again entrusted to a Magyar. But Count István Burián, who was

in charge of foreign affairs until December 1916, failed to keep Italy and Romania out of the war. German attempts to pacify the two states by concessions were unsuccessful, because Francis Joseph was unwilling to cede any territory in response to the irredentist demands of the two nations. How little the outward calm in the Habsburg lands corresponded to the sentiment of the population became apparent when Stürgkh was assassinated by Friedrich Adler, the pacifist son of the leader of Austrian Socialism, in October 1916. Francis Joseph made Ernst Koerber prime minister, but Koerber had no chance to develop a program of his own. On November 21, 1916, Francis Joseph died, at the age of 86, leaving the throne and the shaky empire to his 29-year-old grandnephew, Charles, who had had little preparation for his task until the death of Francis Ferdinand had made him the heir apparent. Full of the best intentions, Charles set out to save the monarchy by searching for peace in foreign affairs and by recognizing the rights of the non-German and non-Magyar nationalities of his empire. Charles relied heavily on the advice of politicians who had had the confidence of Francis Ferdinand. He dismissed Koerber in December 1916 and made Count Heinrich Clam-Martinic, a Czech aristocrat, prime minister. At the foreign office, he replaced Burián with Ottokar, Graf Czernin.

Assassination of Stürgkh

When parliament was reconvened in May 1917, it became manifest how far internal disintegration of the Habsburg monarchy had progressed. Parliament once again became the stage of unrelenting national conflicts. Finding so little support from the Czech side, Charles turned back to the German element, and in June 1917 made Ernst von Seidler, once his tutor in administrative and international law, prime minister. But, even though he tried to appease the Czechs, the stubborn insistence of the Germans not to yield any of their prerogatives made reform of the empire impossible.

Early peace feelers

At the same time, various moves to get Austria-Hungary out of the war ended in failure. After a U.S. offer of general mediation had miscarried in December 1916, Charles tried through secret channels to deal directly with the Entente powers. In the spring of 1917, an exchange of peace feelers took place through the mediation of his brother-in-law, Sixtus of Bourbon-Parma, but Italy's unwillingness to abandon some of the concessions granted to it in the Treaty of London (1915) made these talks abortive. Similarly, negotiations with Allied representatives carried on in Switzerland brought no results.

Since the Austro-Hungarian government was unable to extricate itself from the ties of the Dual Alliance, France and England ceased to have regard for the integrity of the Habsburg monarchy. Furthermore, the revolutionary events in Russia in 1917 and the entry of the United States into the war introduced a new, ideological element into Allied policy toward the Central Powers. The German-directed governments represented an authoritarian system of government, and national agitation in the Habsburg monarchy henceforth assumed the character of a democratic liberation movement, winning the sympathies of west European and American public opinion. From early 1918 on, the Allied governments began officially to promote the activities of the emigrés from Austria, foremost among them Tomáš Masaryk, the Czech leader, and in April 1918 a Congress of Oppressed Nationalities was organized in Rome. But the collapse of the Habsburg monarchy cannot be ascribed to the new Allied policy of supporting the independence claims of the Habsburg nationalities, which was only a belated adjustment to the changed conditions within Austria-Hungary. From the summer of 1917, the activities of the nationalist movements within the empire made the situation increasingly untenable, and, two days before U.S. Pres. Woodrow Wilson proclaimed his Fourteen Points, one of which demanded the reorganization of the Habsburg monarchy in accordance with the principles of national autonomy, the Czechs demanded outright independence (January 6, 1918). Within a month, Polish and South Slav deputies, together with the Czechs, presented to the Reichsrat a program demanding the establishment of independent constituent assemblies for nationally homogeneous areas.

The end of the Habsburg Empire. During the same period in which the national-independence movement reached its final stage, another dangerous development manifested itself. From 1915 on, the supply situation had worsened increasingly, and by January 1918 there were dangerous shortages, especially of food. Prompted by the difficult food situation and inspired by the Bolshevik victory in Russia, a strike movement developed. Demands for more bread and a demand for peace were combined with nationalist claims into open opposition to the government. The strikes among the civilian population were followed by mutinies in the army and navy. In January and February 1918, however, the army and the government succeeded in suppressing social unrest and antiwar demonstrations. But, from the same date, the national opposition movement gathered momentum.

Strikes, mutinies, and demonstrations

The hopes that the government had placed in peace settlements with the eastern states were not fulfilled, either. The peace treaty with the Ukraine (signed in February 1918), the Treaty of Brest-Litovsk with Soviet Russia (March 3, 1918), and the Treaty of Bucharest, which settled the peace with Romania (May 7, 1918), did not alleviate the supply situation and irritated the Poles because of certain provisions of the Ukrainian settlement.

In April 1918, Czernin was replaced as foreign minister by Burián. This change was the result of a conflict between Czernin and Charles over the desirability and possibility of Austria's concluding a separate peace with the Allies. Unknown to Czernin, Charles had in 1917 made certain secret overtures to the Allies, which were revealed by French premier Georges Clemenceau. The Germans were outraged, and Czernin was dismissed on their orders. Burián returned to the foreign office on April 16 and immediately reported to the German high command at Spa, where Emperor Charles and Burián had to assure the German Emperor of their unchanging loyalty. While this act of submission satisfied the German Austrians, it further incensed the Slav opposition. In May 1918 a Slav national celebration in Prague demonstrated the strength of the independence movements. But the Emperor and the German elements in the central government were still not aware of the extent of the disintegration. In July 1918, Prime Minister Seidler resigned, and his successor, Max Hussarek von Heinlein, began a belated effort to reorganize the Habsburg monarchy. Hussarek's efforts to federalize the empire in the moment of imminent military defeat unintentionally turned out to provide the basis for the formal liquidation of the Habsburg monarchy. On October 16, 1918, Emperor Charles issued a manifesto announcing the transformation of Austria into a federal union of four components (German, Czech, South Slav, and Ukrainian). The Poles were to be free to join a Polish state, and Trieste was to be given a special status. The lands of the Hungarian crown were to be excepted from this program. Within a few days, national councils were established in all the provinces of the empire and for all practical purposes acted as national governments. The Poles proclaimed the union of all Poles in a unified state and declared their independence at Warsaw on October 7, 1918. The South Slavs advocated union with Serbia, and on October 28, 1918, the Czechs proclaimed the establishment of an independent republic. The dissolution of the Habsburg monarchy was thus consummated by the end of October 1918—that is, before the war actually ended.

Belated attempts at reform

It was impossible for the country to survive another winter of hostilities, and on September 14, 1918, Burián published an appeal to all belligerents to discuss the possibilities of ending the war. When this move was opposed by the Germans as well as by the Allied powers, Burián tried for a separate peace settlement for Austria-Hungary. On October 14, 1918, he sent a note to President Wilson asking for an armistice on the basis of the Fourteen Points. On October 18, 1918, U.S. Secretary of State Robert Lansing replied that, in view of the political development of the preceding months and, especially, in view of the fact that Czechoslovakia had been recognized as being at war with the Central Powers, the U.S. government was unable to deal on the basis of the Fourteen Points anymore. On October 27, 1918, Count Gyula Andrássy, who had re-

placed Burián three days before as foreign minister, sent a new note to President Wilson; in asking for an armistice, he declared full adherence to the statements set forth in the U.S. note of October 18, thus explicitly recognizing the existence of an independent Czechoslovak state. From this moment on, it remained only to liquidate the war. On October 22, 1918, Heinrich Lammasch, a renowned authority in the field of international law and a respected pacifist, formed a new Cabinet. He hoped to save the Habsburg monarchy by drawing up a federative structure. But instead of saving the state, he found himself charged with the task of supervising the dissolution of the empire and bringing about an orderly transfer of power. The government could not influence events outside of Vienna anymore and from October 30, 1918, was even challenged in the central agencies by the German-Austrian state council. The hostilities were ended by an armistice signed on November 3, 1918. The Austro-Hungarian high command, which had blundered into the war unprepared in 1914, did little better at its conclusion. Due to inaccuracies in the wording of the documents, more than 300,000 soldiers were taken prisoner by the Italian army. For some days the government hoped that, in spite of the secession of the Slav areas, the Habsburg dynasty could survive in the remaining lands. But even the German Austrians had lost faith in the Habsburgs, and, with revolutionary agitation on the rise and republican passion widespread, Charles adhered to the advice of Lammasch and decided to waive his rights to exercise political authority. On November 11, 1918, he issued a proclamation acknowledging "in advance the decision to be taken by German Austria" and stating that he relinquished all part in the administration of the state. The declaration of November 11, 1918, marks the formal act of dissolution of the Habsburg monarchy.

THE FIRST REPUBLIC AND THE ANSCHLUSS

The war's aftermath. On October 21, 1918, the 210 German members of the imperial parliament (Reichsrat) of Austria formed themselves into a national assembly for Deutschösterreich, or German-Austria; and on October 30 they proclaimed this an independent state under the direction of a State Council (Staatsrat) composed of the leaders of the three main parties and other elected members. Revolutionary disturbances in Vienna and, more important, the news of the German revolution forced the State Council on the republican path. On November 12, the day after the emperor Charles's abdication, the National Assembly resolved unanimously that "German-Austria is a democratic republic" and "German-Austria is a component part of the German republic." Karl Renner, a leading Socialist, became head of a coalition government, with Otto Bauer, the acknowledged spokesman of the left wing of the Social Democrats, as foreign secretary. On November 22, the territory of the republic was further defined: the National Assembly claimed for the new state all the Habsburg lands in which a majority of the population was German. It also claimed the German areas of Bohemia and Moravia.

Dangers of a Communist revolution

From the first day, the republic was faced with the disastrous heritage of the war. Four years of war effort and the breakup of the Habsburg Empire had brought economic exhaustion and chaos. The resulting social distress and poverty inspired revolutionary activities, thus making Bolshevism appear the greatest danger to the new republic, especially after a Soviet republic was established in Hungary at the end of March 1919. The Austrian Social Democrats were determined to resist Bolshevism with their own forces without making an alliance (as the German Social Democrats did) with the old order. A Volkswehr (People's Guard) was organized and was twice effective (April 17 and June 15) against Communist attempts at a *Putsch.* Otto Bauer and Friedrich Adler staked their popularity on defeating the Communist agitation in the workers' and soldiers' councils, which had been set up on the Soviet model. By mid-1919, political and social order was restored on parliamentary lines, the Communist Party relapsing into insignificance. More dangerous was the tendency of the *Länder* (provinces, or states) to break away from Vienna or to claim almost complete independence. Though the principal motive of this was reluctance to send food supplies to Vienna, it also represented a genuine social, political, and ideological conflict: the administration of the industrialized capital was Socialist controlled, while the *Länder,* being predominantly agrarian, remained conservative and faithful to the Catholic tradition. This difference was aggravated by the fact that the monarchy had been the only bond between the German Austrian lands; with the abdication of the Emperor, no symbol of loyalty common to all *Länder* remained. Vorarlberg voted for union with Switzerland in May 1919, and Tirol also attempted to secede.

The constitutional settlement. In February 1919, elections for a constitutional assembly were held. The Social Democrats were returned as the largest single party, with 69 seats. Sixty-three were won by the Christian Socialists and 26 by the German Nationalists. When this assembly met (March 4), it had to make wide concessions to federalism in order to appease the provinces. In exchange, Vienna was also elevated to the rank of a state and the mayor made the equivalent of a state governor. This proviso subsequently enabled "Red Vienna" to pursue an autonomous policy, despite the fact that the *Bundesregierung* ("federal government") was controlled by the conservative parties from 1920 to 1934.

Autonomy of Vienna

The constituent assembly also settled the constitution of the federal republic (October 1, 1920). The State Council was abolished, and a bicameral legislative assembly, the Bundesversammlung, was established. The Bundesrat (upper house) was to exercise only a suspensive veto and was to be elected roughly in proportion to the population in each state. This represented a defeat for the federal elements in the states, which had wanted the Bundesrat to exercise an absolute veto and to be composed of equal numbers of members from each state. The lower house, or Nationalrat, was to be elected by universal suffrage on a basis of proportional representation. The Bundesversammlung in full session elected the president of the republic for a four-year term, but the federal government, with the chancellor at its head, was elected in the Nationalrat on a motion submitted by its principal committee; this committee was itself representative of the proportions of the parties in the house.

The foreign policy of Otto Bauer and representatives of the major political parties had insisted firmly on *Anschluss* ("union") with Germany, and as late as 1921 unauthorized plebiscites held in the western provinces returned overwhelming majorities in favour of union with Germany. But article 88 of the peace treaty of Saint-Germain, signed on September 10, 1919, forbade *Anschluss* without the consent of the League of Nations and also stipulated that the republic should cease to call itself Deutschösterreich (German-Austria); it became the Republik Österreich. The Austrian claim for the German-speaking areas of Bohemia and Moravia was denied by the peace conference, and Austria had to recognize the frontiers of Czechoslovakia along slightly rectified historical administrative lines. The southern frontier with Yugoslavia was threatened by Yugoslav armed invasion, and it was finally decided that the question should be settled by a plebiscite, which, on October 10, 1920, returned a majority of 59 percent in favour of Austria. The German-speaking districts of western Hungary were to be ceded to Austria outright; but Austria, in the face of Hungarian resistance, was obliged to hold a plebiscite. The area of Sopron was finally restored to Hungary.

After the elections of February 1919, Renner formed another coalition government, but after a government crisis in the summer of 1920 a caretaker cabinet under the Christian Socialist Michael Mayr was formed. This government prepared the draft of the constitution and introduced it into parliament. After its approval, new elections were held, on October 17, 1920. The Christian Socialists were returned as the strongest party, gaining 82 seats, while the Social Democrats were reduced to 66 and the German Nationalists to 20. Mayr formed a new cabinet composed of Christian Socialists; the Social Democrats went into opposition and never returned to the government throughout the First Republic. This political division hardened, and no decisive change took place during the following years.

Christian Socialist electoral victory (1920)

The system of proportional representation combined with the ideological background of Austrian parties made oscillations of political allegiance unlikely. Of the two mass parties, the Social Democrats had an unshakable majority in Vienna (in which about a third of all the inhabitants of the republic lived), while the Christian Socialists had an equally secure majority among the Catholic peasants and the conservative classes, the latter consisting largely of army officers, landowners, and big business. The urban middle classes, hostile to both workers and peasants, became German nationalists. But German nationalism was not limited to the middle classes. The workers and even the peasants felt themselves to be Germans and also responded to the national appeal.

Economic reconstruction. The main task of the non-Socialist governments in power in Austria from the autumn of 1920 on was to restore financial and economic stability. Between 1919 and 1921, the urban population of Austria lived largely on relief from the United States and Great Britain, and, although production improved, distress was heightened by inflation that threatened financial collapse in 1922. In October 1922, the chancellor, Ignaz Seipel, secured a large loan through the instrumentality of the League of Nations, enabling Austrian finances to be stabilized. In return, Austria had to undertake to remain independent for at least 20 years. The controller general appointed by the League of Nations reported in December 1925 that the the Austrian budget had been balanced satisfactorily, and in March 1926 international financial supervision was withdrawn.

Seipel's success in October 1922 gave Austria some years of stability and made economic reconstruction and relative prosperity possible. In Socialist-controlled Vienna, an ambitious program of working class housing, health schemes, and adult education was carried out under the leadership of Karl Seitz, Hugo Breitner, and Julius Tandler. "Red Vienna" thus acquired a unique reputation in Europe.

Political strife. In 1920 all three major parties spoke in democratic terms. Despite democratic phrases, however, preparations for possible civil war had never been abandoned. The Christian Socialists, led by Seipel, a believer in strong government, were convinced that they had to protect the existing social order against a Marxist revolution. In the provinces, reactionary forces (the Heimwehr, or "home defense forces"), originally formed for defense against the Yugoslavs or merely against international disorder, gradually acquired Fascist tendencies. The Social Democrats felt that their social-reform program was endangered by reaction. They possessed their own armed force, the Schutzbund (Defense League), descended from the People's Guard of 1918, and they and the reactionary forces regularly demonstrated against each other. In 1927, in the course of a clash between members of the Schutzbund and certain reactionary forces at Schattendorf, an old man and a child were accidentally shot by the reactionaries. When the latter were acquitted by a Vienna jury on July 14, the Social Democrats called for a mass demonstration, which got out of hand and ended in the burning down of the ministry of justice. In street fighting between the police and the demonstrators, almost 100 persons were killed. The Social Democrats then launched a general strike, but this had to be called off after four days. Seipel had used the opportunity for a violent assertion of government authority. The balance between Socialist and non-Socialist forces in Austria was never secure after this decisive date.

Emergence of Fascist groups

The Christian Socialists, pressed increasingly by the Heimwehr, now began to take the offensive against the Social Democrats. Wilhelm Miklas, a leading Christian Socialist, was elected president as successor to the non-party Michael Hainisch, who had been in office since December 1920. There were repeated attempts to revise the constitution, principally with the object of strengthening the power of the executive. After protracted negotiations, a compromise was reached late in 1929. On December 7, 1929, a series of constitutional amendments gave increased powers to the president. Of particular importance were the rights to appoint ministers and issue emergency decrees. But Vienna preserved its autonomy, and the democratic principle was preserved against the far-reaching authoritarian demands of the Heimwehr. In the elections of November 1930, the Social Democrats were returned as the largest single party, with 72 seats. The Christian Socialists held 66, the German Nationalists 19, and the Heimwehr, now posing as a Fascist party on the Italian model, 8.

These political events were overshadowed by the great world economic crisis. Though the Social Democratic leaders believed that the crisis should be met by the orthodox means of deflation and spending cuts, they were resolved not to be compromised by supporting these measures and refused to enter a coalition government. On the other hand, in October 1931 they acquiesced in suspending the election of the president by direct popular vote, as had been provided by the constitution of 1929, and agreed to the reelection of Miklas by parliament for a further four years. The government, meanwhile, led by Otto Ender and Johann Schober, was driven to desperate devices in order to stave off collapse. Schober, leader of the middle-class German Nationalists, launched the project for a customs union with Germany in March 1931; this provoked violent opposition from France and the alliance of the Little Entente (Czechoslovakia, Yugoslavia, and Romania) and was subsequently condemned by a majority of the International Court at The Hague. The bankruptcy in May 1931 of the Creditanstalt, the most influential banking house in Austria, brought Austria close to financial and economic disaster. This, together with the rise of the National Socialists in Germany, resulted in considerable support being given to the Nazis in Austria; and the provincial elections in 1932 showed that they were draining off votes from the conservative parties. The Nationalists began to demand a general election, and this demand was taken up by the Social Democrats, who saw a chance of winning a majority in parliament.

Proposed customs union with Germany (1931)

Authoritarianism: Dollfuss and Schuschnigg. After the election, when Engelbert Dollfuss came to form a Christian Socialist government on May 20, 1932, he could count on a majority of only one vote. Dollfuss belonged to a new generation that had been educated in the conservative conviction that the Western form of parliamentary government had been forced upon the central Europeans as a result of military defeat and Socialist revolution and that the political and social order could be restored only by the establishment of some kind of strong authority. The leaders of the Christian Socialist Party found themselves under attack from two ideological enemies, the Marxists and the Nazis, who apparently threatened the very basis of the conservative order. In reaction, Dollfuss determined to replace parliamentary government with an authoritarian system. The opportunity to do this came in March 1933, when, during a debate on a minor bill, an argument arose over alleged irregularities in the voting procedure. The president of the Nationalrat resigned; the two vice presidents followed his example; and Dollfuss declared that parliament had proved unworkable. It never met again in full, and Dollfuss governed thereafter by emergency decree.

The end of parliamentary government

By this time (spring of 1933), Adolf Hitler was in power in Germany, and Nazi propaganda for the incorporation of Austria was greatly increased. Dollfuss turned to Italy for help, convinced that British and French aid would be ineffective. This shift in foreign policy can also be attributed to the fact that Dollfuss had to rely on the help of the anti-Marxist Heimwehr to stay in power.

The Social Democrats were subjected to increasing provocation and on February 12, 1934, took to arms. Civil war followed. After four days of fighting, Dollfuss and the Heimwehr were victorious. The Social Democratic Party was declared illegal and driven underground. In the course of the same year, all political parties were abolished except the Vaterländische Front (Fatherland Front), which Dollfuss had founded in 1933 to unite all conservative groups. In April 1934, the rump of the parliament was brought together and accepted an authoritarian constitution. The executive was given complete control over the legislative branch of government; the elected assemblies disappeared and were replaced by advisory bodies, appointed in a complicated and futile fashion. The rights of man guaranteed

under the democratic constitution were also swept away. "Republic" was removed from the official name of the state, which became merely the Federal State of Austria.

Dollfuss murder

On July 25, 1934, a group of Nazis seized the chancellery and attempted to proclaim a Nazi government. Dollfuss, whom they had taken prisoner, was murdered. The plan, however, miscarried: the Nazis in the chancellery were compelled to surrender, and their leaders were executed; a Nazi rising in Styria was suppressed; and Hitler, faced with the mobilization of an Italian army on the Brenner Pass, repudiated his Austrian followers. Franz von Papen was sent as German ambassador to reduce Austria by other means. Kurt von Schuschnigg, who became chancellor on the death of Dollfuss, was a man of gentler personality and of less violent political passions. His administration of the authoritarian constitution was in the easygoing Austrian fashion, less oppressive than in Italy and Germany. Schuschnigg had a mild preference for restoring the Habsburgs, but he shrank from the international complications that this would involve. The regime drifted on without popular favour, weakened by the personal rivalries and ambitions of its leaders and sustained only by a guarantee from Italy. The temporary accord of Great Britain, France, and Italy in the "Stresa front" (April 1935) seemed to promise new security, but the Ethiopian crisis soon destroyed the unity of the Western powers, and Austria's isolation was complete when Hitler and Mussolini allied themselves in 1936. Schuschnigg had to negotiate a compromise with Germany, which was signed on July 11, 1936; Germany promised to respect Austrian sovereignty, and in return Austria acknowledged itself "a German state." The agreement left Austria open to Nazi infiltration. In January 1938, the Austrian police discovered a new Nazi conspiracy. Schuschnigg hoped to defeat this by a meeting with Hitler, but at Berchtesgaden, where Hitler received him on February 12, 1938, Schuschnigg was faced with threats of military intervention in support of the Austrian Nazis. He had to agree to give them a general amnesty and to include some leading Nazis in his Cabinet; the Ministry of the Interior had to be entrusted to Arthur Seyss-Inquart, the spokesman of Austrian Nazis. The open agitation of the Nazis threatened to destroy the government's authority, and confidential contacts in the European capitals brought Schuschnigg to realize that he could not count on the support of the great powers. He therefore resolved to challenge Hitler alone. On March 9 he announced that a plebiscite would be held on March 13 to decide in favour of Austrian independence.

Anschluss and World War II. Though the Austrian crisis had taken him unaware, Hitler acted with energy and speed. Mussolini's neutrality was assured, there was a ministerial crisis in France, and the British government had made it known for some time that it would not oppose union of Austria with Germany. On March 11, 1938, two peremptory demands were made for the postponement of the plebiscite and for the resignation of Schuschnigg. Schuschnigg gave way, and German troops, accompanied by Hitler himself, entered Austria on March 12. A Nazi government in Austria, headed by Arthur Seyss-Inquart, was established and collaborated with Hitler in proclaiming the *Anschluss* on March 13. France and Great Britain protested against the methods used by Hitler but accepted the *fait accompli,* as did all other governments. A plebiscite on April 10, held throughout greater Germany, recorded a vote of more than 99 percent in favour of Hitler.

The Anschluss

Austria was absorbed into Germany. Immediately after the invasion, the Nazis arrested the leaders of the Austrian political parties, and many Austrians, especially those of Jewish origin, went into exile. But the political antagonism that had previously weakened the status of the republic continued to block cooperation among the émigrés as well as among the resistance groups that formed inside Austria.

The possibility of reestablishing an independent Austria was far from dead, however, and, after the outbreak of World War II the Allied governments began to reconsider their attitude toward the *Anschluss.* In December 1941, Stalin informed the British that the Soviet Union would regard the restoration of an independent Austrian republic as an essential part of the postwar order in central Europe, and, at the meeting of the foreign ministers of Great Britain, the Soviet Union, and the United States in Moscow (October 1943), a declaration was published that declared the union with Germany null and void and pledged the Allies to restore Austrian independence. Though the British prime minister, Winston Churchill, continued to make proposals for setting up a central European federation comprising the former Habsburg lands and even southern Germany, the European Advisory Commission in London assumed that Austria would return to sovereignty within the borders of 1937. When Soviet troops liberated Vienna on April 13, 1945, representatives from the resistance movement and the former political parties were allowed to organize.

THE SECOND REPUBLIC

The Allied occupation. On April 27, 1945, Karl Renner set up a provisional government composed of Social Democrats, Christian Socialists, and Communists and proclaimed the reestablishment of Austria as a democratic republic. The Western powers, afraid that the Renner government might be an instrument of Communist expansion, withheld full recognition until the autumn of 1945. Because of similar suspicions, agreement on the division of Austrian zones of occupation was delayed until July 1945. Shortly before the Potsdam Conference (which stipulated that Austria would not have to pay reparations but assigned the German foreign assets of eastern Austria to the Soviet Union), control machinery was set up for the administration of Austria, giving supreme political and administrative powers to the military commanders of the four occupying armies. In September 1945 a conference of representatives of all provinces extended the authority of the Renner government to all parts of Austria. A general election held in November 1945, in which former Nazis were excluded from voting, returned 85 members of the Austrian People's Party (corresponding to the Christian Socialists of the prewar period), 76 Socialists (corresponding to the Social Democrats and Revolutionary Socialists), and four Communists. Renner was elected president of the republic; Leopold Figl, leader of the Austrian People's Party, became chancellor of a coalition Cabinet. The government decided not to draft a new constitution but to return to the constitution of 1920, as amended by the laws of 1929. In June 1946 the control agreement of July 1945 regulating the machinery of Allied political supervision was modified by restricting Allied interference essentially to constitutional matters. Denazification laws passed in 1946 and 1947 eliminated Nazi influence from the public life of Austria. In the postwar years, the Austrian People's Party and the Socialists were the sole partners in a coalition government that was formed in proportion to the parties' strength in parliament. This principle of proportional representation, originally introduced in 1919, was to be an important factor in Austrian political life after 1945.

General elections of November 1945

From 1945 to 1952, Austria had to struggle for survival. After liberation from Nazi rule, the country faced complete economic chaos. Aid provided by the United Nations Relief and Rehabilitation Administration (UNRRA) and, from 1948, support given by the United States under the Marshall Plan made survival possible. Heavy industry and banking were nationalized in 1946, and, by a series of wage–price agreements, the government tried to control inflation. Interference by military commanders in political and economic affairs in the Soviet zone of occupation caused a considerable migration of capital and industry from Vienna and Lower Austria to the formerly purely agricultural western provinces. This brought about a far-reaching transformation of the economic and social structure of the country. Austria continued to be occupied by U.S., British, French, and Soviet forces until 1955. A treaty restoring Austrian sovereignty was expected early, but the atmosphere of the Cold War made agreement among the former Allied powers impossible. In 1953, however, a heavy burden was removed from the Austrian economy when the Soviet government declared that it would pay its own occupation costs (as the United States had done since 1947). Thereupon, the British and the French followed suit.

In 1949 former Nazis were allowed to participate in the general election. The Union of Independents (later renamed the Freedom Party), corresponding to the former German Nationalist group but free from ideological ties, won 16 seats in parliament. In subsequent elections (1953, 1956, 1959, 1962), the relationship of the three parties remained stable. When Renner died (December 31, 1950) Theodor Körner, the Socialist mayor of Vienna, was elected president of the republic by direct popular vote. He was succeeded in 1957 by the leader of the Socialist Party, Adolf Schärf, followed in 1965 by Franz Jonas, former mayor of Vienna, and in 1974 by Rudolf Kirchschläger, former minister of foreign affairs.

The influence of the Socialists in the coalition government, which had been relatively strong under Leopold Figl's chancellorship, was reduced when the Austrian People's Party replaced Figl with Julius Raab in the spring of 1953 and had Reinhard Kamitz appointed minister of finance. The subsequent economic reconstruction and the advance to a prosperity unknown to Austrians since the years before World War I is generally identified with the so-called Raab–Kamitz course, based on a modified free-market economy. The nationalized steel industry, electrical-power plants, and oil fields, together with the privately owned lumber and textile industries and the tourist traffic, were the major economic assets. The Austrian economy came to be dominated to a disproportionate extent by a trend toward the tertiary sector because of the importance of *Fremdenverkehr* (tourist traffic), which transformed the economic and social character of the rural Alpine areas.

The Raab–Kamitz course

The Berlin conference of the foreign ministers of France, Great Britain, the Soviet Union, and the United States in January 1954 raised Austrian hopes for the conclusion of a peace treaty. For the first time, Austria was admitted as an equal conference partner, but the failure of the foreign ministers to agree on the future of Germany again prejudiced Austria's chances. It appeared that the Soviet government was not prepared to forgo the strategic advantages of maintaining its forces in Austria so long as Germany was not "neutralized." In February 1955 the Soviet government suddenly extended an invitation to the Austrian government for bilateral negotiations. An Austrian delegation visited Moscow in April 1955, and a memorandum was agreed on, according to which the Soviet government declared itself ready to restore full Austrian sovereignty and to evacuate its occupation troops in return for an Austrian promise to declare the country permanently neutral.

Restoration of sovereignty. The treaty was signed in Vienna on May 15, 1955, by the representatives of the four powers and Austria. It formally reestablished the Austrian republic in its pre-1938 frontiers as a "sovereign, independent and democratic state." It prohibited *Anschluss* between Austria and Germany as well as the restoration of the Habsburgs. It guaranteed the rights of the Slovene and Croatian minorities in Carinthia, Styria, and Burgenland. The United Kingdom, the United States, and France relinquished to Austria all property, rights, and interests held or claimed by them as former German assets or war booty. The Soviet Union, however, obtained tangible payment for the restoration of Austrian freedom. This included $150,000,000 for the confiscated former German enterprises, which Austria had to buy back from the Administration of Soviet Property in Austria; $2,000,000 for the confiscated German assets of the First Danube Steamshipping Company; and 10,000,000 metric tons of crude oil as the price of Austrian oil fields and refineries that had been Soviet war booty. The treaty came into force on July 27, 1955, and by October 25 all occupation forces were withdrawn. On October 26 a constitutional law of perpetual Austrian neutrality was promulgated. The Austrian government had never left any doubts that the pledge to neutrality could be interpreted only as a military one and never as an ideological one. Throughout the Soviet occupation, the Austrians had proved their anti-Communist attitude, and the spontaneous reaction of the Austrian people during the suppression of the Hungarian Revolution in 1956 demonstrated their sympathy with Western democratic ideas. Austria preserved political stability; changes in the personal and ideological structure of the government and political parties were effected without major political crisis.

Austrian neutrality

From 1962 disagreement over economic problems generated friction between the coalition parties. The annual budget led to grave disunity in the coalition, and in the autumn of 1965 the government resigned and called new elections. The election, held on March 6, 1966, brought a setback for the Socialist Party, and the People's Party was returned to parliament with an absolute majority. Negotiations for a new coalition government failed. The Socialists, led by a former foreign minister, Bruno Kreisky, went into opposition, and Josef Klaus formed the first one-party Cabinet of the Second Republic. Contrary to widespread misgivings, the political stability of the country was not disturbed, and parliament was given new vigour and influence. In ensuing provincial elections, the Socialist Party demonstrated recovery from the setback of 1966, and in the national elections of 1970 the Socialist Party managed to win a plurality of votes, becoming the strongest party in parliament, with 81 seats, but falling short of a majority. After negotiations for a new coalition Cabinet failed, in May 1970 Bruno Kreisky was appointed chancellor and formed the first Austrian all-Socialist Cabinet. Sensing increased support for the Socialists he called for new elections in October 1971, which gave his party a clear majority of 93 seats. In the subsequent elections of 1975 and 1979 the Austrian voters demonstrated their approval of Kreisky's policy of moderate social reform and economic stability by returning the Socialist Party to parliament in increasing strength: the elections of May 1979 gave the Socialists 95 seats, while the Austrian People's Party, weakened by problems of leadership, received only 77 seats and the Freedom Party only 11. In the elections of 1983 the Socialist Party lost its majority, gaining only 90 seats; it formed a coalition government with the Freedom Party, which won 12 seats.

The stability of Austrian politics is paralleled by an equally stable economy: besides having an elaborate system of social security and health insurance, the Austrians have enjoyed an unbroken prosperity with one of the lowest rates of unemployment in Europe. The Kreisky government carried through a number of reform programs, among which the reorganization of the legal code under the minister of justice, Christian Broda, had truly historical dimensions.

Austria became a member of the United Nations in 1955 and of the Council of Europe in 1956. Major problems of Austrian foreign relations were the conflict with Italy over Südtirol (Bolzano) and the problem of association with the European Economic Community (EEC). During the Paris Peace Conference of 1946 an agreement was signed guaranteeing the rights of the German-speaking population of Südtirol. The Austrian government, claiming that the Italians had not lived up to their obligations, initiated bilateral talks. In the early 1960s, acts of terrorism committed by German-speaking chauvinists interfered with the progress of the negotiations, but in 1969 agreement was finally reached on implementing the guarantees provided in the agreement of 1946. In 1958 Austria joined the European Free Trade Association, but a special arrangement with the European Economic Community, accompanied by prudent dealings with the Socialist neighbours, maintained Austria's status as a neutral nation. For current political history, see the annual issues of the *Britannica Book of the Year*.

Foreign relations

(F.Fe.)

BIBLIOGRAPHY

General works. FEDERAL PRESS SERVICE IN VIENNA, *Austria: Facts and Figures* (1979, rev. periodically), contains comprehensive statistical data as well as other information on economic and, to a certain extent, cultural developments in Austria since 1945. KARL STADLER, *Austria* (1971), provides a general introduction to the country to the 1970s. C.T. GRAYSON, *Austria's International Position, 1938–53: The Reestablishment of an Independent Austria* (1953); and RICHARD HISCOCKS, *The Rebirth of Austria* (1953), review Austria's immediate postwar development. JACQUES HANNAK (ed.), *Bestandsaufnahme Österreich 1945–1963* (1963); and HEINRICH SIEGLER, *Austria:*

Problems and Achievements Since 1945 (1967), carry developments beyond the state treaty. E. WEINZIERL and K. SKALNIK (eds.), *Österreich. Die Zweite Republik,* 2 vols. (1972); and H. FISCHER (ed.), *Das politische System Österreichs* (1974), review the general history and the political system since the war. E.E. BAUMANN, *Crossroads of European Art: A Concise History of Art and Architecture in Austria* (1964), is an extensive history of Austria's cultural history. For statistical information, see the annual publications *Österreichisches Jahrbuch* and *Wirtschaftsstatistisches Handbuch;* and ÖSTERREICHISCHES STATISTISCHES ZENTRALAMT, *Republik Österreich 1945–1975* (1975; Eng. trans. 1976).

Histories. *General:* KARL and MATHILDE UHLIRZ, *Handbuch der Geschichte Österreichs und seiner Nachbarländer Böhmen und Ungarn,* 4 vol. (1927–44; vol. 1, 2nd ed., 1963), with excellent bibliography; H. HANTSCH, *Die Geschichte Österreichs,* 4th ed., 2 vol. (1959–68), a conservative, scholarly, pro-Habsburg account; E. ZOLLNER, *Geschichte Österreichs von den Anfängen bis zur Gegenwart,* 4th ed. (1970), a standard Austrian text covering ancient times to the present; O. SCHULMEISTER (ed.), *Spectrum Austriae* (1957), a comprehensive, positive evaluation; R. RICKETT, *A Brief Survey of Austrian History* (1966); *Austrian History Yearbook,* published by Rice University, devoted exclusively to Austrian history.

Prehistory: R. PITTIONI, *Urgeschichte des österreichischen Raumes* (1954).

Roman period: A. BETZ, *Aus Österreichs römischer Vergangenheit* (1956).

Middle Ages to 1246: A.W.A. LEEPER, *A History of Medieval Austria* (1941); K. LECHNER, *Die Babenberger und Österreich* (1947); H. FICHTENAU, *Von der Mark zum Herzogtum: Grundlagen und Sinn des "Privilegium minus" für Österreich,* 2nd ed. (1965).

1246–1526: A. LHOTSKY, *Geschichte Österreichs seit der Mitte des 13. Jahrhunderts (1281–1358)* (1967).

1526–1648: J. LOSERTH, *Die Reformation und Gegenreformation in den innerösterreichischen Ländern* (1898); G. MECENSEFFY, *Geschichte des Protestantismus in Österreich* (1956); V. BIBL, *Maximilian II., der rätselhafte Kaiser* (1929); H. STURMBERGER, *Georg Erasmus Tschernembl* (1953); *Kaiser Ferdinand II. und das Problem des Absolutismus* (1957).

1648–1740: O. REDLICH, *Weltmacht des Barock,* 4th ed. (1961); *Das Werden einer Grossmacht, Österreich von 1700–1740,* 4th ed. (1962); T.M. BARKER, *Double Eagle and Crescent: Vienna's Second Turkish Siege and Its Historical Setting* (1967); A. CORETH, *Österreichische Geschichtschreibung in der Barockzeit, 1620–1740* (1952); B. GRIMSCHITZ, R. FEUCHTMULLER, and W. MRAZEK, *Barock in Österreich* (1960); R.A. KANN, *A Study in Austrian Intellectual History: From Late Baroque to Romanticism* (1960).

1740–92: F. MAASS (ed.), *Der Josephinismus, Quellen zu seiner Geschichte in Österreich 1760–1790,* 5 vol. (1951–57); R.A. KANN (*op. cit.*).

1792–1848: A.H. SPRINGER, *Geschichte Österreichs seit dem Wiener Frieden 1809–1849,* 2 vol. (1863–65); V. BIBL, *Der Zerfall Österreichs,* 2 vol. (1922–24), see vol. 1, *Kaiser Franz und sein Erbe;* C.A. MACARTNEY, *The Habsburg Empire, 1790–1918* (1968).

1848–67: HEINRICH FRIEDJUNG, *Österreich von 1848–1860,* 2 vol. (1908–11); *Der Kampf um die Vorherrschaft in Deutschland, 1856 bis 1866,* 10th ed. (1916–17; abridged Eng. trans., *The Struggle for Supremacy in Germany,* 2 vol., 1935, reprinted 1966); R. KISZLING, *Die Revolution im Kaisertum Österreich,* 2 vol. (1848–49); R.A. KANN, *The Multinational Empire: Nationalism and National Reform in the Habsburg Monarchy, 1848–1918,* 3rd ed., 2 vol. (1964), rev. and enlarged German ed., *Das Nationalitätenproblem der Habsburgermonarchie,* 2nd ed., 2 vol. (1964), the standard work on the nationality problem, heavily documented; ADOLF BEER, *Die österreichische Handelspolitik im 19. Jahrhundert* (1891).

1867–1918: H. WICKHAM STEED, *The Hapsburg Monarchy,* 4th ed. (1919; 1914 ed. reprinted 1969), a critical contemporary view; A.J.P. TAYLOR, *The Habsburg Monarchy, 1809–1918,* new ed. (1948), prematurely dated but still provocative; FRITZ FELLNER, *Der Dreibund,* 2nd ed. (1963), a reappraisal of Austro-Hungarian Alliance systems; A.J. MAY, *The Hapsburg Monarchy, 1867–1914* (1951), a comprehensive and balanced study; *The Passing of the Hapsburg Monarchy, 1914–1918,* 2 vol. (1966); C.A. MACARTNEY (*op. cit.*), a scholarly work particularly strong on Hungary (excellent bibliography); B. JELAVICH, *The Habsburg Empire in European Affairs, 1814–1918* (1969), concentrates on foreign relations; The first three volumes of *Die Habsburgermonarchie 1848–1918,* ed. by ADAM WANDRUSZKA and PETER URBANITSCH and planned to be an eight-volume standard handbook, appeared 1973–80.

1918 to the present: O. BAUER, *Die Österreichische Revolution* (1923; Eng. trans., *The Austrian Revolution,* 1925), Socialist interpretation from one of the leading protagonists; C.A. MACARTNEY, *The Social Revolution in Austria* (1927), the only social history of Austria available; M. MACDONALD, *The Republic of Austria, 1918–1934* (1946), a concise and objective account; C.A. GULICK, *Austria from Habsburg to Hitler,* 2 vol. (1948), the most comprehensive work on the interwar period (pro-Socialist); KARL R. STADLER, *Austria* (1971), a scholarly history of Austria in the 20th century, especially strong on the period since 1938; E. WEINZIERL and K. SKALNIK (eds.), *Österreich: Die Zweite Republik, 1945–1970,* 2 vol. (1972), a comprehensive account; GERALD STOURZH, *Kleine Geschichte des Österreichischen Staatsvertrages* (1975), a well-documented account of Austria's position between East and West.

Automata Theory

In simple terms an automaton represents a formalization of a set of rules for a computation, and automata theory, which is studied as part of the foundations of mathematics, is used in the building of such machines as all-purpose computers. An example of a typical automaton is a pendulum clock. In such a mechanism the gears can assume only one of a finite number of positions, or states, with each swing of the pendulum. Each state, through the operation of the escapement, determines the next succeeding state, as well as a discrete output, which is displayed as the discrete positions of the hands of the clock. As long as such a clock is wound and its operation is not interfered with, it will continue to operate unaffected by outside influences except the effect of gravity on the pendulum.

More general automata are designed to respond to changes in external conditions or to other inputs. For example, thermostats, automatic pilots of aircraft, missile guidance systems, telephone networks, and controls of certain kinds of automatic elevators are all forms of automata.

The internal states of such devices are not determined solely by their initial state, as is the case of the pendulum clock, but may be determined by an input from a human operator, from another automaton, or by an event or series of events in the environment. A thermostat, for instance, has an "on" or "off" state that depends on the temperature. The best known general automaton is the modern electronic computer, the internal states of which are determined by the data input and which operates to produce a certain output. (B.R./R.J.Ne)

This article is divided into the following sections:

NATURE AND ORIGIN OF MODERN AUTOMATA

The components of automata consist of specific materials and devices, such as wires, transistors, levers, relays, gears, and so forth, and their operation is based on the mechanics and electronics of these parts. The principles of their operation as a sequence of discrete states can, however, be understood independently of the nature or arrangement of their components. In this way, an automaton may be considered, abstractly, as a set of physically unspecified states, inputs, outputs, and rules of operation, and the study of automata as the investigation of what can be accomplished with these. This mode of abstraction yields mathematical systems that in certain respects resemble logical systems. Thus, an automaton can be described as a logically defined entity that can be embodied in the form of a machine, with the term automaton designating both the physical and the logical constructions.

In 1936 an English mathematician, Alan Mathison Turing, in a paper published in the *Proceedings of the London Mathematical Society* ("On Computable Numbers with an Application to the Entscheidungsproblem"), conceived a logical machine the output of which could be used to define a computable number. For the machine, time was considered to be discrete and its internal structure, at a given moment, was described simply as one of a finite set of states. It performed its functions by scanning an unbounded tape divided into squares, each of which either contained specific information in the form of one of a finite number of symbols or was blank. It could scan only one square at a time, and, if in any internal state except one called "passive," it was capable of moving the tape forward or backward one square at a time, erasing a symbol, printing a new symbol if the square was blank, and altering its own internal state. The number it computed was determined by symbols (the "program") provided on a finite portion of the tape and the rules of operation, which included stopping when the passive state was reached. The output number was then interpreted from the symbols remaining on the tape after the machine stopped.

The Turing machine

Automata theory since the middle of the 20th century has been extensively refined and has often found practical application in civilian and military machines. The memory banks of modern computers can store large (though finite) amounts of information. (For further information on computers and their applications, see INFORMATION PROCESSING AND INFORMATION SYSTEMS.) The original Turing machine had no limit to the memory bank because each square on the unbounded tape could hold information. The Turing machine continues to be a standard reference point in basic discussions of automata theory, and many mathematical theorems concerning computability have been proved within the framework of Turing's original proposal.

NEURAL NETS AND AUTOMATA

The finite automata of McCulloch and Pitts. Part of automata theory lying within the area of pure mathematical study is often based on a model of a portion of the nervous system in a living creature and on how that system with its complex of neurons, nerve endings, and synapses (separating gap between neurons) can generate, codify, store, and use information. The "all or none" nature of the threshold of neurons is often referred to in formulating purely logical schemata or in constructing the practical electronic gates of computers. Any physical neuron can be sufficiently excited by an oncoming impulse to fire another impulse into the network of which it forms a part, or else the threshold will not be reached because the stimulus is absent or inadequate. In the latter case, the neuron fails to fire and remains quiescent. When several neurons are connected together, an impulse travelling in a particular part of the network may have several effects. It can inhibit another neuron's ability to release an impulse; it can combine with several other incoming impulses each of which is incapable of exciting a neuron to fire but that, in combination, may provide the threshold stimulus; or the impulse might be confined within a section of the nerve net and travel in a closed loop, in what is called "feedback." Mathematical reasoning about how nerve nets work has been applied to the problem of how feedback in a computing machine can result in an essential ingredient in the calculational process.

Original work on this aspect of automata theory was done by Warren S. McCulloch and Walter Pitts at the Research Laboratory of Electronics at the Massachusetts Institute of Technology starting in the 1940s.

The definitions of various automata as used here are based on the work of two mathematicians, John von Neumann and Stephen Cole Kleene, and the earlier neurophysiological researches of McCulloch and Pitts, which offer a mathematical description of some essential features of a living organism. The neurological model is suggested from studies of the sensory receptor organs, internal neural structure, and effector organs of animals. Certain responses of an animal to stimuli are known by controlled observation, and, since the pioneering work of a Spanish histologist, Santiago Ramón y Cajal, in the latter part of the 19th and early part of the 20th century, many neural structures have been well known. For the purposes of this article, the mathematical description of neural structure, following the neurophysiological description, will be called a "neural net." The net alone and its response to input data are describable in purely mathematical terms.

A neural net may be conveniently described in terms of the kind of geometric configuration that suggests the physical structure of a portion of the brain. The component parts in the geometric form of a neural net are named (after the physically observed structures) neurons. Diagrammatically they could be represented by a circle and a line (together representing the body, or soma, of a physiological neuron) leading to an arrowhead or a solid dot (suggesting an endbulb of a neuron). A neuron may be assumed to have either an excitatory or an inhibitory effect on a succeeding one; and it may possess a threshold, or minimum number of unit messages, so to speak, that must be received from other neurons before it can be activated to fire an impulse. The process of transmission of excitation mimics that which is observed to occur in the nervous system of an animal. Messages of unit excitation are transmitted from one neuron to the next, and excitation is passed along the neural net in quantized form, a neuron either becoming excited or remaining nonexcited, depending on the states (excitatory or quiescent) of neurons whose endbulbs impinge upon it. Specifically, neuron N, with threshold h, will be excited at time t, if and only if h or more neurons whose excitatory endbulbs impinge upon it are excited at time $t-1$ and no neuron whose inhibitory endbulb impinges upon it is excited at time $t-1$. A consistent picture can be made of these conditions only if time and excitation are quantized (or pulsed). It is assumed conventionally that a unit of time is required for the transmission of a message by any neuron.

The neural net

Certain neurons in the configuration mathematically represent the physiological receptors that are excited or left quiescent by the exterior environment. These are called input neurons. Other neurons called output neurons record the logical value, excited or quiescent, of the whole configuration after time delay t and transmit an effect to an exterior environment. All the rest stimulate inner neurons.

Any geometric or logical description of the neural structure of an organism formulated as the basis of physical construction must be sufficiently simple to permit mechanical, electric, or electronic simulation of the neurons and their interconnections.

The basic logical organs. The types of events that can excite the automaton and the kinds of responses that it can make must next be considered. By stripping the description down to the most simple cases, the basic organs from which more complicated robots can be constructed may be discovered. Three basic organs (or elementary automata) are necessary, each corresponding to one of the three logical operations of language: the binary operations of disjunction and conjunction, leading to such propositions as $A \cup B$ (read "A or B"), $A \cap B$ (read "A and B"), and the unary operation of negation or complementation, leading to such propositions as A^c (read "not A" or "complement of A"). First to be considered are the stimulus-response pattern of these elementary automata.

Binary operations of disjunction and conjunction

Assuming that a neuron can be in only one of two possible states—*i.e.*, excited or quiescent—an input neuron at a given instant of time $t-1$ must be either excited or nonexcited by its environment. An environmental message transmitted to two input neurons N_1 and N_2 at time $t-1$ can then be represented numerically in any one of the four following ways, in which binary digit 1 represents excitation and binary digit 0 represents quiescence: (0, 0), (0, 1), (1, 0), (1, 1). The disjunction automaton must be such that a single output neuron M correspondingly registers at time t the response: 0, 1, 1, 1. The conjunction automaton must be such that a single output neuron M correspondingly registers at time t the response: 0, 0, 0, 1. The negation automaton considered as having two input neurons N_1 and N_2, of which N_1 is always excited, must respond to the environmental messages (1, 0) and (1, 1) with 1, 0, respectively, at the output neuron M.

The generalized automaton and Turing's machine. The construction of more complicated robots from these basic building blocks constitutes a large part of the theory of automata. The first step in the direction of generalization is to define the neural nets that correspond to formal expressions in n variables of the propositional calculus—that is, the formal system that concerns "or," "and," "not," and "implies." A single output automaton (of which the above three are simple examples) is a neural net with n input neurons, one output neuron, and with interconnections between neurons that conform to the rule that no neuron stimulated at time t can impinge upon a neuron that could have experienced its first stimulation at the same or an earlier time. The latter rule is the requirement of no feedback. Given this concept of a single output automaton, it is possible to examine the output response at time $t+s$, considered as a function of the configuration of stimuli at the n input neurons at time t. This response can be compared with the truth value of a logical statement (polynomial) from the propositional calculus. A logical statement is formed from n component propositions, each of which can assume the truth value either true or false.

Response of automata and truth value of statements

The comparison between automaton and logical statement is accomplished by matching response at the output neuron at time $t+s$ with truth value of the statement for every one of the 2^n cases in which the configuration of stimuli conforms to the configuration of truth values of the component propositions. If, in this sense of comparison, the functional response of the automaton is identical to the functional value of the logical statement (polynomial), the automaton is then said to compute the statement (polynomial) or the statement is said to be computable. A wider class of computable statements is introduced with the general automaton, yet to be defined, as with the more general Turing machine.

The important distinction between the logical statement and the automaton that computes it is that the first is free of any time ingredient while the second is defined only with reference to a time delay of length s.

A basic theorem states that for any polynomial P of the propositional calculus, there exists a time delay s and a single output automaton A, such that A computes P with time delay s. The proof of the theorem rests on the fact from the propositional calculus that all statements are composed from component propositions with the operations of disjunction, conjunction, and negation and the fact from the automata theory that all single output automata can be composed by interconnecting elementary automata of the disjunctive, conjunctive, and negative types.

A second step of generalization in the construction of robots proceeds from the single output automata to the neural net that possesses more than one output neuron and in which the internal connections may include feedback. Such a construction is called a "general automaton." The class of general automata includes all-purpose, electronic digital computers the memory-storage units of which are of fixed, though possibly of very considerable, size. It is within the context of the general automaton that the purely automated decision-making, computing, controlling, and other sophisticated neural functions so suggestive of the mental ability of human beings may appropriately be discussed.

The Turing machine can be defined not only as it was in the introduction (roughly following Turing's approach) but as a general automaton to which an unbounded memory unit (such as an unbounded tape) is added. Thus, the general automaton and the Turing machine differ in logical design only with respect to the extent of memory storage.

Extent of memory in a Turing machine

The distinction is critical, however, for Turing proposed that the class of numbers computable on his machine (a wider class than can be obtained by general automata) coincide with those that are effectively computable in the sense of constructive logics. A simple convention also makes it possible to interpret the output of a Turing machine as the computation of a function. The class of functions so computed, called "Turing computable" or "computable," are of basic importance at the foundations of mathematics and elsewhere. It can also be stated that a useful class of functions that are definable without reference to machines, namely, the so-called partial recursive functions, has the same membership as the class of computable functions. For the present purposes, then, no effort need be made to define the partial recursive functions.

Turing's approach admitted mathematical formalization to the extent that a finite list of symbols $q_1, q_2, q_3, \cdots, q_n$ could be used to denote internal states and a finite list of symbols $a, b, c, \cdots, \lambda$ could designate abstractly what is called "the alphabet"—that is, the list from which individual members could be chosen and printed into the squares of the machine's tape. If the symbols R and L, respectively, designate a move of the tape one square to the right and one square to the left, it remains only to list in some orderly fashion the alternative possible steps in the machine's operation in order to define it completely. Turing himself chose to list alternate steps, or instructions, in the form of quintuples of the above symbols. It is also possible to use quadruples to define a machine. Such a list, then, of, say, quadruples of instructions is equivalent to a Turing machine, and it is significant that the list is finite.

The finiteness of the list of quadruples of instructions leads to the idea that all Turing machines can be listed—that is, they are at most countable in number. This being the case, it can be proved that there is what Turing called a "universal" machine capable of operating like any given Turing machine. For a given partial recursive function of a single argument, there is a corresponding integer, called the Gödel number, that identifies the Turing machine capable of computing the given function. The Gödel number and the argument value of the function to be computed can be given as input data on the tape of the universal machine. From the Gödel number, the list of instructions, defined in the form of quadruples, that are necessary for the computation of the given recursive function at the specific argument value can be encoded by the universal machine on its own tape, and, from that point on, the universal machine will duplicate the required Turing machine.

The "universal" machine

Input: events that affect an automaton. Once having reached the definition of the general automaton and the more general universal Turing machine, a general definition of the events in the environment that stimulate it may be introduced. The automaton, which computes logical statements, is not defined without reference to time, a characteristic that distinguishes the machine itself from the logic. In the same way, stimuli are not definable, in general, without reference to time. These facts are indicative of the simulation features that the computing machine bears with respect to man.

For an automaton with n input neurons, $N_1, N_2, \cdots, N_n$, an individual history of stimulation, starting with the present moment, $t=0$, and continuing to the remote past, can be recorded as a sequence of n-tuples, $(\beta_1, \beta_2, \cdots, \beta_n)$, in which each binary digit, β_k, is either a 0 or a 1. Thus, the beginning of one such individual history for an automaton of four neurons might be recorded in tabular form as an unending list of quadruples of the type (1, 0, 1, 1) (see Box, display 1).

Events as collections of individual histories

An event is a collection of individual histories. This is a generalization of the idea already used to characterize an environmental message transmitted to the two input

neurons of an elementary automaton at time $t-1$. As an example, the stimulus (0, 1) is the same as the collection of all individual histories in which neuron N_2 was stimulated at time $t-1$ and neuron N_1 was not. As another example, the event that neuron N_2 (of a two-neuron automaton) is presently stimulated and has always been stimulated on alternate second can be represented as the collection of two individual histories (see 2). While some events require an infinite tabulation, others that specify the states of each neuron over a finite past (allowing that anything might have occurred before) permit a finite tabulation. Events of the second kind are called definite events, or stimuli.

The construction (either actual or theoretical) of a general automaton with the help of the logical components and interconnections of a neural net results in an entity that responds in reproducible ways to stimuli. A response becomes recorded as a configuration of binary digits, corresponding to the states of the finite number of output neurons at a specified time t in the future, while a stimulus is a collection of individual histories extending over the past and including the present. The logical construction implies a behaviour in the guise of a listing of responses to all possible stimuli. Reciprocally, for a given behaviour of the type defined, the possible structure of a machine that could produce such behaviour can be investigated.

PROBABILISTIC QUESTIONS

It was traditional in the early treatment of automata theory to identify an automaton with an algorithm, or rule of computation, in which the output of the automaton was a logically determined function of the explicitly expressed input. From the time of the invention of the all-mechanical escapement clock in Europe toward the end of the 13th century, through the mechanistic period of philosophy that culminated in the work of the French mathematician Pierre-Simon Laplace, and into the modern era of the logically defined Turing machine of 1936, an automaton was a mechanical or logical construction that was free of probabilistic components. It was also understood to be immersed in an environment (that is, activated or supplied with input data) that could be logically specified without the concept of chance.

Automata and the idea of chance

After the middle of the 20th century, mathematicians explicitly investigated questions concerning automata that included in their formulation the idea of chance, and in doing so they drew upon earlier applicable mathematical results. While the automata themselves are prototypes of deterministic machines, the U.S. mathematician Norbert Wiener showed that they may be programmed in such a way as to extrapolate certain types of random data that are introduced as input. A prediction of data that are not yet received as input can be accomplished, provided the data are what will later be defined to constitute a stationary time series and provided the prediction is restricted according to a well-defined optimization procedure. In this way a logically defined robot, or automaton, may be placed in an environment that evolves according to both deterministic and random processes (the bifurcation of the environment into deterministic and random processes being mathematically postulated by the designer of the robot) and may be seen to respond to the advantage of its designer: The robot can control a ship's rudder, guide an airplane to its landing, reorient a rocket on its course, predict weather, and so forth. The programming of an automaton so that it will react in a suitable way when placed in a naturalistic environment falls under the heading of prediction theory.

Of the types of probabilistic questions considered, four (which will be listed in arbitrary order) were predominant. The first, that of Wiener, was broached in 1948. It concerned the use of mathematically expressed algorithms or physically constructed computers to predict the future of a system, such as the weather, that includes random components—*i.e.*, an automaton in Turing's logical sense immersed in a random environment. The second, of von Neumann, was concerned with the reliability of large computing machines with many components and sought methods of design, called "multiplexing," that would reduce the chance for unwanted error during the machine calculation of a problem. In this context, the automaton was interpreted as a randomly operating device that in practice approximates the operation of a Turing machine under the influence of better and better design. The third, considered by various researchers, concerned the possibility of computing a wider class of sets than are accessible to Turing machines by adding a random component to the machine itself. In this context, the automaton was being interpreted as a Turing machine modified with the potentiality for injecting the output of a random number generating device into one or more of its operational steps. The fourth concerned the logical possibility of an automaton, such as a Turing machine, actually yielding as output a sequence of random numbers. In this context, the automaton was considered to be simultaneously a Turing machine and a generator of numbers that are indistinguishable from measurements on random phenomena.

Some results that have been achieved from examination of each of these four types of questions will constitute the remainder of this section.

The automaton and its environment. It must first be observed that, just as an automaton is an acceptable description (or model) of a neural structure, an automaton, though frequently thought of as a computing machine, is in general a response mechanism that produces output (or behaviour) as a consequence of the input (or environmental stimuli). "Environment" is then another name for the input and output of an automaton. Some poetic license in identifying automata with living things may justify the use of the term.

During his researches on cybernetics, Wiener recognized that, if computers could be programmed to solve certain mathematical equations, then the data read from physically generated time series (or numerical values indexed consecutively in time and related through a transformation) could be extrapolated. He saw that, if this process could be accomplished with sufficient speed, as would be possible with modern electronic circuits, then the extrapolated values would be obtained faster than the actual physically evolving process that produced the time series, and a prediction of the future would result. Errors would be inevitable because a complete history of data and adequate measurements would be unobtainable. For this reason, the mathematical equations that would be at the heart of such an extrapolation could be deduced, in part, from the objective of minimizing the errors. Thus, the matching of an automaton, or computer, with a real physical environment could result in the anticipation of the future, if certain mathematical equations were derived that minimized prediction error.

Matching an automaton with an environment

Control and single-series prediction. A derivation of the mathematical equations of prediction had been accomplished in a limited sense some years before Wiener's work on cybernetics. In 1931 Wiener had collaborated with an Austrian-born U.S. mathematician, Eberhard Hopf, to solve what is now called the Wiener-Hopf integral equation, an equation that had been suggested in a study of the structure of stars but later recurred in many contexts, including electrical-communication theory, and was seen to involve an extrapolation of continuously distributed numerical values. During World War II, gun- and aircraft-control problems stimulated further research in extrapolation, and Wiener composed a purely mathematical treatise, *Extrapolation, Interpolation, and Smoothing of Stationary Time Series,* which was published in 1949. As early as 1939, a note on extrapolation by a Russian mathematician, A.N. Kolmogorov, had appeared in the French journal *Comptes Rendus.* Although the Wiener-Hopf work was concerned exclusively with astronomy and done without the guiding influence of computers, it was recognized in World War II that high-speed computations could involve input information from a moving object and, through prediction or extrapolation, provide output data to correct its path. This recognition was the seed of the concept of the guided missile and radar-controlled aircraft. Weather prediction was also possible, as was computerized research on brain waves whose traces on the electroencephalograph offered another physical realization of the time series that are predictable. The mathematics

that was necessary for a complete understanding of prediction included the concept of a stochastic process, as described in the article PROBABILITY THEORY.

The Wiener and Kolmogorov research on extrapolation of time series became known as single-series prediction and owed much to the studies (1938) of a Swedish mathematician named Herman Wold, whose work was predicated on the assumption that, if $X_1, X_2, X_3, \cdots$, are successive values of a series identified with discrete points in time $t=1$, $t=2$, $t=3, \cdots$, then the successive values are not entirely unrelated (for if they were, there would be no way for an algorithm or an automaton to generate information about later members of the sequence—that is, to predict). It was assumed, with anticipation that there is frequently such a thing in nature, that a transformation T relates members of the series by successively transforming an underlying space of points ω according to a rule. The rule states that the kth member of the time series is a function of an initial point ω that has migrated in the underlying space $X_k = X(T^k\omega)$. It was also assumed that, if sets of points $\{\omega\}$ constituted a region (of sufficient simplicity called "measurable") in space, then when the set was transformed under the influence of T its volume would not be changed. The last assumption had, in fact, been proved by a French mathematician, Joseph Liouville, a century earlier for a wide class of physical processes whose behaviour is correctly described by the so-called Hamiltonian equations. The clearly stated assumptions of Wiener and Kolmogorov, referred to as the stationarity of the time series, were supplemented with the idea (the linearity restriction) that a solution $S_k(\omega)$ for the predicted value of the series, displaced in time k steps into the future, should be restricted to a linear combination of present and past values of the series (see 3).

(1)
$(1, 0, 1, 0)$
$(1, 0, 1, 1)$
$(0, 0, 0, 1)$
$(1, 0, 1, 0)$
. . . .
. . . .
. . . .

(2)

$(0, 1)$	$(1, 1)$
$(0, 0)$	$(1, 0)$
$(0, 1)$	$(1, 1)$
$(0, 0)$	$(1, 0)$
.	.
.	.
.	.

(3) $$S_k(\omega) = \sum_{n=0}^{\infty} X(T^{-n}\omega)\, P_n^{(k)},$$
with convergence defined in the L_2-metric.

(4)
An algorithm for computing the coefficients $P_n^{(k)}$ in the prediction $S_k(\omega)$ is the following:
From the auto-correlation (X_{-n}, X_0) of the time series compute $|\psi(\theta)|$:
$$\frac{1}{2\pi}\int_{-\pi}^{\pi} |\psi(\theta)|^2 e^{in\theta}\, d\theta = (X_{-n}, X_0).$$
From $|\psi(\theta)|$ compute $\psi(r, \theta)$:
$$\log \psi(r, \theta) = \frac{1}{2\pi}\left\{\sum_{n=1}^{\infty} r^n e^{in\theta}\int_{-\pi}^{\pi} \log |\psi(x)|^2 e^{-inx}\, dx + \int_{-\pi}^{\pi} \log |\psi(x)|\, dx\right\}.$$
From $\psi(r, \theta)$ compute (X, h_{-n}):
$$\psi(r, \theta) = \sum_{n=0}^{\infty} r^n e^{in\theta}(X, h_{-n}), \qquad 0 \leq r < 1.$$
From (X, h_{-n}) compute a_n:
$$(X, h)a_0 = 1$$
$$\sum_{n=0}^{m} (X, h_{n-m})a_n = 0; \qquad m > 0.$$
From a_n and (X, h_{-n}) compute $P_n^{(k)}$:
$$P_n^{(k)} = \sum_{m=0}^{n} a_{n-m}(X, h_{-m-k}), \qquad k > 0.$$

(5) $$\sigma_k^2 = \sum_{n=0}^{k-1} |(X, h_{-n})|^2$$

With the help of one other mathematical assumption, it was then possible to solve the single-series prediction problem by specifying an algorithm that would determine the coefficients in the linear combination for $S_k(\omega)$, in which k is a positive integer (see 4). It was possible also to solve for the error of prediction (see 5)—that is, a measure of the discrepancy between the value predicted and the true value of the series that would occur at time k in the future. This meant that for a variety of circumstances, such as the prediction of atmospheric pressure measured at one weather station, or the prediction of a single parameter in the position specification of a particle (such as a particle of smoke) moving according to the laws of diffusion, an automaton could be designed that could sense and predict the chance behaviour of a sufficiently simple component of its environment.

The assumptions of prediction theory

Multiple-prediction theory. Generalizations of the above limited accomplishments are tantalizing to mathematicians. If animals, and humans in particular, are viewed, even in part, as automata with varying degrees of accomplishment and success that depend on their abilities to cope with their environment, then human beings could be better understood and their potentialities could be further realized by exploring a generalized version of an automaton's ability to predict. Success in generalizations of this kind have already been achieved under the heading of what is called multiple-prediction theory. A reference to the problem of multiple prediction without a complete solution was made as early as 1941 by a Russian mathematician, V. Zasuhin. The first major step forward, after Zasuhin, was taken by Wiener in 1955 under the title "On the Factorization of Matrices." Many significant results soon followed.

If multiple-prediction theory is identified with part of automata theory (which is not always done), it is possible to consider the construction of a computing machine, or automaton, capable of sensing many interdependent elements of its environment at once and, from a long history of such data, of predicting a future that is a function of the same interdependent elements. It is recognized that multiple prediction is the most general approach to the study of the automaton and its environment in the sense that it is a formulation of prediction free of the linearity restriction earlier mentioned with reference to single series (see 3). To express a future point $S_k(\omega)$, for example, as a linear function of its present and past values as well as first derivatives, or rates of change, of its present and past values is to perform a double prediction or prediction based on the two time series $X_1, X_2, X_3, \cdots$; $X'_1, X'_2, X'_3, \cdots$, in which primes indicate derivatives with respect to time. Such double prediction is a first step toward nonlinear prediction.

Automata with unreliable components. In 1956 with the continuing development of faster and more complex computing machines, a realistic study of component misfiring in computers was made. Von Neumann recognized that there was a discrepancy between the theory of automata and the practice of building and operating computing machines because the theory did not take into account the realistic probability of component failure. The number of

Discrepancy between automata theory and practice

component parts of a modern all-purpose digital computer was in the mid-20th century already being counted in millions. If a component performing the logical disjunction (A or B) misfired, the total output of a complex operation could be incorrect. The basic problem was then one of probability: whether given a positive number δ and a logical operation to be performed, a corresponding automaton could be constructed from given organs to perform the desired operation and commit an error in the output with probability less than or equal to δ. Affirmative results have been obtained for this problem by mimicking the redundant structure of parallel channels of communication that is frequently found in nature—*i.e.,* rather than having a single line convey a pulse of information, a bundle of lines in parallel are interpreted as conveying a pulse if a sufficient number of members in the bundle do so. Neumann was able to show that with this redundancy technique (multiplexing) "the number of lines deviating from the correctly functioning majorities of their bundles" could with sufficiently high probability be kept below a critical level.

Automata with random elements. The term algorithm has been defined to mean a rule of calculation that guides an intelligent being or a logical mechanism to arrive at numerical or symbolic results. As discussed above under *Neural nets and automata,* a formalization of the intuitive idea of algorithm has led to what is now called an automaton. Thus, a feature of an automaton is the predictability, or the logical certainty, that the same output would be obtained for successive operations of an automaton that is provided with the same input data. If, as a substitute for the usual input data, random numbers (or results due to chance) are provided, the combination of input data and automaton is no longer completely predictable. It is notable, however, that unpredictable results that might be obtained with the use of uncertain input are not without their practical application. Such a method of combining the operation of a computer with the intentional injection of random data is called the "Monte-Carlo method" of calculation and in certain instances (such as in the numerical integration of functions in many dimensions) has been found to be more efficient in arriving at correct answers than the purely deterministic methods.

Quite apart from the questions of efficiency that might bear upon the addition of an element in a computing machine (automaton) that could produce numbers due to chance, the purely logical question has been asked: Is there anything that can be done by a machine with a random element that cannot be done by a deterministic machine? A number of questions of this type have been investigated, but the first clear enunciation and study of such a question was accomplished in 1956 by the U.S. engineer Claude E. Shannon and others. If the random element in the machine is to produce a sequence of digits 0 and 1 in a random order so that the probability is p for a digit 1 to occur, then (assuming that the number p is, itself, obtainable from a Turing machine as a computable number) the machine can do nothing new, so to speak, as compared to the unmodified Turing machine. This result is precisely expressed in the language of automata theory by saying that the sets enumerated by the automaton with random elements can be enumerated also by the unmodified automaton. The computability of p, however, is critical and is necessary for the result. It is also important to emphasize, in order to distinguish this result from what follows, that the computability of p is under discussion, not the computability of the sequence of random digits.

The computability condition for random elements

Computable probability spaces. Finally, it is to be observed that the concept of chance or random number, wherever it has occurred in the above discussion, submits to the interpretation of result of observation of an experiment or physical phenomenon. The chance ingredients in the weather data to which prediction theory applies could be due to molecular disturbances in the atmosphere that are of diverse and minute origin. The chance failure that might cause a component breakdown in a computing machine is due to the physical structure and environment of the defaulting part. The source of chance that could be used to augment the input of a computer for the purposes of the Monte-Carlo method of calculation may be chosen as the erratic emission of electrons from the cathode of an electronic tube and is frequently so chosen.

An entirely distinct question is involved in relating chance and computers. It would be important to know whether an automaton in the sense defined by Turing can generate random numbers. The question is tantamount to asking whether a Turing machine can logically describe the behaviour of those sources of chance that are found in nature and are the subject of the study of probability theory. Because there are many points of view—too many to consider here—more tightly phrased questions may serve as an introduction to the subject, and a few conclusions that can be brought as answers will be mentioned. At the outset, one limited question can be affirmatively answered: Can a computable sequence of numbers, $S = (a_1, a_2, a_3, \cdots)$, serve as the basic ingredient of a probability theory by providing all of the necessary points in a probability space? In this question the term computable sequence is defined to mean that the numbers a_k are real and there is a Turing machine that, for any pair of positive integers A, B, will print out in order the first A digits of all a_k, for k ranging from 1 to B, in a finite number of steps. It might appear that an affirmative answer to the above question is not striking if simple probability theory alone is considered—that is, a theory of events in which the number of possible outcomes is finite, as in the theory of dice, coins, roulette, and the like. On the other hand, it was shown in 1960 that, although a computable sequence can serve as a set of points in a simple probability space, the mathematical expectations of all random variables X defined on the space can be computed according to an explicit algorithm (see 6) that makes use of the sample values, $X(a_1)$, $X(a_2)$, $X(a_3)$, $\cdots$, which themselves are computable if X is computable. In this algorithm it is evident that the potential number of values to be calculated is infinite, though the number of possible outcomes (distinct values of X) might be finite.

Random numbers that a Turing machine can compute

In the language of the limited question considered, a listing of all sample values (random numbers) of an infinite sequence of statistically independent random variables can be printed out by a Turing machine, at least in the simple case, with strict adherence to the definition of all probabilistic terms as based on measure theory, the theory that generalizes the concept of length.

Extension of such constructions beyond the simple case has also been shown to be possible, provided the concept of a random variable can be extended to a class of functions that are more general than the measure-theoretic class. The most explicit formulation of a suitable generalization was given in 1966, and on the basis of this work it is possible to answer affirmatively a second question: For any sequence of probability distributions, is there a sequence of statistically independent random variables with these respective distributions, each of whose sample values can be computed on a Turing machine and whose mathematical expectations are also attainable by algorithm?

Such results would seem to affront the intuition that tends to divide phenomena into deterministic (or computable) and random (or uncomputable) parts. It is to be observed, however, that in probabilistic matters, passage to the limit and infinite collections are essential ingredients, and such entities are unfamiliar objects in the world in which intuitions are formed. (B.R.)

CLASSIFICATION OF AUTOMATA

All automata referred to from this point on may be understood to be essentially Turing machines classified in terms of the number, length, and movement of tapes and of the reading and writing operations used. The term discrete state automaton is sometimes used to emphasize the discrete nature of the internal states. The principal classes are transducers and acceptors. In automata theory, a transducer is an automaton with input and output; any Turing machine for computing a partial recursive function, as previously described, can stand as an example. An acceptor is an automaton without output that, in a special sense, recognizes or accepts words on the machine alphabet. The input of an acceptor is written on tape in

the usual way, but the tape is blank at the end of the computation, and acceptance of the input word is represented by a special state called a final state. Thus, a word x, or sequence of symbols from an alphabet denoted by the letter S, is said to be accepted by an acceptor A if A computes, beginning in an initial state q_0 with x on tape, and halts in a final state with tape being entirely blank. A subset designated U of the set of words S^* on an alphabet S is called an accepted set if there is an automaton A that accepts any word $x \in U$.

Acceptors. An elementary result of automata theory is that every recursively enumerable set, or range of a partial recursive function, is an accepted set. In general the acceptors are two-way unbounded tape automata.

(6) $$EX = \lim_{N\to\infty} \frac{1}{N} \sum_{k=1}^{N} X(a_k)$$

(7)
1st rule: $\overline{S} \rightarrow \overline{Pr}\ \overline{VP}$
2nd rule: $\overline{NP} \rightarrow \overline{Art}\ \overline{Adj}\ \overline{N}$
3rd rule: $\overline{Adj} \rightarrow \overline{Adv}\ \overline{Adj}$
4th rule: $\overline{VP} \rightarrow \overline{V}\ \overline{NP}$
5th rule: $\overline{Pr} \rightarrow$ she
6th rule: $\overline{Art} \rightarrow$ a
7th rule: $\overline{N} \rightarrow \overline{Adj}\ \overline{N}$
8th rule: $\overline{Adj} \rightarrow$ pretty
9th rule: $\overline{Adj} \rightarrow$ little
10th rule: $\overline{Adv} \rightarrow$ pretty
11th rule: $\overline{N} \rightarrow$ girl
12th rule: $\overline{V} \rightarrow$ is

(8)
Step (i): $\overline{S}$, initial symbol
Step (ii): $\overline{Pr}\ \overline{VP}$, using 1st rule
Step (iii): $\overline{Pr}\ \overline{V}\ \overline{NP}$, using 4th rule
Step (iv): $\overline{Pr}\ \overline{V}\ \overline{Art}\ \overline{Adj}\ \overline{N}$, using 2nd rule
Step (v): $\overline{Pr}\ \overline{V}\ \overline{Art}\ \overline{Adj}\ \overline{Adj}\ \overline{N}$, using 7th rule
Step (vi): she $\overline{V}\ \overline{Art}\ \overline{Adj}\ \overline{Adj}\ \overline{N}$, using 5th rule
Step (vii): she is a pretty little girl, using rules: 12, 6, 8, 9, 11

Generative grammars

A useful classification of acceptors has been introduced in conjunction with a theory of generative grammars developed in the United States by a linguist, Noam Chomsky. A generative grammar is a system of analysis usually identified with linguistics. By its means a language can be viewed as a set of rules, finite in number, that can produce sentences. The use of a generative grammar, in the context of either linguistics or automata theory, is to generate and demarcate the totality of grammatical constructions of a language, natural or automata oriented. A simple grammar for a fragment of English, determined by 12 rules (see 7), can serve to introduce the main ideas.

In this simple grammar, each rule is of the form $g \rightarrow g'$ (read, "g' replaces g") and has the meaning that g' may be rewritten for g within strings of symbols. The symbol $\overline{S}$ that appears in the rules may be understood as standing for the grammatical category "sentence," $\overline{Pr}$ for "pronoun," $\overline{VP}$ for "verb phrase," $\overline{NP}$ for "noun phrase," and so forth. Symbols marked with a vinculum ($^-$) constitute the set V_N of nonterminal symbols. The English expressions "she," etc., occurring in the rules constitute the set V_T of terminal symbols. $\overline{S}$ is the initial symbol.

Beginning with $\overline{S}$, sentences of English may be derived by applications of the rules. The derivation begins with $\overline{S}$; the first rule allows $\overline{Pr}\ \overline{VP}$ to be rewritten for $\overline{S}$, yielding $\overline{Pr}\ \overline{VP}$; the fourth rule allows $\overline{V}\ \overline{NP}$ to be rewritten for $\overline{VP}$, yielding $\overline{Pr}\ \overline{V}\ \overline{NP}$; and so forth (see 8). A last step yields a terminal string or sentence; it consists solely of elements of the terminal vocabulary V_T. None of the rules apply to it; so no further steps are possible.

The set of sentences thus generated by a grammar is called a language. Aside from trivial examples, grammars generate denumerably infinite languages.

Recursively enumerable grammars and Turing acceptors. An elementary result of automata theory is that every recursively enumerable set, or range of a partial recursive function, is an accepted set. In general, the acceptors are two-way unbounded tape automata. On the other hand, a grammar consisting of rules $g \rightarrow g'$, in which g and g' are arbitrary words of $(V_T \cup V_N)^*$ is an unrestricted rewriting system and any recursively enumerable set of words—*i.e.*, language in the present sense—is generated by some such system. These very general grammars thus correspond to two-way acceptors, called Turing acceptors, that accept precisely the recursively enumerable sets.

Finite-state grammars and finite-state acceptors. Acceptors that move tape left only, reading symbol by symbol and erasing the while, are the simplest possible, the finite-state acceptors. These automata have exactly the same capability as McCulloch-Pitts automata and accept sets called regular sets. The corresponding grammars in the classification being discussed are the finite-state grammars. In these systems the rules $g \rightarrow g'$ are restricted so that g is a nonterminal v of V_N (as exemplified above) and g' is of the form us, $u \in V_N$ and $s \in V_T$. The languages generated by finite-state grammars, owing to this correspondence, are called regular languages.

Comparison with McCulloch-Pitts automata

Although these simple grammars and acceptors are of some interest in information theory and in neural network modelling, they are not descriptively adequate for English or for such standard computer languages as Algol because they are not able to account for phrase structure. In particular, finite-state grammars cannot generate self-embedded sentences such as "the man the dog bit ran away," nor can they produce sentences with several readings such as "she is a pretty little girl."

Context-free grammars and pushdown acceptors. Context-free, or phrase-structure, grammars, although apparently not affording completely adequate descriptions of vernacular languages, do have the desirable properties just noted. For this family, the rules $g \rightarrow g'$ contain single nonterminals on the left, as in the finite-state grammars, but allow g' to be any word of $(V_T \cup V_N)^*$. The example discussed above is a context-free grammar. These grammars can account for phrase structure and ambiguity (see 9).

Pushdown acceptors, which play a key role in computer-programming theory, are automata corresponding to context-free grammars. A pushdown acceptor is a finite-state acceptor equipped with an added two-way storage tape, the so-called pushdown store. At the beginning of operation this tape is blank. As the automaton computes, the store is used to analyze the syntactical structure of the sentence being read. The store moves left when printed, and only the last symbol printed may be read, then the next to the last, and so forth. The input is accepted if both the (one-way) input and storage tapes are blank when the automaton halts in a final state.

The representation of Turing machines in quadruple form may be replaced here by a somewhat clearer list of rules that simulate tape action in their application. Rules can be formulated for a pushdown acceptor P for a context-free language L of items xcx^{-1}, in which x is a word on an abstract alphabet $\{a, b\}$ and x^{-1} is x written in reverse. A first such rule can be formulated to mean that, if P is in state q_0 scanning a on input and any (defined) symbol on the pushdown store, it moves tape left, erases a from the input, prints a on the store, and goes into state q_1. A symbolic expression for the rule might be: $q_0a \rightarrow aq_1$. Another rule might be of the form: if P is in state q_1 scanning c on input and anything on store, it moves input left, erases c, and does nothing with respect to the store—briefly, $q_1c \rightarrow q_2$. Another requires that, if P is in q_2 scanning a on input and a on store, then it moves input left, erases a, moves store right, and erases a (see 10). An example is easily constructed to show that under certain rules a set, say, *abcba* is accepted (see 11). If q_0abcba indicates

(9)

At step (v) of the derivation in (8),
the 3rd rule of (7) could have been applied to (iv), yielding

(va) $\overline{\text{Pr}}\ \overline{\text{V}}\ \overline{\text{Art}}\ \overline{\text{Adv}}\ \overline{\text{Adj}}\ \overline{\text{N}}$.

Successive steps would again result in (vii), but now to be read with a different meaning than before because of an immediate constituent structure differing from (v). To emphasize this divergence the replacing word g' flanked by parentheses, (g'), may be inserted into a previously obtained word in a derivation. Thus at (ii),

(ii)′ $(\overline{\text{Pr}}\ \overline{\text{VP}})$

would be written; and at (iii),

(iii)′ $(\overline{\text{Pr}}(\overline{\text{V}}\ \overline{\text{NP}}))$.

Proceeding in this way, (v) and (va) would appear as

(v)′ $(\overline{\text{Pr}}\ (\overline{\text{V}}\ (\overline{\text{Art}}\ \overline{\text{Adj}}\ (\overline{\text{Adj}}\ \overline{\text{N}}))))$

and

(va)′ $(\overline{\text{Pr}}\ (\overline{\text{V}}\ (\overline{\text{Art}}\ (\overline{\text{Adv}}\ \overline{\text{Adj}})\ \overline{\text{N}})))$.

Sentences in which phrase structure is indicated by parentheses are *phrase markers*. Making the further substitutions (vi) and (vii) within (v)′ and (va)′ produces two phrase markers, which represent two readings of (vii).

(10)

1st rule: $q_0a \rightarrow aq_1$
2nd rule: $q_0b \rightarrow bq_1$
3rd rule: $q_1a \rightarrow aq_1$
4th rule: $q_1b \rightarrow bq_1$
5th rule: $q_1c \rightarrow q_2$
6th rule: $aq_2a \rightarrow q_2$
7th rule: $bq_2b \rightarrow q_2$

Accepted sets in specific calculations

the outset of a computation with P in the initial state q_0 scanning the first a in $abcba$ on input tape and blank on store tape, and if q_2 is a final state, then the computation is determined by the rules given above (see 10). At the end of the computation the automaton is in a final state q_2, both tapes are blank, and there is no rule with q_2 alone on the left; P halts and hence $abcba$ is accepted.

Reflection on the example and on others easily constructed shows that a pushdown acceptor is able, in effect, to parse sentences of context-free languages.

Context-sensitive grammars and linear-bounded acceptors. A fourth type of acceptor, which is mainly of mathematical rather than applied interest, is the two-way acceptor with bounded tape—*i.e.*, tape the length of which never exceeds a linear function of the input length. These are the linear-bounded acceptors. They correspond in the present classificatory scheme to context-sensitive grammars. Unlike the context-free grammars, these latter systems use rules $g \rightarrow g'$, in which the nonterminal symbol $v \in V_N$ in g may be rewritten only in a context xvy; thus $g \rightarrow g'$ is of the form $xvy \rightarrow xwy$, $x, y, w \in (V_T \cup V_N)^*$. An example of a context-sensitive language accepted by a linear-bounded automaton is the copy language xcx.

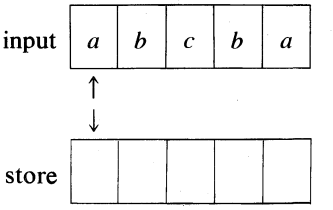
The family of recursively enumerable languages includes the context-sensitive languages, which in turn includes the context-free, which finally includes the regular, or finite-state, languages. No other hierarchy of corresponding acceptors has been intensively investigated.

Finite transducers. The most important transducers are the finite transducers, or sequential machines, which may be characterized as one-way Turing machines with output. They are the weakest with respect to computing power, while the universal machine (see above *The generalized automaton and Turing's machine*) is the most powerful. There are also transducers of intermediate power.

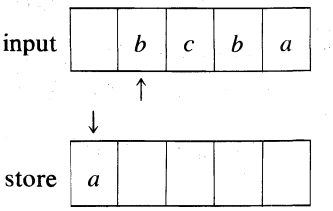
Algebraic definition. Because the tape is one-way with output, a finite transducer T may be regarded as a "black box" with input coming in from the right and output being emitted from the left. Hence, T may be taken to be a quintuple $\langle S, Q, O, M, N\rangle$, in which S, Q, and O are finite,

(11)

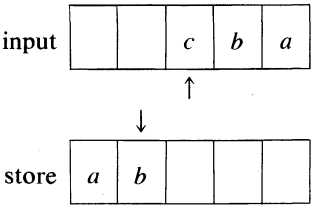
(i) q_0abcba initial tape
This represents the following situation



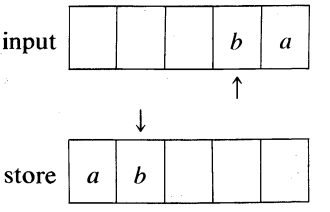
(ii) aq_1bcba by 1st rule (see 10)



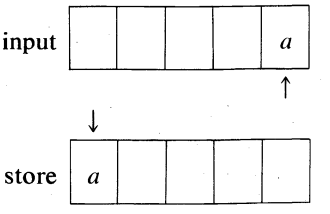
(iii) abq_1cba by 4th rule (see 10)



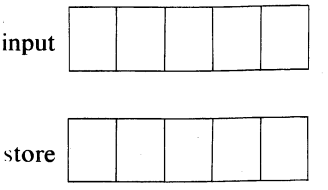
(iv) abq_2ba by 5th rule (see 10)



(v) aq_2a by 7th rule (see 10)



(vi) q_2 by 6th rule (see 10)



(12)

$M(q, \Lambda) = \Lambda$
$M(q, xs) = M(M(q, x), s)$
$N(q, \Lambda) = \Lambda$
$N(q, xs) = N(M(q, x), s)$,

in which the empty word $\Lambda \in S^*$, Λ is adjoined to 0, $x \in S^*$, and $q \in Q$.

(13) Let $\psi: Q_a \to Q_b$ be a map from the states of T_a into those of T_b. The map ψ is a homomorphism if furthermore,
$\psi M_a(q, s) = M_b(\psi(q), s)$
and
$N_a(q, s) = N_b(\psi(q), s)$ for all $s \in S$.

(14) $\phi_s(q) = q'$ if and only if
$M(q, s) = q' \qquad (q, q' \in Q \text{ of } T)$

(15) The *serial connection* of two transducers $T_a = \langle S, Q_a, O_a, M_a, N_a \rangle$ and $T_B = \langle O_a, Q_b, O_b, M_b, N_b \rangle$ is the transducer
$T = T_A \to T_B = \langle S, Q_a \times Q_b, O, M, N \rangle$.
$Q_a \times Q_b$ is the Cartesian product of states, and M, N are given by
$M(\langle q_a, q_b \rangle, s) = \langle M_a(q_a, s), M_b(q_b, N_a(q_a, s)) \rangle$
and
$N(\langle q_a, q_b \rangle, s) = N_b(q_b, N_a(q_a, s))$.

(16) The *parallel connection* of T_A and T_B is the system
$T = T_A \times T_B$
in which
$M(\langle q_a, q_b \rangle, s) = \langle M_a(q_a, s)\, M_b(q_b, s) \rangle$
and
$N(\langle q_a, q_b \rangle, s) = \langle N_a(q_a, s), N_b(q_b, s) \rangle$.

nonempty sets of inputs, states, and outputs, respectively, and M is a function on the product $Q \times S$ into Q and N is a function on the same domain into O. The values are written in the usual functional notation $M(q, s)$, and $N(q, s)$, $s \in S$ and $q \in Q$. M and N may be extended to the domain $Q \times S^*$ by four relations (see 12).

Equivalence and reduction. The most natural classification is by equivalence. If two machines (finite transducers) share the same inputs, then representative states from each are equivalent if every sequence x belonging to the set of words on the alphabet causes the same output from the two machines. Two finite transducers are equivalent if for any state of one there is an equivalent state of the other, and conversely. Homomorphisms between transducers can also be defined (see 13). If two automata are onto homomorphic they are equivalent, but not conversely. For automata that are in a certain sense minimal, however, the converse holds.

Homomorphisms between transducers

Each equivalence class of transducers contains a smallest or reduced transducer—that is, one having the property that equivalence between its states implies equality. There is an algorithm for finding the reduced transducer of a class, which proceeds in a natural way from equivalence classes or blocks of states of a given transducer, each such block being defined as a state of the reduced transducer. Reduced equivalent finite transducers are unique up to an isomorphism—that is to say, if two finite transducers are reduced and equivalent, they differ only in the notations for their alphabets.

Classification by semi-groups. A mathematically significant classification of transducers may be obtained in terms of the theory of semi-groups. In outline, if the transducer T is reduced, the functions φ_s given in terms of M, for fixed input, as maps from and to the space of states Q constitute a semi-group termed the semi-group of T (see 14). By a certain procedure these semi-groups and their associated transducers T may be decomposed into more elementary systems called serial-connected and parallel-connected transducers. In explanation, the next state (starting from state q_a, q_b in the serially connected machine $T_A \to T_B$ is the pair of states made up of the next state in T_A from q_a with input s and the next state in T_B from q_b with input N_a (q_a, s)—which latter is the output of T_a (see 15). Schematically, the connection may be depicted, indicating that in a serial connection the output of T_A is the input to T_B.

The parallel connection of two transducers is a system that may be rigorously defined (see 16) and that may be schematically depicted with input leading in parallel to both machines and output leading in parallel out of both machines. It has been shown that any finite transducer whatsoever can be decomposed into a system of series-parallel-connected automata, such that each element is either a two-state automaton or one whose semi-group is a simple group. This affords a classification of machines that depends ultimately on the determination of the simple groups of finite order.

Serial and parallel connections between transducers

An earlier decomposition scheme was based on a generalization of the concept of congruence relations over sets of states, but discussion of it is omitted here.

Post machines. Types of automata have been investigated that are structurally unlike Turing machines though the same in point of computational capability. The mathematician E.L. Post (U.S.) proposed in 1936 a kind of automaton (or algorithm) that is a finite sequence of pairs $\langle 1, a_1 \rangle, \langle 2, a_2 \rangle, \cdots, \langle m, a_m \rangle$, such that a_i is either an instruction to move an associated two-way tape one square right or left, an instruction to print a symbol, including a blank, from a finite alphabet, or an integer. A Post machine begins at 1 and at step n obeys the instruction a_n and then goes to step $n + 1$, unless a_n is an integer m, in which case it goes to step m if the square scanned at n is marked or to step $n + 1$ if that square is blank. Post machines are prototypes of the program schemes developed 10 years later by von Neumann and his associates. For any partial recursive function a Post machine can be found that is capable of computing it.

Generalizations to automata or information processors in which the restriction to finiteness on sets is dropped or in which additional information from arbitrary sets is available to a machine during computation continue to be considered in the literature. (R.J.Ne.)

BIBLIOGRAPHY. MARVIN L. MINSKY, *Computation: Finite and Infinite Machines* (1967); and R.J. NELSON, *Introduction to Automata* (1967), are the most comprehensive elementary introductions. MICHAEL A. ARBIB, *Theories of Abstract Automata* (1969), is an advanced introduction. MARTIN DAVIS, *Computability and Unsolvability* (1958); and HARTLEY ROGERS, JR., *Theory of Recursive Functions and Effective Computability* (1967), are concerned with the concepts of Turing computability and the theory of recursive functions. C.E. SHANNON and JOHN MCCARTHY (eds.), *Automata Studies* (1956), contains some of the original basic material concerning neural nets and automata with unreliable components or with random elements. BAYARD RANKIN (ed.), *Differential Space, Quantum Systems, and Prediction* (1966), discusses the automaton and its environment in the sense of prediction theory and gives reference to other literature in this area as well as the area of computable probability spaces. A good account of automata theory and its relations to switching theory is MICHAEL A. HARRISON, *Introduction to Switching Automata and Theory* (1965). The best introduction to machine decomposition theory is J. HARTMANIS and R.E. STEARNS, *Algebraic Structure Theory of Sequential Machines* (1966). NOAM CHOMSKY, "Formal Properties of Grammars," in R. DUNCAN LUCE, ROBERT R. BUSH, and EUGENE GALANTER (eds.), *Handbook of Mathematical Psychology*, vol. 2 (1963), is still the best survey of the field of automata and generative grammars. Articles presenting approaches to languages and automata from very general mathematical points of view are SEYMOUR GINSBURG and SHEILAH GREIBACH, "Abstract Families of Languages," *Mem. Am. Math. Soc.*, no. 87, pp. 1–32 (1969); and GENE F. ROSE, "Abstract Families of Processors," *J. Comput. & Syst. Sci.*, 4:193–204 (1970).

(B.R./R.J.Ne.)

Automation

The term automation was coined around 1946 by the automobile industry to describe the increased use of automatic devices and controls in mechanized production lines. Today, it is widely used in a manufacturing context but is also applied outside of manufacturing in connection with a variety of systems in which there is a significant substitution of mechanical, electrical, or computerized action for human effort and intelligence. An operation is commonly described as automated if it is substantially more automatic than its predecessor.

In its most general usage, automation can be defined as a technology concerned with carrying out a process by means of programmed commands combined with the automatic feedback of data relating to the execution of those commands. The resulting system is capable of operating without human intervention. The development of this technology has become increasingly dependent on the use of computers and computer-related technologies. As a consequence, automated systems have become sophisticated and complex. Advanced systems of this sort now represent a level of capability and performance that surpass in many ways the abilities of humans to accomplish the same activities.

Automation technology has matured to a point where a number of other technologies have developed from it and have achieved a recognition and status of their own. Robotics is one such technology. It is a specialized branch of automation in which the automated machine possesses certain anthropomorphic, or humanlike, characteristics. The most typical humanlike characteristic of a modern industrial robot is its powered mechanical arm. The robot's arm can be programmed to do a sequence of motions to perform useful tasks, such as loading and unloading parts at a production machine or making a sequence of spot-welds on the body of an automobile. The robot will repeat the motion pattern until it is reprogrammed to perform some alternative task. As these examples of robot applications suggest, an industrial robot is typically used to replace a human worker in a factory operation.

This article covers the fundamentals of automation and robotics, including the basic components of an automated (or robotic) system, the technical and economic advantages and drawbacks of such systems, and their key applications. The article also reviews the historical development of automation and considers its impact on society. For related topics, see COMPUTERS and INFORMATION PROCESSING AND INFORMATION SYSTEMS.

This article is divided into the following sections:

THE DEVELOPMENT OF AUTOMATION

The technology of automation has evolved from mechanization, which had its beginnings in the Industrial Revolution. (The term mechanization is used to refer to the replacement of human [or animal] power with mechanical power of some form.) Mechanization, in turn, developed out of the human propensity to create tools and mechanical devices.

The first tools made of stone represented prehistoric man's attempts to direct his own physical strength under the control of human intelligence. Thousands of years were undoubtedly required for the development of simple mechanical devices and machines such as the wheel, the lever, and the pulley, by which the power of human muscle could be magnified. The next extension was the development of powered machines that did not require human strength to operate. Examples of these powered machines include waterwheels, windmills, and simple steam-driven devices. More than 2,000 years ago, the Chinese developed trip-hammers powered by flowing water and waterwheels. The early Greeks experimented with simple reaction motors powered by steam. Windmills, with mechanisms for automatically turning the sails, were developed during the Middle Ages in Europe and the Middle East. The steam engine represented a major step forward in the development of these powered machines and marked the beginning of the Industrial Revolution. Since the appearance of the Watt steam engine (developed in 1765 by the Scottish inventor James Watt), powered engines and machines have been devised that obtain their energy from steam, electricity, and chemical, mechanical, and nuclear sources.

Each new development in the history of powered machines has brought with it an increased requirement for control devices to harness the power of the machine. The earliest steam engines required a person to open and close the valves, first to admit steam into the piston chamber and then to exhaust it. Later, the development of a slide valve mechanism that was coupled to the piston shaft made it possible to accomplish these functions automatically. The only need of the human operator was then to regulate the amount of steam that controlled the engine's speed and power. Finally, the requirement for human attention in the operation of the steam engine was eliminated by James Watt's flyball governor introduced during the late 1780s. This device consisted of a weighted ball on a hinged arm, which was mechanically coupled to the output shaft of the engine. As the rotational speed of the shaft increased, centrifugal force caused the weighted ball to be moved outward. This motion controlled a valve that reduced the steam being fed to the engine, thus slowing the engine. The flyball governor remains an elegant early example of a negative feedback control system, in which the increasing output of the system is used to decrease (*i.e.*, add a negative value to) the activity of the system.

Need for control devices

Negative feedback is a widely applied means of automatic control used to achieve a constant operating level for a system. A common example of a modern negative feedback control system is the home thermostat. In this

device a rise in room temperature causes a bimetallic strip to flex, opening an electrical switch that turns off the furnace. As the room cools down, the bimetallic strip flexes in the opposite direction, closing the switch and turning on the furnace. The switch can be set to start up the furnace at a particular set point (*e.g.*, 70° F).

Another important development in the history of automation was the Jacquard loom, which demonstrated the concept of a programmable machine. This automatic loom, introduced by the French inventor Joseph-Marie Jacquard in 1805, was capable of producing complex patterns in textiles by controlling the motions of many shuttles of different coloured threads. The selection of the different patterns was determined by a program contained in steel cards in which holes were punched. These cards were the ancestors of the paper cards and tapes that control various modern automatic machines. The concept of machine programming was further developed later in the 19th century when Charles Babbage, an English mathematician, proposed a complex, mechanical device that could perform arithmetic operations, data processing, and limited decision making. Although Babbage was never able to complete his so-called Analytical Engine, it is generally considered the precursor of the modern programmable digital computer.

ıilding ocks auto- ation

The historical developments described above provided the four basic building blocks of automation: (1) a source of power to perform some action, (2) feedback controls, (3) machine programming, and (4) decision making. Over the years these fundamental elements have been refined and enhanced, so that modern automated systems can operate virtually without human intervention.

Some of the most significant refinements and enhancements of the four building blocks of automation have occurred during the 20th century and include developments in electronics leading to the electronic digital computer, improvements in program storage technology, development of new software used to write the programs, advances in sensor technology, and the derivation of a mathematical theory of control systems.

The development of the electronic digital computer (the ENIAC [Electronic Numerical Integrator and Calculator] in 1946 and UNIVAC I [Universal Automatic Computer] in 1951) has permitted the control function in automation to become much more sophisticated and the associated calculations to be executed much faster than previously possible. The trend in computer technology has led to machines that are markedly smaller and less expensive than their predecessors yet capable of operating at much greater speeds. This trend is represented today by the microprocessor (and corresponding microcomputer), miniature multicircuited devices capable of performing all of the logic and arithmetic functions of large digital computers.

npact ımputer ıd lated chnolo- es

Along with the advances in computer technology that have occurred since the mid-1940s, there have been parallel improvements in program storage technology for holding programming commands. Modern storage mediums include magnetic tapes and disks; magnetic bubble memories; optical data storage read by lasers; videodisks; and electron beam-addressable memory systems. In association with such developments in computer and storage technology, improvements have been made in the methods by which computers (and other programmable machines) are programmed.

Advances in sensor technology have provided a vast array of measuring devices that can be used as components in automatic feedback control systems. These devices include highly sensitive electromechanical probes, scanning laser beams, systems involving the use of electrical field techniques, and machine vision. Some devices and systems of this type require computer technology for their implementation. Machine vision, for example, requires the processing of enormous amounts of data, and this can only be accomplished by high-speed digital computers. This particular technology is proving to be a versatile sensory capability for accomplishing various sophisticated industrial tasks (*e.g.*, part identification and product inspection), as well as providing the basis for robot guidance systems.

Finally, there has evolved since World War II a highly advanced mathematical and logical theory of control systems. The theory includes traditional negative feedback control, optimal control, adaptive control, and artificial intelligence. Traditional feedback control theory makes use of linear ordinary differential equations to analyze problems such as that of Watt's flyball governor. This theory is basic in modern mechanical, electrical, and chemical engineering curricula. Both optimal control theory and adaptive control theory are concerned with the problem of defining an appropriate index of performance for the process of interest and then operating the process in such a manner as to optimize its performance. The difference between optimal and adaptive control is that the latter must be implemented under conditions of a continuously changing and unpredictable environment. Artificial intelligence is an advanced field of computer science in which the computer is programmed to exhibit characteristics that are commonly associated with human intelligence. These characteristics include the capacity for learning, understanding language, reasoning, solving problems, and rendering expert diagnoses of a condition or situation. Developments in artificial intelligence are expected to provide robots and other "intelligent" machines with the ability to communicate with humans and to accept very high-level instructions rather than the detailed step-by-step programming statements typically required by present-day programmable machines. In the future, a robot endowed with artificial intelligence may, for example, be capable of accepting and executing the terse command "Assemble radio." By contrast, existing industrial robots would have to be provided with a detailed set of instructions that would specify the locations of the radio components, the particular components to assemble first, and so forth.

Artificial intelligence and robotics

The various developments described here have made possible a wide range of automated systems both in industrial and nonindustrial applications. Examples of automated systems in industry are numerically controlled machine tools, industrial robots, automated guided-vehicle systems, and automated storage and retrieval systems. Automated systems commonly encountered in everyday life include programmable household appliances, microprocessor-controlled automobile engines, and automatic bank teller machines. (M.P.G./M.T.)

THE DEVELOPMENT OF ROBOTICS

The field of robotics has its roots in the development of automation technology. Numerical control (NC) and telecherics are two important areas of technology that constitute the foundations of robotics technology.

Importance of numerical control and telecherics

Numerical control is a method of controlling machine tool axes by means of numbers and other symbols that have been coded on a medium such as punched paper tape. It was developed during the late 1940s and early 1950s. The first NC machine tool was demonstrated in 1952 at the Massachusetts Institute of Technology, Cambridge. Subsequent research there led to the development of the APT (Automatically Programmed Tools) language for programming machine tools.

Telecherics is concerned with the use of remote manipulators controlled by humans. A remote manipulator is a mechanical arm and hand that translates the motions of a human being at one location into motions at a remote location. Such a device is sometimes called a teleoperator. Initial work on the design of teleoperators can be traced to the development of methods for handling radioactive materials in the early 1940s.

Industrial robotics might be regarded as a combination of numerical control and telecherics. Numerical control provided the concept of a programmable industrial machine, while telecherics contributed the notion of a mechanical arm that could be utilized to perform useful work. The first industrial robot was installed in 1961 to unload parts from a die-casting operation. Its development was due largely to the efforts of two Americans—George C. Devol, an inventor, and Joseph F. Engelberger, a businessman. Devol originated the design for a programmable manipulator, the patent for which he titled "Programmed Article Transfer." The U.S. patent for the device was issued in

1961. Engelberger teamed with Devol to promote the use of robots in industry and, together, the two founded the first corporation in robotics, Unimation, Inc., Danbury, Conn.

PRINCIPLES AND THEORY OF AUTOMATION

As noted earlier, the four basic building blocks of an automated system are (1) a source of power to perform some action, (2) feedback controls, (3) machine programming, and (4) decision making. Any automated system must exhibit at least the first three of these elements. Modern automated systems often possess the fourth element as well.

Power source. An automated system is designed to accomplish some useful action, and that action invariably requires power. There are many sources of power available, but the most commonly used form of energy in present-day automated systems is electricity. Electrical energy is the most versatile because it can be generated readily from many sources (*e.g.,* fossil fuel, hydroelectric, solar, and nuclear) and to perform useful work can be converted readily into several types of power (*e.g.,* mechanical, hydraulic, and pneumatic). In addition, electrical energy can be stored in high-performance, long-lived batteries.

Types of actions performed by automated systems

The actions performed by automated systems are generally of two types: (1) processing, and (2) transfer and positioning. In the first case, energy is applied to accomplish some processing operation on some entity. The process may involve shaping metal, molding plastic, switching electrical signals in a communication system, or processing data in a computerized information system. All of these actions entail the use of energy to transform the entity (*e.g.,* metal, plastic, electrical signals, or data) from one state or condition into another more valuable state or condition. The second type of action performed by automated systems, transfer and positioning, is most readily conceptualized in automated manufacturing systems designed to perform work on a product. In such cases, the product must generally be moved (transferred) from one location to another during the series of processing steps. At each processing location, positioning of the product is often required. The transfer and positioning actions are called materials handling in automated production systems. In automated communications and information systems, the terms transfer and positioning refer to the movement of data (or electrical signals) among various processing units and the delivery of information to output terminals (printers, video display units, etc.) for interpretation and use by humans.

Feedback controls. Feedback controls are widely used in modern automated systems. A feedback control system consists of five basic components: (1) input, (2) process being controlled, (3) output, (4) sensing elements, and (5) controller and actuating devices. These five components are illustrated in the diagram of Figure 1. The term closed-loop feedback control is often used to describe this kind of system.

Basic components of a feedback control system

The input to the system is the reference value, or set point, for the system output. This represents the desired operating value of the output. With the previous example of the home heating system as an illustration, the input is the desired temperature setting for the room. The process being controlled is the furnace that provides heat to the room. In other feedback systems, the process might be a manufacturing operation, the rocket engines of the U.S. space shuttle orbiter, the automobile engine in a so-called cruise control system, or any of a variety of other operating mechanisms to which the action power is applied. The output is the variable of the process that is being measured and compared to the input. In the example, room temperature is the variable of interest.

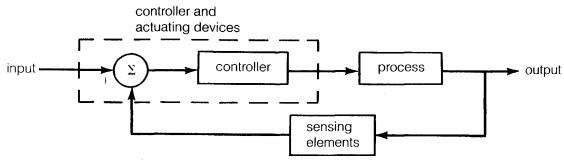


Figure 1: The five components of a feedback control system.

The sensing elements are the measuring devices used in the feedback loop to monitor the value of the output variable. In the home heating system example, a bimetallic strip in the thermostat performs the sensing function. The two different metals in the bimetallic strip possess different coefficients of thermal expansion; accordingly, the flex (deflection) of the strip is directly related to temperature. As such, the strip is capable of measuring temperature. Many different kinds of sensors are used in feedback control systems for automation.

The purpose of the controller and actuating devices in the feedback system is to compare the measured output value with the reference input value and to reduce the difference between them. In general, the controller and actuator of the system are the mechanisms by which changes in the process are accomplished to influence the output variable. These mechanisms are usually designed specifically for the system and consist of devices such as motors, valves, solenoid switches, piston cylinders, gears, power screws, pulley systems, chain drives, and other mechanical and electrical components. The switch connected to the bimetallic strip of the thermostat is the controller and actuating device for the home heating system. When the output (room temperature) is below the set point, the switch turns on the furnace. When the temperature reaches or slightly exceeds the set point, the furnace is turned off.

Machine programming. The programmed commands determine the actions that are to be accomplished automatically by the system. These commands specify what the automated system should do and how the various components of the system must function to accomplish the desired result. The content of the program varies considerably from one automated system to the next. In relatively simple systems, the program specifies a limited number of well-defined actions that are performed continuously and repeatedly in the proper sequence with no deviation from one cycle to the next. In more complex systems, the number of commands could be large and the level of detail in each command could be significantly greater. In relatively sophisticated systems, it is also possible to readily change the program to alter the sequence of actions to be performed by the system.

Programming commands are related to feedback control in an automated system in the sense that the program establishes the sequence of values for the inputs (set points) of the various feedback control loops that make up the system. A given programming command may specify the set point for the feedback loop, which in turn controls some action that the system is to accomplish. In effect, the purpose of the feedback loop is to verify that the programmed step has been carried out. In a robot controller, for example, the program might indicate that the arm is to move to a specified position. The feedback control system in this case would be used to verify that the move has been correctly made. The relationship of program control and feedback control in an automated system is illustrated in Figure 2.

Feedback loop

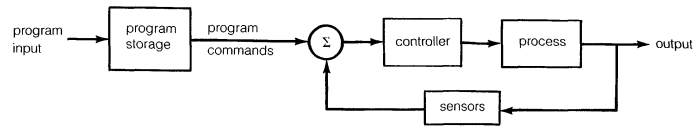


Figure 2: Relationship of program control and feedback control in an automated system.

Some of the programmed commands may be executed in an open-loop fashion—*i.e.,* without the need for a feedback loop to verify that the command has been properly carried out. For instance, a command to flip an electrical switch may not require feedback. An example of the need for feedback control in an automated system exists in a situation where there are variations in the raw materials being fed into a production process, and the system must take these variations into consideration by making adjustments in its controlled actions. Without feedback, the system would not be able to exercise a sufficient level of control over the quality of the process output.

The programmed commands may be contained on me-

chanical devices (*e.g.,* mechanical cams and linkages), punched paper tape, magnetic tape, magnetic disk, computer memory, or any of a variety of other mediums that have been developed over the years for particular applications. It is common today for automated equipment to use computer storage technology as the means for storing the programmed commands and converting them into controlled actions. One of the advantages of using computer storage technology is that that program can be easily changed or improved. Altering a program contained on a set of mechanical cams involves considerably more work.

Decision making. Most highly sophisticated automated systems are capable of making decisions during operation. The decision-making capacity is generally contained in the control program in the form of logical instructions that govern the operation of such a system under varying circumstances. Under one set of circumstances, the system responds one way; and under a different set of circumstances, it responds in another. There are several reasons for providing an automated system with decision-making capability. These reasons include (1) error detection and recovery, (2) safety protection, (3) interaction with humans, and (4) process optimization.

Error detection and recovery is concerned with decisions that must be made by the system in response to undesirable operating conditions. In the operation of any automated system, some form of corrective action must be taken to restore the system when malfunctions and errors occur during the normal cycle of operations. The typical response to a system malfunction has been to call for human assistance. There is a growing trend in automation and robotics to enable the system itself to sense these malfunctions and to correct them in some manner without human intervention. This sensing and correction (referred to as error detection and recovery) can be realized by programming a decision-making capability into the system.

ensing
nd
orrecting
alfunc-
ons
ithout
uman in-
ervention

Safety protection is a special case of error detection and recovery in which the malfunction involves a safety hazard. Decisions are required when the automated system detects, through its sensors, that there has developed a safety condition that is hazardous to either the equipment or humans in the vicinity of the equipment. The terms safety monitoring system and hazard monitoring system are typically used to refer to that portion of the automated system that includes the safety sensors and decision-making apparatus. The purpose of the safety monitoring system is to detect the hazardous condition and to take the most appropriate action to remove or reduce it. This may involve stopping the operation and alerting the maintenance personnel of the condition, or it may involve a more complex set of actions to eliminate the hazard.

Some automated systems are required to interact with humans in some way. An automatic bank teller machine, for example, must receive instructions from customers and make decisions according to these instructions. In some automated systems, a variety of different instructions from humans is possible, and the decision-making capability of the system must be sophisticated to deal with the array of possibilities.

rocess op-
mization

A fourth reason to endow an automated system with decision-making capacity is to optimize the process being controlled. This need for optimization occurs most commonly in production situations in which there is an economic performance criterion for the process and it is desirable to optimize this criterion. For example, minimizing cost is usually a key objective in manufacturing. An automated system would make use of optimal control or adaptive control principles in its program to receive appropriate sensor signals and other inputs and make decisions to drive the process toward the optimal state.

TECHNOLOGY OF ROBOTICS

The most widely accepted definition of an industrial robot is one developed by the Robotic Industries Association:

> An industrial robot is a reprogrammable, multifunctional manipulator designed to move materials, parts, tools, or specialized devices through variable programmed motions for the performance of a variety of tasks.

The technology of robotics is concerned with the design of the mechanical manipulator and the computer systems used to control it. The technology is also concerned with the industrial applications of robots. These will be dealt with below in the section *Manufacturing applications of automation and robotics.*

The mechanical manipulator of an industrial robot is made up of a sequence of link and joint combinations. The links are rigid members connecting the joints. The joints, also called axes, are the movable components of the robot that cause relative motion between adjacent links. Four principal types of mechanical joints are used to construct the manipulator—namely, a linear joint and three types of rotational joints. Figure 3 illustrates the four types. One way to define a robot is by the number of joints used in its construction. Typically three joints are used for a robot's arm and body and two or three joints for its wrist. This

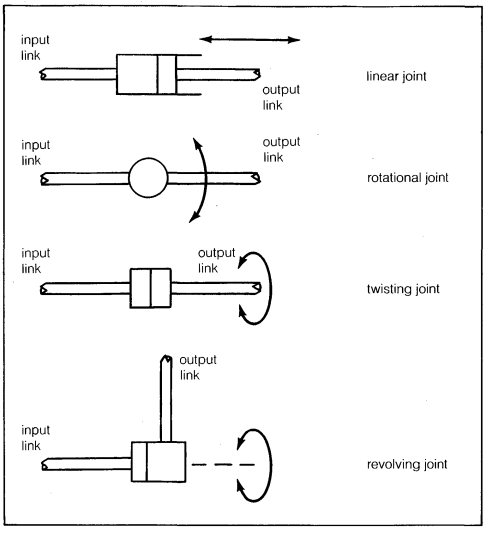


Figure 3: Four types of robot manipulator joints.

permits the robot to position and orient parts and tools in the work space. One possible configuration for a six-axis robot is pictured in Figure 4.

The computer system that controls the manipulator must be programmed to teach the robot the particular motion sequence and other actions it must perform to accomplish

By courtesy of Cincinnati Milacron, Inc.

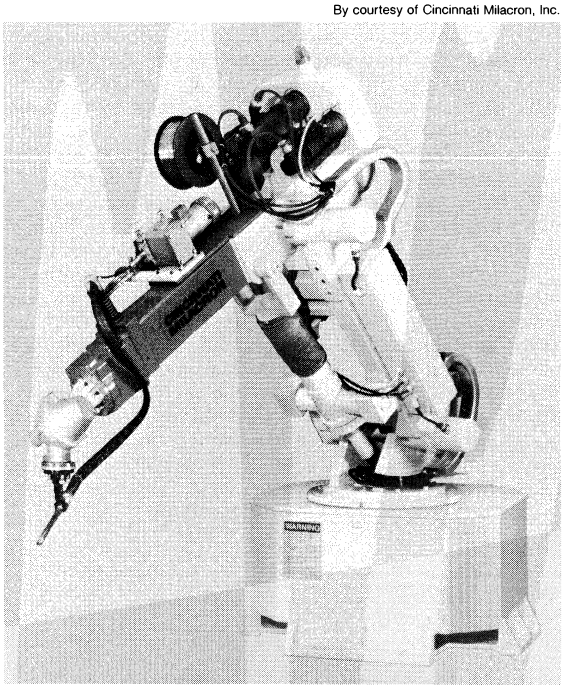


Figure 4: A six-axis, computer-controlled robot equipped for arc welding.

Programming industrial robots

its task. There are several ways that industrial robots are programmed. One method, called lead-through programming, requires that the manipulator be driven through the various motions needed to perform a given task and that those motions be recorded into the robot's computer memory. This can be done either by physically moving the manipulator through the motion sequence or by power driving the manipulator through the sequence using a control box called a teach pendant.

A second method of programming involves the use of a textual programming language very much like a computer programming language. However, in addition to many of the capabilities of a computer programming language (*i.e.,* data processing, communicating with other computer devices, and decision making), the robot language also includes statements specifically designed for robot control. The latter involves motion control and input/output commands. Motion control commands direct the robot to move its manipulator to some defined position in space. The statement "Move P1," for example, might be used to direct the robot to a point in space called P1. Input/output commands are employed to control the receipt of signals from sensors and other devices in the work cell, as well as to initiate control signals to other pieces of equipment in the cell. For instance, the statement "Signal 3, On" might be used to turn on a motor in the cell by means of output line number 3 in the robot's controller.

Robots of the future are likely to receive instructions (programming commands) by voice input/output. They would therefore have to have artificial intelligence, which would make possible the proper interpretation and execution of the commands. The commands would be expressed in a very high-level language. Unlike today's programming languages in which detailed instructions must be provided in precise syntax, those of the future would be less demanding in terms of format, and the commands would be more task-oriented. That is to say, the programmer would give a command such as "Assemble radio," and the robot would need to develop its own step-by-step procedure for doing the task. The robot would have to possess a high level of intelligence for it to be programmed in this manner. Robots of the future also will necessarily make greater use of sensors for determining positions of objects in the work cell, for safety monitoring, and for error detection and recovery. Machine vision is expected to be an important sensor technology for coming generations of robots.

MANUFACTURING APPLICATIONS OF AUTOMATION AND ROBOTICS

One of the most important application areas for automation technology is manufacturing. To many people, automation means manufacturing automation. Three different types of automation in production can be distinguished: (1) fixed automation, (2) programmable automation, and (3) flexible automation.

Fixed automation

The term fixed automation refers to an automated production facility in which the sequence of processing operations is fixed by the equipment configuration. This is sometimes called "hard automation." The programmed commands are, in effect, contained in the machines in the form of cams, gears, wiring, and other hardware that is not easily changed over from one type of product to another. This form of automation is characterized by high initial investment and high production rates. It is therefore suitable for products that are made in large volumes. Examples of fixed automation include machining transfer lines found in the automotive industry, automatic assembly machines, and certain chemical processes.

Programmable automation

Programmable automation is a form of automation used in the production of batches (or quantities) of products. The products are made in batches ranging from several dozen to several thousand units at a time. For each different batch of product, the production equipment must be reprogrammed and converted to accommodate the new product configuration. This reprogramming and changeover (called the setup in many industries) take time to accomplish, and there is thus a nonproductive period followed by a production run for each new batch. Production rates in programmable automation are generally lower than in fixed automation because the equipment is designed to facilitate product changeover rather than for product specialization. A numerical control machine tool is a good example of programmable automation. The program is coded on punched paper tape for each different product style, and the machine tool actions are controlled by the punched tape. Industrial robots are another example of programmable automation.

Flexible automation

Flexible automation is an extension of programmable automation. The difficulty with programmable automation is the time required to reprogram and change over the production equipment for each new batch of product. This process takes time, and time is expensive. In flexible automation, the variety of products is sufficiently limited that the changeover of the physical setup of the equipment can be done quickly and automatically. The reprogramming of the equipment in flexible automation is done off-line (*i.e.,* the programming can be accomplished without using the production equipment itself). Accordingly, there is no need to group identical products into batches; instead, a mixture of different products can be produced one right after the other. Flexible automation is a relatively new concept in automation. Consequently, there are not many examples of this form of automation compared to the other two types. The economics of flexible automation, however, are advantageous enough that it is expected to become an important method of production in future automated factories.

Applications of automation can be found in nearly all types of production. Many of the important examples are reviewed here.

Machining. The shaping of metal by means of cutting tools was one of the first manufacturing processes to be mechanized and then automated. There are three examples of automation in the machining process that relate to the previous descriptions of the three types of production automation.

Transfer line

The first of these is the transfer line, which is used for machining metal parts in large volumes at high production rates. This is an example of fixed automation, because transfer lines are typically set up for long production runs, perhaps for making millions of parts and running for several years between changeovers. A transfer line is divided into a series of workstations, with each station designed to perform some specific machining operation on a part. The workstations are connected by a parts handling system that moves the parts from one operation to the next. The raw work part enters at one end of the transfer line, proceeds through each workstation, and emerges at the other end as a completed part. In the normal operation of the line, there is a work part being processed at each station, so that many parts are processed simultaneously and a finished part is produced with each cycle of the line. The origins of present-day transfer lines date back to the 1940s and earlier.

Numerical control

The second example of automation in machining is numerical control. The first NC machine tool was invented during the early 1950s. Numerical control is a form of programmable automation in which the machine is controlled by means of numbers (and other symbols) that have been coded on punched paper tape or some alternative storage medium. The program represents the set of machining instructions for a particular part; it is therefore called the NC part program. The coded numbers in the program (specifying x-y-z coordinates in a Cartesian axis system) indicate to the machine tool the various positions of the cutting tool relative to the work part. By means of a sequencing of these positions in the program, the machine tool is directed to accomplish the machining of the part. A position feedback control system is used in most machines to determine that the coded instructions have been correctly performed. Since the early 1980s, a microcomputer has typically been used as the controller in an NC machine tool, and the program is actuated from computer memory rather than from punched paper tape. Initial entry of the program into the computer memory is still accomplished by means of punched tape in many cases. This form of numerical control is called computer numerical control (CNC). Another variation in the imple-

mentation of numerical control involves the capability of sending the part program over telecommunications lines from a central computer to individual machine tools in a factory, thereby eliminating the use of punched tape altogether. This form of numerical control is known as direct numerical control (DNC).

The third example of automation in machining is the flexible manufacturing system (FMS). The FMS is a form of flexible automation in which several machine tools are linked together by a materials handling system, and all aspects of the system are controlled by a central computer. The materials handling system is capable of delivering parts to any machine in the FMS. Each machine is controlled by CNC, and a central computer sends programs to each controller according to a preplanned schedule. The FMS represents a high level of technological sophistication and a highly integrated form of production automation.

Chemical processing. Some of the most highly automated production facilities are found in the chemical-processing industries. These industries include petroleum refining, food processing, and other operations in which the products are processed in gas, liquid, or powdered form. Such forms facilitate the movement of products through the various steps in the production process. In addition, these products are usually made in large quantities. Because of the ease of handling the products and the large volumes involved, a high level of automation has been accomplished in these industries.

The typical modern process plant is computer-controlled. In one advanced petrochemical system that produces more than 20 different products, the facility is divided into three areas each with several chemical-processing units. Each of the three areas has its own process-control computer to perform scanning, control, and alarm functions. The three computers are connected to a central computer in a hierarchical configuration as illustrated in Figure 5. The central computer calculates how to obtain maximum yield from each process and generates management reports on process performance.

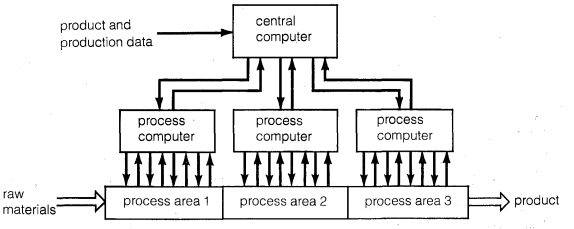


Figure 5: Hierarchical configuration typical of computer process control in a modern plant.

Each process computer monitors up to 2,000 process parameters, such as temperature, pressure, flow rate, liquid levels, chemical concentrations, and other variables required to control the process. These measurements are taken on a sampling basis; the time between samples varies between 2 and 120 seconds, depending on the relative need for the data. Each computer controls approximately 400 feedback control loops. Under normal operation, the control computers maintain operation of the process at or near optimum performance levels. If process parameters exceed the specified normal or safe ranges, the control computer actuates a signal light and alarm horn and prints a message for the technician indicating the nature of the problem.

The central computer receives data from the process computers and performs calculations to optimize the performance of each chemical-processing unit. The results of these calculations are then passed to the individual process computers in the form of commands to change the set points for the various control loops.

ıtegration f data ·om ıultiple ɔntrol ɔops

Substantial economic advantages are obtained from this type of computer control in the process industries. The computer hierarchy is capable of integrating all of the data from the many individual control loops far better than humans can, thus permitting a higher level of performance. Advanced control algorithms can be applied by the computer for optimal process control. In addition, much more quickly than humans, the computer is capable of sensing process conditions that indicate unsafe or abnormal operation. All of these improvements increase productivity, efficiency, and safety during process operation.

Basic metals industries. Like the chemical-processing industries, the basic metals industries (aluminum, iron and steel, etc.) have adopted automation for many of their processes. Similar to the chemical industries, the metals industries deal in large volumes of products, and so there is a substantial economic incentive to invest in automation. Metals, however, are typically produced in batches rather than continuously, and handling metals in bulk form is generally more difficult than handling chemicals that flow. Consequently, automation in the basic metals industries has not achieved as high a level as that in the chemical-processing industries.

One example of process automation in the metals industry is the rolling of hot metal ingots into their final shapes (*e.g.,* coils and strips). This was first done in the steel industry; similar processing is also accomplished in aluminum and other metals. In a modern steel plant, hot-rolling is accomplished with the aid of computer control. The rolling process involves the forming of a large, hot metal billet by passing it through a rolling mill consisting of one or more sets of large cylindrical rolls that squeeze the metal and reduce its cross section. In most cases, several passes are required to gradually reduce the ingot to the desired shape. Sensors and automatic instruments measure the dimensions and temperature of the ingot after each pass through the rolls, and the control computer calculates and regulates the roll settings for the next pass.

Control programs have been developed to schedule the sequence and rate at which the hot metal ingots are fed through the rolling mills. In a large plant several different orders for rolled products with different specifications may be in process at any given time. The production control task of scheduling and keeping track of the different customer orders in the mill requires rapid massive data gathering and analysis. In the most modern plants this production control task has been effectively integrated with the computer control of the rolling mill operations to achieve a highly automated production system.

Assembly. Assembly operations have traditionally been performed manually, either at single assembly workstations or on assembly lines with multiple workstations. Because of the high labour content and the high cost of manual labour, greater attention has been given in recent years to the use of automation and robotics for assembly work.

For high production work, automated assembly machines have been developed that operate in a manner similar to machining transfer lines. Instead of performing machining operations at the workstations, assembly operations are performed. A typical assembly machine consists of several stations, each equipped with a supply of components and a mechanism for delivering the components into position for assembly. A workhead at each station performs the actual adding and fastening of a component. Typical workheads include automatic screwdrivers, staking or riveting machines, welding heads, and other joining devices. A new component is added to the partially completed product at each workstation, thus building up the product gradually as it proceeds through the line. Assembly machines of this general type are considered to be examples of fixed automation because they are usually configured for a particular product that is made in large quantities.

Typical assembly machine

Since the late 1970s, there has been a growing effort to use robots in assembly operations. Because robots are programmable, the objective in using them for assembly work is to produce parts in medium quantities and to produce a mixture of different models on the same production line. An important research and development project in this general area was the Adaptable Programmable Assembly System (APAS) Project, conducted from 1980 to 1983 in the United States by the Westinghouse Electric Corporation under sponsorship of the National Science Foundation. The assembly of small electric motors was selected as the basis for the APAS development study. The assembly system for the project consisted of six workstations, four

of which used robots to perform the assembly operations. A conveyor system was utilized to move partially assembled motors between the stations. Control of the system was accomplished by means of a master supervisory computer, which communicated with local control computers at each workstation. The master computer scheduled and coordinated production and informed the local workstation computers of the tasks that had to be performed.

One of the significant lessons learned from the APAS Project was the importance of designing products for ease of automated assembly. The assembly methods that are satisfactory for humans are not necessarily the most suitable methods for automated workheads. For example, using a screw and nut as a fastening method is suitable for manual assembly but not for automated assembly. Designing the components to be added from the same direction through the use of snap fits and other one-step fastening procedures are much more easily accomplished by automated and robotic assembly methods.

Electronics manufacturing. A basic feature of the electronics industry is the need to achieve a high level of coordination in the design, fabrication, and testing of a product. A good example of the need for coordination is in large-scale integration (LSI) and very large-scale integration (VLSI), both of which involve the fabrication of electrical and electronic circuits on small semiconductor chips. LSI and VLSI products are extremely miniaturized and very complex. Computers must be used to perform much of the design analysis and synthesis and then to transfer the design specifications directly to the equipment that produces the chips. Also, electronic components such as LSI and VLSI chips are often produced in relatively low yields, meaning that the proportion of good units to total units made is significantly less than 100 percent. To separate the good units from those that are defective, testing of each unit must be done as part of the production sequence. This need for integration has motivated the introduction of automation into the electronics industry.

Design analysis and synthesis by means of computers

Two examples of automation in electronics manufacturing are part insertion machines and wire wrap machines. Part insertion machines are used to position electronic components (LSI chip modules and other devices) relative to a printed circuit board. It is basically an x = y positioning table that moves the printed circuit board relative to the part insertion head, which places the individual components into position on the board. Wire wrap machines are used to make wired connections between terminals on the back of an electrical panel automatically. The wiring head is positioned at the desired terminal pin, after which it wraps the wire around the pin, leads the unconnected end of the wire through a prescribed path to the second terminal pin, and cuts and wraps the wire around the second pin. An electrical test is then made to ensure that the proper terminals have been connected. Faulty connections are identified on a computer-generated report.

Part insertion and wire wrap machines both are production machines that utilize numerical control principles in which programmed commands are communicated to the machine from the design data base for the product. In the part insertion machine, the positioning of the table and the determination of the component to be inserted are contained in the numerical control program. In the wire wrap machine, such a program defines the connections that are to be made. Both the part insertion machine and the wire wrap machine are examples of programmable automation because they can be reprogrammed to accommodate different product configurations.

Application of numerical control principles

New inspection and testing equipment is being designed to perform a variety of checks automatically on electronic parts and products. Machine-vision systems are being developed to check the circuits and other quality features of microelectronic chips. These vision systems scan the chips to detect any irregularities from a given standard. Other automated test equipment is designed to perform electrical testing of the circuits more accurately than humans and in a fraction of the time that humans would require to make the tests.

Computer-aided design/computer-aided manufacturing (CAD/CAM). Since about 1970, there has been a growing trend in manufacturing firms toward the use of computers to perform many of the functions related to design and production. This trend is popularly known as CAD/CAM. The technology of CAD/CAM is based largely on the capability of a computer system to process, store, and display large amounts of data representing part and product specifications. For mechanical products, the data represent graphic models of the components; for electrical products, they represent circuit information; and so forth. CAD/CAM technology has been applied in many industries, including those involved in machined components, electronics products, and equipment design and fabrication for chemical processing. It reflects not only the automation of the manufacturing operations but the automation of elements in the entire design-and-manufacturing procedure as well.

Computer-aided design entails the use of computer systems to assist in the creation, modification, analysis, and optimization of a design. The designer, working with the CAD system rather than with the traditional drafting board, creates the lines and surfaces that form the designed object (product, part, structure, etc.) and stores this model in the computer data base. By invoking the appropriate CAD software, the designer can perform various analyses on the object, such as stress-strain analysis and heat-transfer calculations. By making adjustments in the design on the basis of these analyses, he develops the final design of the object. Once the design procedure has been completed, the computer-aided design system can be used to prepare the detailed drawings required to make the object.

Computer-aided manufacturing is defined as the use of computer systems to assist in the planning, control, and management of production operations. This can be accomplished by either a direct or an indirect connection between the computer system and the production operations. In the case of the direct connection, the computer is used to monitor or control the process in the factory. Computer process monitoring involves the use of the computer system to observe the manufacturing process, collect data on process performance, and report the performance to plant management. In this way, the plant can be operated more efficiently. Computer process control involves the use of the computer system to execute control actions to operate the plant automatically without significant human intervention. In most cases, process control includes observing the process (*i.e.*, taking feedback measurements) as part of the control function. Several applications of computer process control have been described above, as, for example, in the discussions of the chemical-processing and basic metals industries.

The indirect connections between the computer system and the process are those applications in which the computer is used to support the production operations without actually monitoring or controlling the operations. These applications typically involve planning and management functions that can be performed by the computer (or by humans working with the computer) more efficiently than by humans alone. Examples of these functions include planning of the step-by-step processes for the product, part programming in numerical control, and scheduling the production operations in the factory.

Robots in manufacturing. Today, robotics technology is limited almost exclusively to industrial applications, usually as direct contributors to the production operation. Industrial robot applications can be divided into three categories: (1) material transfer and machine loading/unloading, (2) processing operations, and (3) assembly and inspection.

Material transfer consists of applications in which robots are used to move materials or work parts from one location to another. Many of these tasks are relatively simple, requiring robots to pick parts from one conveyor and place them on another conveyor. Other transfer operations are more complex materials handling tasks, such as placing parts onto pallets in an arrangement that the robot must calculate. In machine-loading and machine-unloading operations, a robot tends a production machine in place of a human worker. Depending on the production operation, the robot either loads parts into the machine or unloads

Material transfer and machine loading/unloading operations

them from it. To perform the task, the robot has to be equipped with a special hand, called a gripper, to grasp a part. In most cases, the gripper must be designed for the particular geometrical shape of the part. An example of a robot load/unload application is an automated machining cell, in which the robot works with a numerically controlled machine tool. The robot loads raw work parts into the machine tool and unloads finished parts from it, and the machine performs the production operation under automatic cycle.

Robot processing operations

By contrast, processing operations require the robot to manipulate some tool to perform a process on the work part. Examples of these applications include spot welding, continuous arc welding, and spray painting. As of the mid-1980s, spot welding of automobile bodies was the most common application of industrial robots in the United States. In this operation, a robot is used to position a spot-welding apparatus against automobile panels and frames to complete the assembly of the basic car body. Arc welding is a continuous process in which the robot must move the welding rod along the seam to be welded. Spray painting involves the manipulation of a spray-painting gun over the surface of the object to be painted. Other processing operations that utilize robots include grinding, polishing, and routing processes in which a rotating spindle serves as the robot's tool.

Assembly and inspection applications

Robot assembly applications were discussed above with particular attention to APAS. The use of robots in assembly operations is expected to increase because of the high cost of manual labour in this area. Inspection is another area of factory operations in which the use of robots is growing. In a typical inspection application, the robot positions a sensor with respect to the work part and determines whether the part is consistent with the quality specifications.

In nearly all of these applications, the robot performs a repetitive operation as a substitute for human labour. Another characteristic of industrial robot applications is that the tasks are hazardous or unpleasant for human workers, as in the case of spray painting, spot welding, arc welding, and certain machine-loading/unloading tasks. Other situations where robots prove attractive as alternatives to humans are those in which the work part or tool is heavy and awkward to handle and in which robots can be used on two or three shifts. (M.P.G.)

NONMANUFACTURING APPLICATIONS OF AUTOMATION

In addition to the manufacturing applications of automation technology, there have been significant achievements in such areas as communications, transportation, the military, and the service industries. Some of the more significant are described in this section.

Telephone switching

Communications. One of the earliest practical applications of automation was in telephone switching. The first switching machines, invented near the end of the 19th century, were simple mechanical switches that were remotely controlled by the user pushing buttons or turning a dial on the telephone. The next generation of automatic switching equipment consisted of electromechanical control systems that were developed for public use around the 1920s and 1930s. They were made up of electromechanical relays and switches that performed functions such as monitoring thousands of telephone lines, determining which were demanding service, providing the dial tone, remembering the digits of each telephone number as it was being dialed, setting up the required connections, sending electrical signals to ring the receiver's unit, monitoring the call during its progress, and disconnecting the telephone when the call was completed. These systems also were used to time and bill toll calls and to transmit billing information and other data relative to the business operations of the telephone company. The introduction of these electromechanical systems into the nation's telephone system required many years to complete, but they gradually assumed nearly all of the functions of the human telephone operator.

Modern electronic telephone switching machines are based on highly sophisticated digital computers that perform all of the functions of their electromechanical predecessors with much greater speed, reliability, functionality, and economy. In addition to the functions mentioned above, the newest electronic systems automatically transfer calls to alternate numbers, call the user back when a busy line becomes free, and perform other customer services in response to dialed codes. These systems also perform function tests on their own operations, diagnose problems when they arise, and print out detailed instructions for maintenance personnel.

Other applications of automation in communications systems include local area networks, communications satellites, and automated mail-sorting machines. A local area network (LAN) operates like an automated telephone company within a single building or group of buildings. Such networks are generally capable of carrying not only voice messages but also large quantities of digital data between terminals in the system. Communications satellites have become essential for relaying telephone or video signals across great distances. These satellite communications would not be possible without the automated guidance systems that place and maintain the satellites in a predetermined orbit around the Earth. Finally, automatic mail-sorting machines are being used or installed in post offices throughout the United States to read ZIP codes on envelopes and sort mail according to destination.

Transportation. Automation has been applied in various ways in the transportation industries. Major applications include airline reservation systems, automatic pilots in aircraft and locomotives, and automated transit systems.

Airlines utilize automated reservation systems to continuously monitor the status of flights. With these systems, ticket agents at widely separated locations can obtain information about the availability of seats on any flight in a matter of seconds. The reservation systems compare requests for space with the status of each flight, grant space when available, and automatically update the reservation status files. With contemporary systems passengers can secure their seat assignments well in advance of flight time.

Automatic pilots

Nearly all commercial aircraft used by the airlines are equipped with instrument systems called automatic pilots. Under normal flying conditions, these systems can guide an airplane over a predetermined route by detecting changes in the aircraft's orientation and heading from gyroscopes and similar instruments and by providing appropriate control signals to the craft's steering mechanism. Automatic navigation systems and instrument landing systems operate by using radio signals from ground beacons that provide the aircraft with course directions for guidance. When an airplane is within the traffic pattern for ground control, its human pilot normally assumes control.

The BART system

The Bay Area Rapid Transit (BART) system in the San Francisco–Oakland area of California represents an ambitious undertaking in automated rail transportation. It has been a model for rapid transit railways in Washington, D.C., Atlanta, Ga., and various other U.S. cities. The BART system consists of more than 75 miles (120 kilometres) of track, with about 100 trains operating at peak hours between roughly 30 stations. The trains sometimes attain speeds of 80 miles (128 kilometres) per hour with intervals between trains as short as 90 seconds. Each train carries one operator whose role is that of an observer and communicator capable of overriding the automatic system in case of an emergency. The automatic system protects the trains by assuring a safe distance between them and by controlling their speed. Another function of the system is to control train routings and make adjustments in the operation of each train to keep the entire system operating on schedule.

As a BART train enters the station, it automatically transmits its identification and destination, which are flashed on a display board for passengers. Such information is also transmitted to the control centres, from which signals are automatically returned to the train to regulate its time in the station and its running time to the next station. At the beginning of the day an ideal schedule is determined. As the day progresses the performance of each train is compared with the schedule, and adjustments are made to each train's performance as required. If an unexpected situation such as a train breakdown arises, the system can automatically adjust itself to minimize the effect. The

entire system is controlled by two identical computers so that if one malfunctions the other assumes control. In the event of a complete failure of the computer control system, the aboard train operators resort to manual control.

Military applications. The possibility of using robotics and other advanced automation technologies in military operations has much appeal. As yet, few actual applications have been developed. Many of the routine and sometimes dangerous tasks conducted during land and naval operations and the logistical support of these operations, however, could be performed by sophisticated robots of the future. Some examples include driving trucks in a convoy following a lead vehicle operated by a human, refueling tanks and other vehicles on the battlefield, loading artillery, and working in the engine room on board ship. The possibility of sending soldier robots on a suicide mission deep inside enemy territory is surely a prospect that would interest any military strategist.

Service industries. Automation of the service industries encompasses an assortment of applications as diverse as the services themselves. The services include health care, banking and other financial services, government, and retail trade.

In health care the use of computer systems is increasing dramatically and is helping to improve services and relieve the burden on medical staffs. In hospitals computer terminals on each nursing care floor are used to record data on patient status, medications administered, and other relevant information. Besides providing an official record of the nursing care given to individual patients, the computer system facilitates the nurse's task of updating reports at the time of shift changes. Moreover, the system is connected to the hospital's business office so that proper charges can be made to each patient's account for services rendered and medicines provided.

Banks and other financial institutions have embraced automation in their operations because of the large volume of documents and data that have to be processed in their daily business. The sorting of checks and verification of account balances have been computerized by virtually all banks and savings and loans. An increasing number of institutions have gone a step further by establishing systems of electronic banking, including the use of the automatic teller machine (ATM). Located in shopping centres, business buildings, and other convenient places, ATM's permit customers to carry out basic transactions without the assistance of bank personnel.

Electronic banking

Credit card transactions also have become highly automated. Many restaurants and retailers employ systems that automatically check the validity of a credit card and the credit standing of the cardholder as the latter waits for the transaction to be finalized. It is anticipated that future credit card transactions may involve an immediate transfer of funds from the cardholder's account into the merchant's account via high-speed communications lines linking computerized point-of-sale terminals and banks.

Increased reliance on computers and computerized data bases have highly automated the operations of government services. The Internal Revenue Service (IRS) of the U.S. government must review and approve the tax returns of millions of taxpayers each year. The detailed checking of returns—a process known as auditing—is a task that has traditionally been done on a sampled basis by large staffs of professional auditors. In 1985 the IRS began using a computerized system to automate the auditing procedure for the 1984 returns. This system is programmed to perform complex tax calculations on each return being audited. As tax laws change, it can be reprogrammed so as to audit new returns in accordance with the revisions. The IRS expects that the computer-automated auditing system will substantially increase the work capacity of its auditing department without a corresponding increase in manpower.

The retail trade has seen a number of changes in its operations as a result of automation. Selling merchandise has typically been a labour intensive activity, in which salesclerks need to assist customers with their selections and then finalize transactions at the cash register. Each transaction depletes the inventory of the store, and so the item purchased must be identified for reorder. Computer-automated systems have been installed in most large department stores and supermarkets to speed up sales transactions and to provide efficient inventory control. Such a system features a laser light pen or similar sensing device designed to read an identification symbol consisting of a series of bars printed on or affixed to each product. By scanning the symbol with the optical reading unit at the register or checkout counter, the salesclerk quickly identifies the item being sold, records its price into the total of the sale, and enters the transaction into the inventory files of the store. The store's central computer automatically updates these records, subtracting the item sold from the total number of items of the same kind and brand still in stock.

Automated checkout and inventory control

Consumer products. A wide variety of consumer products have been automated to enhance performance and user convenience. Microwave ovens, washing machines, dryers, compact disc players, videocassette recorders, and a number of other modern home appliances are equipped with a microprocessor that serves as the computer controller for the machine. By simply pressing a series of buttons in proper sequence, the user can program the operation of the appliance. Many videocassette recorders, for instance, can be programmed to automatically tape several television programs transmitted on different channels at different times during the course of a week or longer.

The automobile is another example of a highly automated consumer product. Most recent model cars come equipped with several microprocessors that regulate the operation of the engine (fuel-air ratio and other functions) and of the clock, radio, and automatic speed control (popularly known as cruise control). Special options on some models include the capability to monitor sensors in the car to alert the driver to specific problems (*e.g.,* low fuel, door ajar, and engine temperature) and the ability to compute such information as average gasoline mileage and driving range on remaining fuel supply. One automotive manufacturer has pioneered the use of an electronic voice system, in which the car communicates problems to the driver by means of simple verbal messages delivered through its radio speakers. (M.P.G./M.T.)

AUTOMATION AND SOCIETY

Over the years, the social merits of automation have been argued by labour leaders, government officials, business executives, and college professors. No doubt the biggest controversy has focused on the employment issue: What is the effect of automation on employment? There are other important aspects of the automation issue as well, including its effect on productivity, economic competition, education, and quality of life. These social issues will be explored here.

Impact on the individual. Nearly all industrial installations of automation, robotics in particular, involve a replacement of human labour by an automated system. Therefore, one of the direct effects of automation in factory operations is the dislocation of human labour from the workplace. The long-term effects of automation on employment and unemployment rates are debatable. Most studies in this area have been controversial and inconclusive. Workers have indeed been lost by automation, but population increases and consumer demand for the products of automation have compensated for these losses. Labour unions have argued, and many companies have adopted the policy, that workers who are displaced by automation should be retrained for other positions, perhaps increasing their skill levels in the process. This argument succeeds so long as the company and the economy in general are growing at a fast enough rate to create new jobs as the jobs replaced by automation are lost.

Of particular concern for many labour specialists is the impact of industrial robots on the work force, since robot installations involve a direct substitution of machines for humans at a ratio of from two to three humans per robot. During the first half of the 1980s, however, the effect of robotics on labour was relatively minor in the United States at least, because the number of robots in the nation's factories was small compared to the number of human

workers. By the end of 1984, only about 10,000 robots had been installed, while the total work force stood at more than 100,000,000 persons, approximately 19,000,000 of whom worked in factories. This would indicate that the impact of robots on unemployment has been relatively modest to date.

Automation affects not only the number of workers in factories but also the type of work that is done. An automated factory is oriented toward the use of computer systems and sophisticated programmable machines rather than manual labour. Greater emphasis is placed on knowledge-based work and technical skill than on physical work. The types of jobs that must be done in modern factories include more machine maintenance, improved scheduling and process optimization, systems analysis, and computer programming and operation. Thus workers in automated facilities must be technologically proficient to perform such jobs. Professional and semiprofessional positions, as well as traditional labour jobs, are affected by this shift in emphasis toward factory automation.

Impact on society. Besides affecting the individual worker, automation has an impact on society in general. Productivity is a fundamental social and economic issue that is influenced by automation. The productivity of a process is traditionally defined as the ratio of output units to the units of labour input. A properly justified automation installation will provide an increase in productivity due to increases in production rate and reductions in labour content. Over the years productivity gains have led to reduced prices for products and increased prosperity for society.

ncreased
roduc-
ivity

A number of issues related to education and training have been raised by the increased use of automation, robotics, computer systems, and related technologies. As automation has increased, there has developed a shortage of technically trained personnel to implement these technologies competently. This shortage has had a direct influence on the rate at which automated systems can be introduced. The shortage of skilled staffing in automation technologies increases the need for vocational and technical training to develop the required work force skills. Unfortunately the educational system is also in need of technically qualified instructors to teach these subjects. The laboratory equipment available in schools, moreover, all too often does not represent the state-of-the-art technology typically used in industry.

Advantages and disadvantages of automation. The advantages commonly attributed to automation include increased production rates, more efficient use of materials, better product quality, improved safety, shorter workweeks for labour, and reduction of factory lead times.

Automated machines are usually designed to operate at higher production rates than humans are capable of achieving. This increased productivity has been one of the biggest reasons for justifying the adoption of automated systems. Notwithstanding the claims of high quality from good workmanship by humans, automated systems are generally capable of carrying out the manufacturing processs with less variability than humans, thus yielding greater control and consistency of product quality. In addition, the increased process control makes possible more efficient use of materials, resulting in less waste.

Automated manufacturing systems often remove workers from the workplace, thus safeguarding them against hazards in the work environment. In the United States, the Occupational Safety and Health Act of 1970 (OSHA) was enacted with the objective of making work safer and protecting the physical well-being of the worker on a national scale. OSHA has had the effect of promoting the use of automation and robotics in the factory.

lenefits for
he factory
/orker

Another of the benefits of automation noted above is the reduction in the number of hours worked on average per week by factory workers. Around the turn of the century, the average workweek was about 70 hours. This has gradually been reduced, so that today the standard workweek in the United States is about 40 hours. Mechanization and automation have played a significant role in this reduction. Finally, the time required to process a typical job through the factory is generally reduced with automation. Referred to as the manufacturing lead time, this is the amount of time between the beginning of a project or process and the appearance of its results.

Among the major disadvantages associated with automation is worker dislocation. This problem was touched upon earlier. Whatever the ultimate social benefits that might result from retraining displaced workers for other (perhaps more skilled) jobs, in almost all cases the worker whose job is taken over by a machine suffers considerable personal stress. In addition to displacement from work, the worker may be displaced geographically. To find other work, an individual may have to relocate, which is itself another source of emotional stress.

Another significant disadvantage associated with automation is high capital expenditure. Automated systems can cost millions of dollars to design, fabricate, and put into operation. Still other problems characteristic of automated equipment include a higher level of maintenance than is required by manually operated hardware and an inherent lack of flexibility in terms of the variety of products that can be produced.

Also, automation is not without its potential dangers. As often suggested in works of science fiction, the technology associated with automation might ultimately subjugate rather than serve humankind. There is a possibility that workers will become slaves to automated machines, that the privacy of humans will be invaded by vast computer data banks, that human error in the management of technology will somehow endanger the health of civilization, and that society will become completely dependent on automation for its economic well-being.

These dangers notwithstanding, there are substantial opportunities that may arise from the wise and effective use of automation technology. There is an opportunity to relieve humans from boring, repetitive, hazardous, and unpleasant labour in all forms. Then, too, there is an opportunity for future automation technologies to provide a growing social and economic environment in which humans can enjoy a higher standard of living and a better way of life.

BIBLIOGRAPHY. Works on the technology of automation and its applications include MIKELL P. GROOVER, *Automation, Production Systems, and Computer-Aided Manufacturing* (1980), an informative survey; N. CAPTOR *et al., Adaptable-Programmable Assembly Research Technology Transfer to Industry: Phase 2* (1982), a final report on technological innovations in manufacturing processes; MIKELL P. GROOVER and EMORY W. ZIMMERS, JR., *CAD/CAM: Computer-Aided Design and Manufacturing* (1984), a comprehensive reference source; and "Automation U.S.A.," *High Technology* 5(5):24–47 (May 1985), a series of articles on factory automation.

Robotics technology and its applications receive focal attention in the following: MARVIN MINSKY (ed.), *Robotics* (1985); V. DANIEL HUNT, *Smart Robots: A Handbook of Intelligent Robotic Systems* (1985); and MIKELL P. GROOVER, M. WEISS, R.N. NAGEL, and N.G. ODREY, *Industrial Robotics: Technology, Programming, and Applications* (1986). A general introduction to robotics applications can be found in ROBERT U. AYERS and STEVEN M. MILLER, *Robotics, Applications and Social Implications* (1983). For more technical material, see JOSEPH F. ENGELBERGER, *Robotics in Practice: Management and Applications of Industrial Robots* (1980). Precise descriptions of robotics applications and a glossary of robotics terminology can be found in DAVID F. TVER and ROGER W. BOLZ, *Robotics Sourcebook and Dictionary* (1983). JOHN HARTLEY, *Flexible Automation in Japan* (1984), is a collection of articles describing Japanese applications of robotics and flexible manufacturing systems. ALAN PUGH (ed.), *Robot Vision* (1983), brings together research papers on this aspect of manufacturing technology.

The impact of automation and robotics technology on the individual and society are discussed in MIKELL P. GROOVER, JOHN E. HUGHES, JR., and NICHOLAS G. ODREY, "The Societal Impact of Factory Automation," *Industrial Engineering* 16(4):50–59 (April 1984); HARLEY SHAIKEN, *Work Transformed: Automation and Labor in the Computer Age* (1985); and ROBERT J. MILLER (ed.), *Robotics: Future Factories, Future Workers* (1983), which also includes the impact on public policy. See also *Exploratory Workshop on the Social Impacts of Robotics: Summary and Issues, a Background Paper* (1982), and *Computerized Manufacturing Automation: Employment, Education, and the Workplace* (1984), with *Working Papers,* 2 vol., published by the Office of Technology Assessment.

(M.P.G.)

Bach

Although he was admired by his contemporaries primarily as an outstanding harpsichordist, organist, and expert on organ building, Johann Sebastian Bach is now generally regarded as one of the greatest composers of all time and is celebrated as the creator of the *Brandenburg Concertos, The Well-Tempered Clavier,* the *Mass in B Minor,* and numerous other masterpieces of church and instrumental music. Appearing at a propitious moment in the history of music, Bach was able to survey and bring together the principal styles, forms, and national traditions that had developed during preceding generations and, by virtue of his synthesis, enrich them all.

By courtesy of the Osterreichische Nationalbibliothek, Vienna

Bach, lithograph by Rudolf Hoffmann.

He was a member of a remarkable family of musicians who were proud of their achievements, and about 1735 he drafted a genealogy, *Ursprung der musicalisch-Bachischen Familie* ("Origin of the Musical Bach Family"), in which he traced his ancestry back to his great-great-grandfather Veit Bach, a Lutheran baker (or miller), who was driven from Hungary to Wechmar in Thuringia, a historic region of Germany, by religious persecution late in the 16th century and died in 1619. There were Bachs in the area before that, and it may be that, when Veit moved to Wechmar, he was returning to his birthplace. He used to take his cittern to the mill and play it while grinding was going on. Johann Sebastian remarked, "A pretty noise they must have made together! However, he learnt to keep time, and this apparently was the beginning of music in our family."

Until the birth of Johann Sebastian, his was the least distinguished branch of the family; its members had been competent practical musicians, but not composers, such as Johann Christoph and Johann Ludwig. In later days the most important musicians in the family were Johann Sebastian's sons, Wilhelm Friedemann, Carl Philipp Emanuel, and Johann Christian (the "English Bach").

LIFE

Early years. J.S. Bach was born at Eisenach, Thuringia (now in East Germany), on March 21, 1685, the youngest child of Johann Ambrosius Bach and Elisabeth Lämmerhirt. Ambrosius was a string player, employed by the town council and the ducal court of Eisenach. Johann Sebastian started school in 1692 or 1693 and did well in spite of frequent absences. Of his musical education at this time, nothing definite is known; but he may have picked up the rudiments of string playing from his father, and no doubt he attended the Georgenkirche, where Johann Christoph Bach was organist until 1703.

First keyboard lessons

By 1695 both his parents were dead, and he was looked after by his eldest brother, also named Johann Christoph (1671–1721), organist at Ohrdruf. This Christoph had been a pupil of the influential keyboard composer Johann Pachelbel, and he apparently gave Johann Sebastian his first formal keyboard lessons. The young Bach again did well at school, until in 1700 his voice secured him a place in a select choir of poor boys at the school at the Michaelskirche, Lüneburg (now in West Germany).

His voice must have broken soon after this, but he remained at Lüneburg for a time, making himself generally useful. No doubt he studied in the school library, which had a large and up-to-date collection of church music; he probably heard Georg Böhm, organist of the Johanneskirche; and he visited Hamburg to hear the renowned organist and composer Johann Adam Reinken at the Katharinenkirche, contriving also to hear the French orchestra maintained by the Duke of Celle.

He seems to have returned to Thuringia in the late summer of 1702. By this time he was already a reasonably proficient organist. His experience at Lüneburg, if not at Ohrdruf, had turned him away from the secular string-playing tradition of his immediate ancestors; thenceforth, he was, chiefly, though not exclusively, a composer and performer of keyboard and sacred music. The next few months are wrapped in mystery, but, by March 4, 1703, he was a member of the orchestra employed by Johann Ernst, Herzog (Duke) von Weimar (brother of Wilhelm Ernst, whose service Bach entered in 1708). This post was a mere stopgap; he probably already had his eye on the organ then being built at the Neukirche in Arnstadt; for, when it was finished, he helped to test it, and in August 1703 he was appointed organist—all this at the age of 18. Arnstadt documents imply that he had been court organist at Weimar; this is incredible, though it is likely enough that he had occasionally played there.

The Arnstadt period. At Arnstadt, on the northern edge of the Thuringian forest, where he remained until 1707, Bach devoted himself to keyboard music, the organ in particular. While at Lüneburg, he had apparently had no opportunity of becoming directly acquainted with the spectacular, flamboyant playing and compositions of Dietrich Buxtehude, the most significant exponent of the north German school of organ music. In October 1705 he repaired this gap in his knowledge by obtaining a month's leave and walking to Lübeck (more than 200 miles [300 kilometres]). His visit must have been profitable, for he did not return until about the middle of January 1706. In February his employers complained about his absence and about other things as well: he had harmonized the hymn tunes so freely that the congregation could not sing to his accompaniment, and, above all, he had produced no cantatas. Perhaps the real reasons for his neglect were that he was temporarily obsessed with the organ and was on bad terms with the local singers and instrumentalists, who were not under his control and did not come up to his standards. In the summer of 1705 he had made some offensive remark about a bassoon player, which led to an unseemly scuffle in the street. His replies to these complaints were neither satisfactory nor even accommodating; and the fact that he was not dismissed out of hand suggests that his employers were as well aware of his exceptional ability as he was himself and were reluctant to lose him.

Acquaintance with Buxtehude's music

During these early years Bach inherited the musical culture of the Thuringian area, a thorough familiarity with the traditional forms and hymns (chorales) of the orthodox Lutheran service, and, in keyboard music, perhaps (through his brother, Johann Christoph) a bias toward the formalistic styles of the south. But he also learned eagerly from the northern rhapsodists, Buxtehude above all. By 1708 he had probably learned all that his German pre-

decessors could teach him and arrived at a first synthesis of northern and southern German styles. He had also studied, on his own and during his presumed excursions to Celle, some French organ and instrumental music.

Among the few works that can be ascribed to these early years with anything more than a show of plausibility are the *Capriccio sopra la lontananza del suo fratello dilettissimo* (*Capriccio on the Departure of His Most Beloved Brother,* 1704, BWV 992); the chorale prelude on *Wie schön leuchtet* (*How Brightly Shines, c.* 1705, BWV 739); the fragmentary early version of the organ *Prelude and Fugue in G Minor* (before 1707, BWV 535a). (The "BWV" numbers provided are the standard catalog numbers of Bach's works as established in the *Bach-Werke-Verzeichnis,* prepared by the German musicologist Wolfgang Schmieder.)

The Mühlhausen period. In June 1707 Bach obtained a post at the Blasiuskirche in Mühlhausen in Thuringia. He moved there soon after and married his cousin Maria Barbara Bach at Dornheim on October 17. At Mühlhausen things seem, for a time, to have gone more smoothly. He produced several church cantatas at this time; all of these works are cast in a conservative mold, based on biblical and chorale texts and displaying no influence of the "modern" Italian operatic forms that were to appear in Bach's later cantatas. The famous organ *Toccata and Fugue in D Minor* (BWV 565), written in the rhapsodic northern style, and the *Prelude and Fugue in D Major* (BWV 532) may also have been composed during the Mühlhausen period, as well as the organ *Passacaglia in C Minor* (BWV 582), an early example of Bach's instinct for large-scale organization. Cantata No. 71, *Gott ist mein König* (*God is My King*), of February 4, 1708, was printed at the expense of the city council and was the first of Bach's compositions to be published. While at Mühlhausen, Bach copied music to enlarge the choir library, tried to encourage music in the surrounding villages, and was in sufficient favour to be able to interest his employers in a scheme for rebuilding the organ (February 1708). His real reason for resigning on June 25, 1708, is not known. He himself said that his plans for a "well-regulated [concerted] church music" had been hindered by conditions in Mühlhausen and that his salary was inadequate. It is generally supposed that he had become involved in a theological controversy between his own pastor Frohne and Archdeacon Eilmar of the Marienkirche. Certainly, he was friendly with Eilmar, who provided him with librettos and became godfather to Bach's first child; and it is likely enough that he was not in sympathy with Frohne, who, as a Pietist, would have frowned on elaborate church music. It is just as possible, however, that it was the dismal state of musical life in Mühlhausen that prompted Bach to seek employment elsewhere. At all events, his resignation was accepted, and shortly afterward he moved to Weimar, some miles west of Jena on the Ilm River. He continued, nevertheless, to be on good terms with Mühlhausen personalities, for he supervised the rebuilding of the organ, is supposed to have inaugurated it on October 31, 1709, and composed a cantata for February 4, 1709, which was printed but has disappeared.

Cantatas of the Mühlhausen period

The Weimar period. Bach was, from the outset, court organist at Weimar and a member of the orchestra. Encouraged by Wilhelm Ernst, he concentrated on the organ during the first few years of his tenure. From Weimar, Bach occasionally visited Weissenfels; in February 1713 he took part in a court celebration there that included a performance of his first secular cantata, *Was mir behagt,* or the *Hunt Cantata* (BWV 208).

Late in 1713 Bach had the opportunity of succeeding Friedrich Wilhelm Zachow at the Liebfrauenkirche, Halle; but the Herzog raised his salary, and he stayed on at Weimar. On March 2, 1714, he became concertmaster, with the duty of composing a cantata every month. He became friendly with a relative, Johann Gottfried Walther, a music lexicographer and composer who was organist of the town church, and, like Walther, Bach took part in the musical activities at the Gelbes Schloss (Yellow Castle), then occupied by Herzog Wilhelm's two nephews, Ernst August and Johann Ernst, both of whom he taught. The latter was a talented composer who wrote concertos in the Italian manner, some of which Bach arranged for keyboard instruments; the boy died in 1715, in his 19th year.

Unfortunately, Bach's development cannot be traced in detail during the vital years 1708–14, when his style underwent a profound change. There are too few datable works. From the series of cantatas written in 1714–16, however, it is obvious that he had been decisively influenced by the new styles and forms of the contemporary Italian opera and by the innovations of such Italian concerto composers as Antonio Vivaldi. The results of this encounter can be seen in such cantatas as numbers 182, 199, and 61 in 1714; 31 and 161 in 1715; and 70 and 147 in 1716. His favourite forms appropriated from the Italians were those based on refrain (ritornello) or da capo schemes in which wholesale repetition—literal or with modifications—of entire sections of a piece permitted him to create coherent musical forms with much larger dimensions than had hitherto been possible. These newly acquired techniques henceforth governed a host of Bach's arias and concerto movements, as well as many of his larger fugues (especially the mature ones for organ) and profoundly affected his treatment of chorales.

Italian influence on Bach

Among other works almost certainly composed at Weimar are most of the *Orgelbüchlein* (*Little Organ Book*); all but the last of the so-called 18 "Great" chorale preludes; the earliest organ trios; and most of the organ preludes and fugues. The "Great" *Prelude and Fugue in G Major* for organ (BWV 541) was finally revised about 1715, and the *Toccata and Fugue in F Major* (BWV 540) may have been played at Weissenfels.

On December 1, 1716, Johann Samuel Drese, musical director at Weimar, died. He was then succeeded by his son, who was rather a nonentity. Bach presumably resented being thus passed over; and in due course he accepted an appointment as musical director to Prince Leopold of Köthen, which was confirmed in August 1717. Herzog Wilhelm, however, refused to accept his resignation—partly, perhaps, because of Bach's friendship with the Herzog's nephews, with whom the Herzog was on the worst of terms. About September a contest between Bach and the famous French organist Louis Marchand was arranged at Dresden. The exact circumstances are not known; but Marchand avoided the contest by leaving Dresden a few hours before it should have taken place. By implication, Bach won. Perhaps this emboldened him to renew his request for permission to leave Weimar; at all events he did so but in such terms that the Herzog imprisoned him for a month (November 6–December 2). A few days after his release, Bach moved to Köthen (in modern East Germany), some 30 miles (50 kilometres) north of Halle.

The Köthen period. There, as musical director, he was concerned chiefly with chamber and orchestral music. Even though some of the works may have been composed earlier and revised later, it was at Köthen that the sonatas for violin and clavier, viola da gamba and clavier, and the works for unaccompanied violin and cello were put into something like their present form. The *Brandenburg Concertos* were finished by March 24, 1721; in the sixth concerto—so it has been suggested—Bach bore in mind the technical limitations of the Prince, who played the gamba. Bach played the viola by choice; he liked to be "in the middle of the harmony." He also wrote a few cantatas for the Prince's birthday and other such occasions; most of these seem to have survived only in later versions, adapted to more generally useful words. And he found time to compile pedagogical keyboard works: the *Clavierbüchlein* for W.F. Bach (begun January 22, 1720); some of the *French Suites;* the *Inventions* (1720); and the first book (1722) of *Das Wohltemperierte Klavier* (*The Well-Tempered Clavier,* eventually consisting of two books, each of 24 preludes and fugues in all keys and known as the Forty-eight). This remarkable collection systematically explores both the potentials of a newly established tuning procedure—which, for the first time in the history of keyboard music, made all the keys equally usable—and the possibilities for musical organization afforded by the system of "functional tonality," a kind of musical syntax consolidated in the music of the Italian concerto composers of the preceding generation and a system that

Influence of *The Well-Tempered Clavier* on keyboard music

was to prevail for the next 200 years. At the same time *The Well-Tempered Clavier* is a compendium of the most popular forms and styles of the era: dance types, arias, motets, concertos, etc., presented within the unified aspect of a single compositional technique: the rigorously logical and venerable fugue.

Maria Barbara Bach died unexpectedly and was buried on July 7, 1720. About November, Bach visited Hamburg; his wife's death may have unsettled him and led him to inquire after a vacant post at the Jacobikirche. Nothing came of this, but he played at the Katharinen-kirche in the presence of Reinken. After hearing Bach improvise variations on a chorale tune, the old man said, "I thought this art was dead; but I see it still lives in you."

On December 3, 1721, Bach married Anna Magdalena Wilcken, daughter of a trumpeter at Weissenfels. Apart from his first wife's death, these first four years at Köthen were probably the happiest of Bach's life. He was on the best terms with the Prince, who was genuinely musical; and in 1730 Bach said that he had expected to end his days there. But the Prince married on December 11, 1721, and conditions deteriorated. The Princess—described by Bach as "an *amusa*" (that is to say, opposed to the muses)—required so much of her husband's attention that Bach began to feel neglected. He also had to think of the education of his elder sons, born in 1710 and 1714, and he probably began to think of moving to Leipzig as soon as the cantorate fell vacant with the death of Johann Kuhnau on June 5, 1722. Bach applied in December, but the post—already turned down by Bach's friend, Georg Philipp Telemann—was offered to another prominent composer of the day, Christoph Graupner, the musical director at Darmstadt. As the latter was not sure that he would be able to accept, Bach gave a trial performance (Cantata No. 22, *Jesu nahm zu sich die Zwölfe* [*Jesus called unto Him the Twelve*]) on February 7, 1723; and, when Graupner withdrew (April 9), Bach was so deeply committed to Leipzig that, although the Princess had died on April 4, he applied for permission to leave Köthen. This he obtained on April 13, and on May 13 he was sworn in at Leipzig.

He was appointed honorary musical director to Köthen, and both he and Anna were employed there from time to time until the Prince died, on November 19, 1728.

Years at Leipzig. As Director of Church Music for the city of Leipzig, Bach had to supply performers for four churches. At the Peterskirche the choir merely led the hymns. At the Neukirche, Nikolaikirche, and Thomaskirche, part singing was required; but Bach himself conducted, and his own church music was performed, only at the last two. His first official performance was on May 30, 1723, the first Sunday after Trinity Sunday, with Cantata No. 75, *Die Elenden sollen essen.* New works produced during this year include many cantatas and the *Magnificat* in its first version. The first half of 1724 saw the production of the *St. John Passion,* which was subsequently revised. The total number of cantatas produced during this ecclesiastical year was about 62, of which about 39 were new works.

Production of the *St. John Passion*

On June 11, 1724, the first Sunday after Trinity, Bach began a fresh annual cycle of cantatas, and within the year he wrote 52 of the so-called chorale cantatas, formerly supposed to have been composed over the nine-year period 1735–44. The Sanctus of the *Mass in B Minor* was produced at Christmas.

During his first two or three years at Leipzig, Bach had produced a large number of new cantatas, sometimes, as recent research has revealed, at the rate of one a week. This phenomenal pace raises the question of Bach's approach to composition. Bach and his contemporaries, subject to the hectic pace of production, had to invent or discover their ideas quickly and could not rely on the unpredictable arrival of "inspiration." Nor did the musical conventions and techniques or the generally rationalistic outlook of the time necessitate this reliance, as long as the composer was willing to accept them. The Baroque composer who submitted to the regimen inevitably had to be a traditionalist who willingly embraced the conventions.

Symbolism. A repertory of melody types existed, for example, that was generated by an explicit "doctrine of figures" that created musical equivalents for the figures of speech in the art of rhetoric. Closely related to these "figures" are such examples of pictorial symbolism in which the composer writes, say, a rising scale to match words that speak of rising from the dead or a descending chromatic scale (depicting a howl of pain) to sorrowful words. Pictorial symbolism of this kind occurs only in connection with words—in vocal music and in chorale preludes, where the words of the chorale are in the listener's mind. There is no point in looking for resurrection motifs in *The Well-Tempered Clavier.* Pictorialism, even when not codified into a doctrine, seems to be a fundamental musical instinct and essentially an expressive device. It can, however, become more abstract, as in the case of number symbolism, a phenomenon observed too often in the works of Bach to be dismissed out of hand. Number symbolism is sometimes pictorial; in the *St. Matthew Passion* it is reasonable that the question "Lord, is it I?" should be asked 11 times, once by each of the faithful disciples. But the deliberate search for such symbolism in Bach's music can be taken too far. Almost any number may be called "symbolic" (3, 6, 7, 10, 11, 12, 14, and 41 are only a few examples); any multiple of such a number is itself symbolic; and the number of sharps in a key signature, notes in a melody, measures in a piece, and so on may all be considered significant. As a result, it is easy to find symbolic numbers anywhere, but ridiculous to suppose that such discoveries invariably have a meaning.

Use of pictorial symbolism

Besides the melody types, the Baroque composer also had at his disposal similar stereotypes regarding the further elaboration of these themes into complete compositions, so that the arias and choruses of a cantata almost seem to have been spun out "automatically." One is reminded of Bach's delightfully innocent remark "I have had to work hard; anyone who works just as hard will get just as far," with its implication that everything in the "craft" of music is teachable and learnable. The fact that no other composer of the period, with the arguable exception of Handel, even remotely approached Bach's achievement indicates clearly enough that the application of the "mechanical" procedures was not literally "automatic" but was controlled throughout by something else—artistic discrimination, or taste. "Taste," a most respected attribute in the culture of the 18th century, is an utterly individual compound of raw talent, imagination, psychological disposition, judgment, skill, and experience. It is unteachable and unlearnable.

As a result of his intense activity in cantata production during his first three years in Leipzig, Bach had created a supply of church music to meet his future needs for the regular Sunday and feast-day services. After 1726, therefore, he turned his attention to other projects. He did, however, produce the *St. Matthew Passion* in 1729, a work that inaugurated a renewed interest in the mid-1730s for vocal works on a larger scale than the cantata: the now-lost *St. Mark Passion* (1731), the *Christmas Oratorio,* BWV 248 (1734), and the *Ascension Oratorio* (Cantata No. 11, *Lobet Gott in seinen Reichen;* 1735).

St. Matthew Passion* and *Christmas Oratorio

Nonmusical duties. In addition to his responsibilities as director of church music, Bach also had various nonmusical duties in his capacity as the cantor of the school at the Thomaskirche. Since he resented these latter obligations, Bach frequently absented himself without leave, playing or examining organs, taking his son Friedemann to hear the "pretty tunes," as he called them, at the Dresden opera and fulfilling the duties of the honorary court posts that he contrived to hold all his life. To some extent, no doubt, he accepted engagements because he needed money; he complained in 1730 that his income was less than he had been led to expect (he remarked that there were not enough funerals); but, obviously, his routine work must have suffered. Friction between Bach and his employers thus developed almost at once. On the one hand, Bach's initial understanding of the fees and prerogatives accruing to his position—particularly regarding his responsibility for musical activities in the University of Leipzig's Paulinerkirche—differed from that of the town council and the university organist, Johann Gottlieb Görner. On the other

hand, Bach remained, in the eyes of his employers, their third (and unenthusiastic) choice for the post, behind Telemann and Graupner. Furthermore, the authorities insisted on admitting unmusical boys to the school, thus making it difficult for Bach to keep his churches supplied with competent singers; they also refused to spend enough money to keep a decent orchestra together. The resulting ill feeling had become serious by 1730. It was temporarily dispelled by the tact of the new rector, Johann Matthias Gesner, who admired Bach and had known him at Weimar; but Gesner stayed only until 1734 and was succeeded by Johann August Ernesti, a young man with up-to-date ideas on education, one of which was that music was not one of the humanities but a time-wasting sideline. Trouble flared up again in July 1736; it then took the form of a dispute over Bach's right to appoint prefects and became a public scandal. Fortunately for Bach, he became court composer to the Elector of Saxony in November 1736. As such, after some delay, he was able to induce his friends at court to hold an official inquiry, and his dispute with Ernesti was settled in 1738. The exact terms of the settlement are not known; but, thereafter, Bach did as he liked.

Instrumental works. In 1726, after he had completed the bulk of his cantata production, Bach began to publish the clavier *Partitas* singly, with a collected edition in 1731, perhaps with the intention of attracting recognition beyond Leipzig and thus securing a more amenable appointment elsewhere. The second part of the *Clavierübung,* containing the *Concerto in the Italian Style* and the *French Overture* (*Partita*) *in B Minor* appeared in 1735. The third part, consisting of the *Organ Mass* with the *Prelude and Fugue* ["St. Anne"] *in E Flat Major* (BWV 552), appeared in 1739. From *c.* 1729 to 1736 Bach was honorary musical director to Weissenfels; and, from 1729 to 1737 and again from 1739 for a year or two, he directed the Leipzig Collegium Musicum. For these concerts he adapted some of his earlier concertos as harpsichord concertos, thus becoming one of the first composers in history—if not the very first—of concertos for keyboard instrument and orchestra, just as he was one of the first to use the harpsichordist's right hand as a true melodic part in chamber music. These are just two of several respects in which the basically conservative and traditional Bach, as is becoming increasingly recognized, was a significant innovator as well.

nnova-ions as a :eyboard :omposer

About 1733 Bach began to produce cantatas in honour of the Elector of Saxony and his family, evidently with a view to the court appointment he secured in 1736; many of these secular movements were adapted to sacred words and re-used in the *Christmas Oratorio.* The Kyrie and Gloria of the *Mass in B Minor,* written in 1733, were also dedicated to the Elector, but the rest of the *Mass* was not put together until Bach's last years. On his visits to Dresden, Bach had won the regard of Graf Hermann Karl von Keyserlingk, the Russian envoy, who commissioned the so-called *Goldberg Variations;* these were published as part four of the *Clavierübung* about 1742, and Book Two of the "Forty-eight" seems to have been compiled about the same time. In addition, he wrote a few cantatas, revised some of his Weimar organ works, and published the so-called *Schübler Chorale Preludes* in or after 1746.

Last years. In May 1747 he visited his son Emanuel at Potsdam and played before Frederick II the Great of Prussia; in July his improvisations, on a theme proposed by the King, took shape as *The Musical Offering.* In June 1747 he joined a Sozietät der Musikalischen Wissenschaften (Society of the Musical Sciences) that had been founded by his former pupil Lorenz Christoph Mizler; he presented the canonic variations on the chorale *Vom Himmel hoch da komm' ich her* (*From Heaven Above to Earth I Come*) to the society, in manuscript, and afterward published them.

Of Bach's last illness little is known, except that it lasted several months and prevented him from finishing *The Art of the Fugue.* His constitution was undermined by two unsuccessful eye operations performed by John Taylor, the itinerant English quack who numbered Handel among his other failures; and he died on July 28, 1750, at Leipzig. His employers proceeded with relief to appoint a successor; Burgomaster Stieglitz remarked, "The school needs a cantor, not a musical director—though certainly he ought to understand music." Anna Magdalena was left badly off. For some reason, her stepsons did nothing to help her, and her own sons were too young to do so. She died on February 27, 1760, and was given a pauper's funeral.

Unfinished as it was, *The Art of the Fugue* was published in 1751. It attracted little attention and was reissued in 1752 with a laudatory preface by Friedrich Wilhelm Marpurg, a well-known Berlin musician, who later became director of the royal lottery. In spite of Marpurg and of some appreciative remarks by Johann Mattheson, the influential Hamburg critic and composer, only about 30 copies had been sold by 1756, when Emanuel Bach offered the plates for sale. As far as is known, they were sold for scrap.

Publication of *The Art of the Fugue*

Emanuel Bach and the organist-composer Johann Friedrich Agricola (a pupil of Sebastian's) wrote an obituary; Mizler added a few closing words and published the result in the journal of his society (1754). There is an English translation of it in *The Bach Reader.* Though incomplete and inaccurate, the obituary is of very great importance as a firsthand source of information.

Bach appears to have been a good husband and father. Indeed, he was the father of 20 children, only 10 of whom survived to maturity. There is amusing evidence of a certain thriftiness, a necessary virtue; for he was never more than moderately well off, and he delighted in hospitality. Living as he did at a time when music was beginning to be regarded as no occupation for a gentleman, he occasionally had to stand up for his rights both as a man and as a musician; he was then obstinate in the extreme. But no sympathetic employer had any trouble with Bach, and with his professional brethren he was modest and friendly. He was also a good teacher and from his Mühlhausen days onward was never without pupils.

REPUTATION AND INFLUENCE

For about 50 years after Bach's death, his music was neglected. This was only natural; in the days of Haydn and Mozart, no one could be expected to take much interest in a composer who had been considered old-fashioned even in his lifetime—especially since his music was not readily available, and half of it (the church cantatas) was fast becoming useless as a result of changes in religious thought.

At the same time, musicians of the late 18th century were neither so ignorant of Bach's music nor so insensitive to its influence as some modern authors have suggested. Emanuel Bach's debt to his father was considerable and Bach exercised a profound and acknowledged influence directly on Haydn, Mozart, and Beethoven.

Revival of music. After 1800 the revival of Bach's music gained momentum. The German writer Johann Nikolaus Forkel published a *Life, Genius and Works* in 1802 and acted as adviser to the publishers Hoffmeister and Kühnel, whose collected edition, begun in 1801, was cut short by the activities of Napoleon. By 1829 a representative selection of keyboard music was nonetheless available, although very few of the vocal works were published. But in that year the German musician Eduard Devrient and the German composer Felix Mendelssohn took the next step with the centenary performance of the *St. Matthew Passion.* It and the *St. John Passion* were both published in 1830; the *Mass in B Minor* followed (1832–45). The Leipzig publisher Peters began a collected edition of "piano" and instrumental works in 1837; the organ works followed in 1844–52.

Publication of the *Mass in B Minor*

Encouraged by Robert Schumann, the Bach-Gesellschaft (BG) was founded in the centenary year 1850, with the purpose of publishing the complete works. By 1900 all the known works had been printed, and the BG was succeeded by the Neue Bach-Gesellschaft (NBG), which exists still, organizing festivals and publishing popular editions. Its chief publication is its research journal, the *Bach-Jahrbuch* (from 1904). By 1950 the deficiencies of the BG edition had become painfully obvious, and the Bach-Institut was founded with headquarters at Göttingen (West Germany) and Leipzig, to produce a new standard edition (the *Neue Bach-Ausgabe* or NBA) expected to comprise 84 volumes.

In retrospect, the Bach revival, reaching back to 1800, can be recognized as the first conspicuous example of the deliberate exhumation of old music, accompanied by biographical and critical studies; and it served as an inspiration and a model for subsequent work of that kind.

Among the biographical and critical works on Bach, the most important was the monumental study *Johann Sebastian Bach* (2 vol., Leipzig, 1873–80), by the German musicologist Philipp Spitta, covering not only Bach's life and works but also a good deal of the historical background. Although wrong in many details, the book is still indispensable to the Bach student.

Editions of Bach's works. The word *Urtext* (original text) may lead the uninitiated to suppose that they are being offered an exact reproduction of what Bach wrote. It must be understood that the autographs of many important works no longer exist. Therefore, Bach's intentions often have to be pieced together from anything up to 20 sources, all different. Even first editions and facsimiles of autograph manuscripts are not infallible guides to Bach's intentions. In fact, they are often dangerously misleading, and practical musicians should take expert advice before consulting them. Editions published between 1752 and *c.* 1840 are little more than curiosities, chiefly interesting for the light they throw on the progress of the revival.

No comprehensive edition is trustworthy throughout: neither Peters nor the BG nor even the NBA. Nevertheless, it is advisable to begin by finding out whether the music desired has been published by the NBA.

MAJOR WORKS

Vocal music (sacred)

MASSES: *Mass in B Minor,* BWV 232 (1724–46); 4 Lutheran masses (*i.e.,* containing only settings of the Kyrie and the Gloria).

ORATORIOS: *Christmas Oratorio,* BWV 248 (1734); *Easter Oratorio* (*Kommt, eilet und laufet,* BWV 249; 1725); *Ascension Oratorio* (1735).

PASSIONS: *Passion According to St. John,* BWV 245 (1724); *Passion According to St. Matthew,* BWV 244 (1729).

CANTATAS: About 200 for different Sundays in the church year (1707–after 1735; mainly 1714–16, 1723–27), mostly for soloist(s), chorus, and orchestra.

OTHER WORKS: *Magnificat in D Major,* BWV 243; 7 motets; 2 Sanctus settings (3 others based on works by other composers); 186 independent chorale harmonizations.

Vocal music (secular)

CANTATAS: 24, mostly for soloists, chorus, and orchestra—all on German texts, except two Italian. They include the *Coffee Cantata* (*Schweigt stille, plaudert nicht,* BWV 211; *c.* 1732) and the *Peasant Cantata* (*Mer hahn en neue Oberkeet,* BWV 212; 1742).

OTHER WORKS: 5 songs for voice and continuo and 1 quodlibet for four voices and continuo.

Orchestral music

CONCERTI: 6 *Brandenburg Concertos* (pre-1721); 2 concertos for violin and orchestra and one for two violins (1717–23); 7 for one harpsichord, 3 for two harpsichords, 2 for three, and 1 for four harpsichords; 1 concerto for harpsichord, flute, and violin.

OTHER ORCHESTRAL WORKS: 4 overtures (suites); *Sinfonia in D Major* (incomplete).

Chamber music

SONATAS: 2 for violin and continuo; 2 for flute and continuo; 1 for two flutes and harpsichord; 2 for flute, violin, and continuo; 3 for harpsichord and flute; 3 for harpsichord and viola da gamba; 6 for harpsichord and violin.

OTHER CHAMBER MUSIC: *Das musikalisches Opfer* (1747) for strings, flute, and continuo; 6 unaccompanied sonatas (partitas) for violin (*c.* 1720); 6 unaccompanied suites (sonatas) for cello (*c.* 1720).

Organ music

CHORALE PRELUDES: 140 chorale preludes including the *Orgelbüchlein* (mainly 1714–16); *Clavierübung,* vol. 3 (1739), and *Schübler Chorale Preludes* (1746 or later).

FUGUES: 18 preludes and fugues (1708–17, 1729–39), including the "St. Anne" in E flat major and the "Wedge" in E minor; 5 toccatas and fugues (1700–17), including the "Dorian" in D minor; 3 fantasies and fugues; 4 other fugues.

OTHER ORGAN COMPOSITIONS: Variations on the chorale *Vom Himmel hoch* (1747); *Passacaglia in C Minor,* BWV 582 (1708–17); 4 concertos; 7 fantasies; 4 preludes; 6 sonatas (trios); 3 trios.

Harpsichord music

COLLECTIONS: *Clavierübung*: vol. 1 (1726–31), 6 partitas; vol. 2 (1735), *French Overture in B Minor* and *Concerto in the Italian Style;* vol. 3 (1739) is organ music with 4 "duets" for harpsichord; and vol. 4 (1742), *Goldberg Variations; The Well-Tempered Clavier,* 2 vol. (1722 and 1742), containing 48 preludes and fugues, one in each key in each book; *Clavierbüchlein* (1720), for Wilhelm Friedemann Bach, containing 15 two-part and 15 three-part inventions, 20 preludes, 2 chorale preludes, 2 allemandes, 4 minuets, a fugue, and an "applicatio"; *Clavierbüchlein* (1722) and *Notenbuch* (1725), both for Anna Magdalena Bach, containing marches, minuets, a musette, polonaises, etc.; 6 *French Suites* and 6 *English Suites.*

OTHER HARPSICHORD WORKS: *Aria variata* in A minor; 2 capriccios; *Chromatic Fantasy and Fugue;* 5 fantasies, 2 with fugues; 12 *Little Preludes;* 4 preludes and 6 for beginners; 4 preludes and fughettas, 3 preludes and fugues; 2 sonatas; 4 miscellaneous suites; 7 toccatas and arrangements.

For unspecified instrument(s)

Die Kunst der Fuge (1749); 16 fugues and 4 canons.

(W.Em/Ro.Ma.)

BIBLIOGRAPHY

Catalogs: WOLFGANG SCHMIEDER, *Thematisch-systematisches Verzeichnis der musikalischen Werke von Johann Sebastian Bach. Bach-Werke-Verzeichnis* (BWV), (1950, reprinted 1981), the standard catalog of Bach's music, including a bibliography for each work; PAUL KAST, *Die Bach-Handschriften der Berliner Staatsbibliothek* (1958), a descriptive catalog of the Bach manuscripts in the possession of the Deutsche Staatsbibliothek, East Berlin, the largest single repository, with more than 75 percent of the surviving Bach sources; MAY DEFOREST MCALL (comp.), *Melodic Index to the Works of Johann Sebastian Bach,* rev. and enl. ed. (1962), containing some 3,872 themes; WALTER KOLNEDER, *Lübbes Bach Lexikon* (1982), an illustrated dictionary.

Collections of correspondence, sketchbooks, and reminiscences: Schriftstücke von der Hand Johann Sebastian Bachs (1963), a critical edition of all surviving nonmusical documents, such as letters and receipts, in Bach's hand, and *Fremdschriftliche und gedruckte Dokumente zur Lebensgeschichte Johann Sebastian Bachs 1685–1750* (1969), a critical edition of all known printed and handwritten discussions of and references to Bach dating from his lifetime, both edited by WERNER NEUMANN and HANS-JOACHIM SCHULZE and published as supplements to the *Neue Ausgabe sämtlicher Werke* (hereafter referred to as *Neue Bach-Ausgabe*), are vol. 1 and 2 in the series *Bach-Dokumente.* Volume 3 of the series, ed. by HANS-JOACHIM SCHULZE, is *Dokumente zum nachwirken Johann Sebastian Bachs: 1750–1800* (1972); and vol. 4, ed. by WERNER NEUMANN, is *Bilddokumente zur Lebensgeschichte Johann Sebastian Bachs* (1979), with text and captions in both English and German. See also HANS T. DAVID and ARTHUR MENDEL (eds.), *The Bach Reader: A Life of Johann Sebastian Bach in Letters and Documents,* rev. ed. (1966, reprinted 1972); and ROBERT L. MARSHALL, *The Compositional Process of J.S. Bach,* 2 vol. (1972), a study of the autograph scores of the vocal works, with transcriptions of all surviving musical sketches and drafts included in vol. 2.

Biography and criticism: PHILIPP SPITTA, *Johann Sebastian Bach: His Work and Influence on the Music of Germany, 1685–1750,* 3 vol. (1884, reprinted 1951; originally published in German, 2 vol., 1873–80, reprinted 1979), a monumental study that is still the standard biography, although no longer valid in many particulars. Further important full-length studies are ALBERT SCHWEITZER, *J.S. Bach,* 2 vol. (1911, reprinted 1966; originally published in French, 1905), an influential, if highly subjective and personal, interpretation; CHARLES SANFORD TERRY, *Bach: A Biography;* 2nd rev. ed. (1933, reprinted 1972), a useful supplement (based on new archival researches) to the biographical portions of Spitta's work; KARL GEIRINGER, *Johann Sebastian Bach: The Culmination of an Era* (1966), a full-length account of the life and works that uses results of research in the 1950s by Alfred Dürr and Georg von Dadelsen bearing on the chronology of Bach's works, and *The Bach Family: Seven Generations of Creative Genius* (1954, reprinted 1981); PERCY M. YOUNG, *The Bachs: 1500–1850* (1970). Other useful sources include ALEC ROBERTSON, *Bach* (1977), a biography, with a survey of books, published editions of the works, and recordings; NORMAN CARRELL, *Bach the Borrower* (1967, reprinted 1980), on Bach's use of preexisting material; EVA MARY GREW and SYDNEY GREW, *Bach* (1947, reprinted 1977); C. HUBERT H. PARRY, *Johann Sebastian Bach: The Story of the Development of a Great Personality,* rev. ed. (1934, reprinted 1977); ROBERT L. WEAVER (ed.), *Essays on the Music of J.S. Bach and Other Divers Subjects* (1981).

On the vocal music: ALFRED DÜRR, *Die Kantaten von Johann Sebastian Bach,* 2 vol. (1971, reprinted 1979), a general survey plus individual essays on each cantata by one of the principal editors of the *Neue Bach-Ausgabe;* WERNER NEUMANN, *Handbuch der Kantaten, Johann Sebastian Bachs,* 4th rev. ed. (1971), a handbook of useful factual data and schematic analyses of all the cantatas, which is complemented by WERNER

NEUMANN (ed.), *Sämtliche von Johann Sebastian Bach vertonte Texte* (1974), complete texts of the works set to music by Bach, and *Sämtliche Kantatentexte* (1956, reissued 1967), a complete critical edition of the cantata texts; WILLIAM G. WHITTAKER, *The Cantatas of Johann Sebastian Bach, Sacred and Secular*, 2 vol. (1959, reprinted 1978), a stimulating appreciation, but one that should be used with caution, and *Fugitive Notes on Certain Cantatas and the Motets of J.S. Bach* (1924, reprinted 1976). Detailed research is given in CHARLES SANFORD TERRY, *Bach's Chorals*, 3 vol. (1915–21, reprinted in 2 vol., 1979), *Bach: The Cantatas and Oratorios*, 3 vol. (1925–29, reprinted in one vol., 1972), and *Bach: The Passions*, 2 vol. (1926, reprinted 1980); BASIL SMALLMAN, *The Background of Passion Music: J.S. Bach and His Predecessors*, 2nd rev. ed. (1970); PAUL STEINITZ, *Bach's Passions* (1978), with an overview of the history of performances of the Passions; JAMES DAY, *The Literary Background to Bach's Cantatas* (1961, reprinted 1966); HOWARD E. SMITHER, *A History of the Oratorio*, 2 vol. (1977), a survey of sacred dramatic music of the 17th century; ALEC ROBERTSON, *The Church Cantatas of J.S. Bach* (1972), with information on the religious significance of Bach's treatment of the texts; AUGUSTA RUBIN, *J.S. Bach: The Modern Composer* (1976), an analysis of his harmonic methods, with more than 1,200 quotations from the chorales.

On the instrumental music: HERMANN KELLER, *The Organ Works of Bach* (1967; originally published in German, 1948), *Die Klavierwerke Bachs* (1950), and *The Well-Tempered Clavier by Johann Sebastian Bach* (1976; originally published in German, 1965, reprinted 1981), the historical context of Bach's organ and keyboard works, and individual analyses of the compositions; DONALD F. TOVEY, *A Companion to "The Art of Fugue"* (1931, reprinted 1960), an analysis; HANS T. DAVID, *J.S. Bach's Musical Offering: History, Interpretation, and Analysis* (1945, reissued 1982); CHARLES SANFORD TERRY, *The Music of Bach* (1933, reissued 1963), and *Bach's Orchestra* (1932, reprinted 1972); ANDRÉ PIRRO, *J.S. Bach* (1957; originally published in French, 1907), and *Johann Sebastian Bach: The Organist and His Works for the Organ* (1902, reprinted 1978; trans. from French); WERNER MENKE, *History of the Trumpet of Bach and Handel* (1934, reprinted 1972; originally published in German, 1934); CECIL GRAY, *The Forty-Eight Preludes and Fugues of J.S. Bach* (1948); ALAN E.F. DICKINSON, *Bach's Fugal Works, with an Account of Fugue Before and After Bach* (1956, reprinted 1979); NORMAN CARRELL, *Bach's Brandenburg Concertos* (1963, reprinted 1977).

On performance: ERWIN BODKY, *The Interpretation of Bach's Keyboard Works* (1960, reprinted 1976), a controversial but stimulating approach; WALTER EMERY, *Bach's Ornaments* (1953), a discussion of the problems and suggested solutions, and *Notes on Bach's Organ Works*, 2 vol. (1952–57), with facsimiles of scores; ROBERT DONINGTON, *Tempo and Rhythm in Bach's Organ Music* (1960); THOMAS HARMON, *The Registration of J.S. Bach's Organ Works*, 2nd ed. (1981); HARVEY GRACE, *The Organ Works of Bach* (1922); FRITZ ROTHSCHILD, *The Lost Tradition in Music: Rhythm and Tempo in J.S. Bach's Time* (1953, reprinted 1979); FREDERICK NEUMANN, *Ornamentation in Baroque and Post-Baroque Music: With Special Emphasis on J.S. Bach* (1978), with a glossary of terms and symbols and a bibliography of primary sources. See also ARTHUR MENDEL, the preface of his edition of the vocal score of *The Passion According to St. John* (1951).

Bach's place in history is discussed in WILFRID MELLERS, *Bach and the Dance of God* (1980), focussing on the creative process and the relationship of music, word, and drama in his music; JAN CHIAPUSSO, *Bach's World* (1968, reprinted 1980), a historical study, with musical analyses; DOUGLAS R. HOFSTADTER, *Gödel, Escher, Bach: An Eternal Golden Braid* (1979, reprinted 1980), a metaphorical, philosophical work for the reader interested in the structure of Bach's music.

(W.Em./Ro.Ma./Ed.)

Francis Bacon

Francis Bacon, lawyer, courtier, statesman, philosopher, and master of the English tongue, is remembered in literary terms for the sharp worldly wisdom of a few dozen essays; by students of constitutional history for his power as a speaker in Parliament and in some famous trials and as James I's lord chancellor; and intellectually as a man who claimed all knowledge as his province and, after a magisterial survey, proceeded urgently to advocate new ways by which men might establish a legitimate command over nature for the relief of man's estate.

Francis Bacon, oil painting by an unknown artist. In the National Portrait Gallery, London.
By courtesy of the National Portrait Gallery, London

LIFE

Bacon's family

Youth and early maturity. Bacon was born Jan. 22, 1561, at York House off the Strand, London, the younger of the two sons of the lord keeper, Sir Nicholas Bacon, by his second marriage. Nicholas Bacon, born in comparatively humble circumstances, had risen to become lord keeper of the great seal. Francis' cousin through his mother was Robert Cecil, later earl of Salisbury and chief minister of the crown at the end of Elizabeth I's reign and the beginning of James I's. From 1573 to 1575 Bacon was educated at Trinity College, Cambridge, but his weak constitution caused him to suffer ill health there. His distaste for what he termed "unfruitful" Aristotelian philosophy began at Cambridge. From 1576 to 1579 Bacon was in France as a member of the English ambassador's suite. He was recalled abruptly after the sudden death of his father, who left him relatively little money. Bacon remained financially embarrassed virtually until his death.

Early legal career and political ambitions. In 1576 Bacon had been admitted as an "ancient" (senior governor) of Gray's Inn, one of the four Inns of Court that served as institutions for legal education, in London. In 1579 he took up residence there and after becoming a barrister in 1582 progressed in time through the posts of reader (lecturer at the Inn), bencher (senior member of the Inn), and queen's (from 1603 king's) counsel extraordinary to those of solicitor general and attorney general. Even as successful a legal career as this, however, did not satisfy his political and philosophical ambitions.

Bacon occupied himself with the tract "Temporis Partus Maximus" ("The Greatest Part of Time") in 1582; it has not survived. In 1584 he sat as member of Parliament for Melcombe Regis in Dorset and subsequently represented Taunton, Liverpool, the County of Middlesex, Southamp-

ton, Ipswich, and the University of Cambridge. In 1589 a "Letter of Advice" to the Queen and *An Advertisement Touching the Controversies of the Church of England* indicated his political interests and showed a fair promise of political potential by reason of their levelheadedness and disposition to reconcile. In 1593 came a setback to his political hopes: he took a stand objecting to the government's intensified demand for subsidies to help meet the expenses of the war against Spain. Elizabeth took offense, and Bacon was in disgrace during several critical years when there were chances for legal advancement.

Relationship with Essex. Meanwhile, sometime before July 1591, Bacon had become acquainted with Robert Devereux, the young earl of Essex, who was a favourite of the Queen, although still in some disgrace with her for his unauthorized marriage to the widow of Sir Philip Sidney. Bacon saw in the Earl the "fittest instrument to do good to the State" and offered Essex the friendly advice of an older, wiser, and more subtle man. Essex did his best to mollify the Queen, and when the office of attorney general fell vacant, he enthusiastically but unsuccessfully supported the claim of Bacon. Other recommendations by Essex for high offices to be conferred on Bacon also failed.

By 1598 Essex' failure in an expedition against Spanish treasure ships made him harder to control; and although Bacon's efforts to divert his energies to Ireland, where the people were in revolt, proved only too successful, Essex lost his head when things went wrong and he returned against orders. Bacon certainly did what he could to accommodate matters but merely offended both sides; in June 1600 he found himself as the Queen's learned counsel taking part in the informal trial of his patron. Essex bore him no ill will and shortly after his release was again on friendly terms with him. But after Essex' abortive attempt of 1601 to seize the Queen and force her dismissal of his rivals, Bacon, who had known nothing of the project, viewed Essex as a traitor and drew up the official report on the affair. This, however, was heavily altered by others before publication.

The traitorous actions of Essex

After Essex' execution Bacon, in 1604, published the *Apologie in Certaine Imputations Concerning the Late Earle of Essex* in defense of his own actions. It is a coherent piece of self-justification, but to posterity it does not carry complete conviction, particularly since it evinces no personal distress.

Career in the service of James I. When Elizabeth died in 1603, Bacon's letter-writing ability was directed to finding a place for himself and a use for his talents in James I's services. He pointed to his concern for Irish affairs, the union of the kingdoms, and the pacification of the church as proof that he had much to offer the new king.

Through the influence of his cousin Robert Cecil, Bacon was one of the 300 new knights dubbed in 1603. The following year he was confirmed as learned counsel and sat in the first Parliament of the new reign in the debates of its first session. He was also active as one of the commissioners for discussing a union with Scotland. In the autumn of 1605 he published his *Advancement of Learning,* dedicated to the King, and in the following summer he married Alice Barnham, the daughter of a London alderman. Preferment in the royal service, however, still eluded him, and it was not until June 1607 that his petitions and his vigorous though vain efforts to persuade the Commons to accept the King's proposals for union with Scotland were at length rewarded with the post of solicitor general. Even then, his political influence remained negligible, a fact that he came to attribute to the power and jealousy of Cecil, by then earl of Salisbury and the King's chief minister. In 1609 his *De Sapientia Veterum* ("The Wisdom of the Ancients"), in which he expounded what he took to be the hidden practical meaning embodied in ancient myths, came out and proved to be, next to the *Essayes,* his most popular book in his own lifetime. In 1614 he seems to have written *The New Atlantis,* his far-seeing scientific utopian work, which did not get into print until 1626.

Political advancement

After Salisbury's death in 1612, Bacon renewed his efforts to gain influence with the King, writing a number of remarkable papers of advice upon affairs of state and, in particular, upon the relations between Crown and Parliament. The King adopted his proposal for removing Coke from his post as chief justice of the common pleas and appointing him to the King's Bench, while appointing Bacon attorney general in 1613. During the next few years Bacon's views about the royal prerogative brought him, as attorney general, increasingly into conflict with Coke, the champion of the common law and of the independence of the judges. It was Bacon who examined Coke when the King ordered the judges to be consulted individually and separately in the case of Edmond Peacham, a clergyman charged with treason as the author of an unpublished treatise justifying rebellion against oppression. Bacon has been reprobated for having taken part in the examination under torture of Peacham, which turned out to be fruitless. It was Bacon who instructed Coke and the other judges not to proceed in the case of commendams (*i.e.,* holding of benefices in the absence of the regular incumbent) until they had spoken to the King. Coke's dismissal in November 1616 for defying this order was quickly followed by Bacon's appointment as lord keeper of the great seal in March 1617. The following year he was made lord chancellor and baron Verulam, and in 1620/21 he was created viscount St. Albans.

The main reason for this progress was his unsparing service in Parliament and the court, together with persistent letters of self-recommendation; according to the traditional account, however, he was also aided by his association with George Villiers, later duke of Buckingham, the King's new favourite. It would appear that he became honestly fond of Villiers; many of his letters betray a feeling that seems warmer than timeserving flattery.

Among Bacon's papers a notebook has survived, the *Commentarius Solutus* ("Loose Commentary"), which is revealing. It is a jotting pad "like a Marchant's wast booke where to enter all maner of remembrance of matter, fourme, business, study, towching my self, service, others, eyther sparsim or in schedules, without any maner of restraint." This book reveals Bacon reminding himself to flatter a possible patron, to study the weaknesses of a rival, to set intelligent noblemen in the Tower of London to work on serviceable experiments. It displays the multiplicity of his concerns: his income and debts, the King's business, his own garden and plans for building, philosophical speculations, his health, including his symptoms and medications, and an admonition to learn to control his breathing and not to interrupt in conversation. Between 1608 and 1620 he prepared at least 12 draftings of his most celebrated work, the *Novum Organum,* and wrote several minor philosophical works.

The major occupation of these years must have been the management of James, always with reference, remote or direct, to the royal finances. The King relied on his lord chancellor but did not always follow his advice. Bacon was longer sighted than his contemporaries and seems to have been aware of the constitutional problems that were to culminate in civil war; he dreaded innovation and did all he could, and perhaps more than he should, to safeguard the royal prerogative. Whether his policies were sound or not, it is evident that he was, as he later said, "no mountebank in the King's services."

Fall from power. By 1621 Bacon must have seemed impregnable, a favourite not by charm (though he was witty and had a dry sense of humour) but by sheer usefulness and loyalty to his sovereign; lavish in public expenditure (he was once the sole provider of a court masque); dignified in his affluence and liberal in his household; winning the attention of scholars abroad as the author of the *Novum Organum,* published in 1620, and the developer of the *Instauratio Magna* ("Great Instauration"), a comprehensive plan to reorganize the sciences and to restore man to that mastery over nature that he was conceived to have lost by the fall of Adam. But Bacon had his enemies. In 1618 he fell foul of George Villiers when he tried to interfere in the marriage of the daughter of his old enemy, Coke, and the younger brother of Villiers. Then, in 1621, two charges of bribery were raised against him before a committee of grievances over which he himself presided. The shock appears to have been twofold because Bacon,

Charges of bribery

who was casual about the incoming and outgoing of his wealth, was unaware of any vulnerability and was not mindful of the resentment of two men whose cases had gone against them in spite of gifts they had made with the intent of bribing the judge. The blow caught him when he was ill, and he pleaded for extra time to meet the charges, explaining that genuine illness, not cowardice, was the reason for his request. Meanwhile, the House of Lords collected another score of complaints. Bacon admitted the receipt of gifts but denied that they had ever affected his judgment; he made notes on cases and sought an audience with the King that was refused. Unable to defend himself by discriminating between the various charges or cross-examining witnesses, he settled for a penitent submission and resigned the seal of his office, hoping that this would suffice. The sentence was harsh, however, and included a fine of £40,000, imprisonment in the Tower of London during the King's pleasure, disablement from holding any state office, and exclusion from Parliament and the verge of court (an area of 12 miles radius centred on where the sovereign is resident). Bacon commented to Buckingham: "I acknowledge the sentence just, and for reformation's sake fit, *the justest Chancellor that hath been in the five changes since Sir Nicolas Bacon's time.*" The magnanimity and wit of the epigram sets his case against the prevailing standards.

Bacon did not have to stay long in the Tower, but he found the ban that cut him off from access to the library of Charles Cotton, an English man of letters, and from consultation with his physician more galling. He came up against an inimical lord treasurer, and his pension payments were delayed. He lost Buckingham's goodwill for a time and was put to the humiliating practice of roundabout approaches to other nobles and to Count Gondomar, the Spanish ambassador; remissions came only after vexations and disappointments. Despite all this his courage held, and the last years of his life were spent in work far more valuable to the world than anything he had accomplished in his high office. Cut off from other services, he offered his literary powers to provide the King with a digest of the laws, a history of Great Britain, and biographies of Tudor monarchs. He prepared memorandums on usury and on the prospects of a war with Spain; he expressed views on educational reforms; he even returned, as if by habit, to draft papers of advice to the King or to Buckingham and composed speeches he was never to deliver. Some of these projects were completed, and they did not exhaust his fertility. He wrote: "If I be left to myself I will graze and bear natural philosophy." Two out of a plan of six separate natural histories were composed—*Historia Ventorum* ("History of the Winds") appeared in 1622 and *Historia Vitae et Mortis* ("History of Life and Death") in the following year. Also in 1623 he published the *De Dignitate et Augmentis Scientiarum,* a Latin translation, with many additions, of the *Advancement of Learning.* He also corresponded with Italian thinkers and urged his works upon them. In 1625 a third and enlarged edition of his *Essayes* was published.

Profuse literary production of his last years

Bacon in adversity showed patience, unimpaired intellectual vigour, and fortitude. Physical deprivation distressed him but what hurt most was the loss of favour; it was not until Jan. 20, 1622/23, that he was admitted to kiss the King's hand; a full pardon never came. Finally, in March 1626, driving one day near Highgate (a district to the north of London) and deciding on impulse to discover whether snow would delay the process of putrefaction, he stopped his carriage, purchased a hen, and stuffed it with snow. He was seized with a sudden chill, which brought on bronchitis, and he died at the Earl of Arundel's house nearby on April 9, 1626. (K.M.L./A.M.Q.)

THOUGHT AND WRITINGS

The intellectual background. Bacon appears as an unusually original thinker for several reasons. In the first place he was writing, in the early 17th century, in something of a philosophical vacuum so far as England was concerned. The last great English philosopher, William of Ockham, had died in 1347, two and a half centuries before the *Advancement of Learning;* the last really important philosopher, John Wycliffe, had died not much later, in 1384.

The 15th century had been intellectually cautious and torpid, leavened only by the first small importations of Italian humanism by such cultivated dilettantes as Humphrey Plantagenet, duke of Gloucester, and John Tiptoft, earl of Worcester. The Christian Platonism of the Renaissance became more established at the start of the 16th century in the circle of Erasmus' English friends: the so-called Oxford Reformers—John Colet, William Grocyn, and Thomas More. But that initiative succumbed to the ecclesiastical frenzies of the age. Philosophy did not revive until Richard Hooker in the 1590s put forward his moderate Anglican version of Thomist rationalism in the form of a theory of the Elizabethan church settlement. This happened a few years before Bacon began to write.

In England three systems of thought prevailed in the late 16th century: Aristotelian Scholasticism, scholarly and aesthetic humanism, and occultism. Aristotelian orthodoxy had been reanimated in Roman Catholic Europe after the Council of Trent and the Counter-Reformation had lent authority to the massive output of the 16th-century Spanish theologian and philosopher Francisco Suárez. In England learning remained in general formally Aristotelian, even though some criticism of Aristotle's logic had reached Cambridge at the time Bacon was a student there in the mid-1570s. But such criticism sought simplicity for the sake of rhetorical effectiveness and not, as Bacon's critique was to do, in the interests of substantial, practically useful knowledge of nature.

Prevailing systems of thought in England

The Christian humanist tradition of Petrarch, Lorenzo Valla, and, more recently, of Erasmus was an active force. In contrast to orthodox asceticism, this tradition, in some aspects, inclined to glorify the world and its pleasures and to favour the beauty of art, language, and nature, while remaining comparatively indifferent to religious speculation. Attraction to the beauty of nature, however, if it did not cause was at any rate combined with neglect and disdain for the knowledge of nature. Educationally it fostered the sharp separation between the natural sciences and the humanities that has persisted ever since. Philosophically it was skeptical, nourishing itself, notably in the case of Montaigne, on the rediscovery in 1562 of Sextus Empiricus' comprehensive survey of the skepticism of Greek thought after Aristotle.

The third important current of thought in the world into which Bacon was born was that of occultism, or esotericism, that is, the pursuit of mystical analogies between man and the cosmos, or the search for magical powers over natural processes, as in alchemy and the concoction of elixirs and panaceas. Although its most famous exponent, Paracelsus, was German, occultism was well rooted in England, appealing as it did to the individualistic style of English credulity. Robert Fludd, the leading English occultist, was an approximate contemporary of Bacon. Bacon himself has often been held to have been some kind of occultist, and, even more questionably, to have been a member of the Rosicrucian order, but the sort of "natural magic" he espoused and advertised was altogether different from that of the esoteric philosophers.

There was a fourth mode of Renaissance thought outside England to which Bacon's thinking bore some affinity. Like that of the humanists it was inspired by Plato, at least to some extent, but by another part of his thought, namely its cosmology. This was the boldly systematic nature-philosophy of Nicholas of Cusa and of a number of Italians, in particular Bernardino Telesio, Francesco Patrizzi, Tommaso Campanella, and Giordano Bruno. Nicholas of Cusa and Bruno were highly speculative, but Telesio and, up to a point, Campanella affirmed the primacy of sense perception. In a way that Bacon was later to elaborate formally and systematically, they held knowledge of nature to be a matter of extrapolating from the findings of the senses. There is no allusion to these thinkers in Bacon's writings. But although he was less metaphysically adventurous than they were, he shared with them the conviction that the human mind is fitted for knowledge of nature and must derive it from observation, not from abstract reasoning.

Related Renaissance thought outside England

Bacon's scheme. Bacon drew up an ambitious plan for a comprehensive work that was to appear under the title of *Instauratio Magna* ("The Great Instauration"), but like many of his literary schemes, it was never completed. Its first part, *De Augmentis Scientiarum,* appeared in 1623 and is an expanded, Latinized version of his earlier work the *Advancement of Learning,* published in 1605 (the first really important philosophical book to be written in English). The *De Augmentis Scientiarum* contains a division of the sciences, a project that had not been embarked on to any great purpose since Aristotle and, in a smaller way, since the Stoics. The second part of Bacon's scheme, the *Novum Organum,* which had already appeared in 1620, gives "true directions concerning the interpretation of nature," in other words, an account of the correct method of acquiring natural knowledge. This is what Bacon believed to be his most important contribution and is the body of ideas with which his name is most closely associated. The fields of possible knowledge having been charted in *De Augmentis Scientiarum,* the proper method for their cultivation was set out in *Novum Organum.*

Instauratio Magna

Third, there is natural history, the register of matters of observed natural fact, which is the indispensable raw material for the inductive method. Bacon wrote "histories," in this sense, of the wind, of life and death, and of the dense and the rare, and, near the end of his life, he was working on his *Sylva Sylvarum: Or A Natural Historie* ("Forest of Forests"), in effect, a collection of collections, a somewhat uncritical miscellany.

Fourth, there is the "ladder of the intellect," consisting of thoroughly worked out examples of the Baconian method in application, the most successful one being the exemplary account in *Novum Organum* of how his inductive "tables" show heat to be a kind of motion of particles. Fifth, there are the "forerunners," or pieces of scientific knowledge arrived at by pre-Baconian, common sense methods. Sixth and finally, there is the new philosophy, or science itself, seen by Bacon as a task for later generations armed with his method, advancing into all the regions of possible discovery set out in the *Advancement of Learning.* The wonder is not so much that Bacon did not complete this immense design but that he got as far with it as he did.

The idols of the mind. In the first book of *Novum Organum* Bacon discusses the causes of human error in the pursuit of knowledge. Aristotle had discussed logical fallacies, commonly found in human reasoning, but Bacon was original in looking behind the forms of reasoning to underlying psychological causes. He invented the metaphor of "idol" to refer to such causes of human error.

Causes of human error

Bacon distinguishes four idols, or main varieties of proneness to error. The idols of the tribe are certain intellectual faults that are universal to mankind, or, at any rate, very common. One, for example, is a tendency toward oversimplification, that is, toward supposing, for the sake of tidiness, that there exists more order in a field of inquiry than there actually is. Another is a propensity to be overly influenced by particularly sudden or exciting occurrences that are in fact unrepresentative.

The idols of the cave are the intellectual peculiarities of individuals. One person may concentrate on the likenesses, another on the differences, between things. One may fasten on detail, another on the totality.

The idols of the marketplace are the kinds of error for which language is responsible. It has always been a distinguishing feature of English philosophy to emphasize the unreliable nature of language, which is seen, nominalistically, as a human improvisation. Nominalists argue that even if the power of speech is given by God, it was Adam who named the beasts and thereby gave that power its concrete realization. But language, like other human achievements, partakes of human imperfections. Bacon was particularly concerned with the superficiality of distinctions drawn in everyday language, by which things fundamentally different are classed together (whales and fishes as fish, for example) and things fundamentally similar are distinguished (ice, water, and steam). But he was also concerned, like later critics of language, with the capacity of words to embroil men in the discussion of the meaningless (as, for example, in discussions of the deity Fortune). This aspect of Bacon's thought has been almost as influential as his account of natural knowledge, inspiring a long tradition of skeptical rationalism, from the Enlightenment to Comtian positivism of the 19th and logical positivism of the 20th centuries.

The fourth and final group of idols is that of the idols of the theatre, that is to say mistaken systems of philosophy in the broadest, Baconian sense of the term, in which it embraces all beliefs of any degree of generality. Bacon's critical polemic in discussing the idols of the theatre is lively but not very penetrating philosophically. He speaks, for example, of the vain affectations of the humanists, but they were not a very apt subject for his criticism. Humanists were really anti-philosophers who not unreasonably turned their attention to nonphilosophical matters because of the apparent inability of philosophers to arrive at conclusions that were either generally agreed upon or useful. Bacon does have something to say about the skeptical philosophy to which humanists appealed when they felt the need for it. Insofar as skepticism involves doubts about deductive reasoning, he has no quarrel with it. Insofar as it is applied not to reason but to the ability of the senses to supply the reason with reliable premises to work from, he brushes it aside too easily.

Bacon's attack on Scholastic orthodoxy is surprisingly rhetorical. It may be that he supposed it to be already sufficiently discredited by its incurably contentious or disputatious character. In his view it was a largely verbal technique for the indefinite prolongation of inconclusive argument by the drawing of artificial distinctions. He has some awareness of the central weakness of Aristotelian science, namely its attempt to derive substantial conclusions from premises that are intuitively evident, and argues that the apparently obvious axioms are neither clear or indisputable. Perhaps Bacon's most fruitful disagreement with Scholasticism is his belief that natural knowledge is cumulative, a process of discovery, not of conservation. Living in a time when new worlds were being found on Earth, he was able to free himself from the view that everything men needed to know had already been revealed in the Bible or by Aristotle.

Bacon's criticism of Scholasticism

Against the fantastic learning of the occultists Bacon argued that individual reports are insufficient, especially since men are emotionally predisposed to credit the interestingly strange. Observations worthy to substantiate theories must be repeatable. Bacon defended the study of nature against those who considered it as either base or dangerous. He argued for a cooperative and methodical procedure and against individualism and intuition.

The classification of the sciences. Book II of the *Advancement of Learning* and Books II to IX of the *De Augmentis Scientiarum* contain an unprecedentedly thorough and detailed systematization of the whole range of human knowledge. Bacon begins with a distinction of three faculties—memory, imagination, and reason—to which are respectively assigned history, "poesy," and philosophy. History has an inclusive sense and means all knowledge of singular, individual matters of fact. "Poesy" is "feigned history" and not taken to be cognitive at all and so really irrelevant. After subdividing poesy perfunctorily into narrative, representative (or dramatic), and allusive (or parabolical) forms, Bacon gives it no further consideration.

History is divided into natural and civil, the civil category also including ecclesiastical and literary history (which for Bacon is really the history of ideas). History supplies the raw material for philosophy, in other words for the general knowledge that is inductively derived from it. Although Bacon proclaims the universal applicability of induction, he himself treats it almost exclusively as a means to natural knowledge and ignores its civil (or social) application.

Two further general distinctions should be mentioned. The first is between the divine and the secular. Most divine knowledge must come from revelation, and reason has nothing to do with it. There is such a thing as divine philosophy (what was later called rational, or natural, theology), but its sole task and competence is to prove that there is a God. The second, more pervasive distinction is between theoretical and practical disciplines, that is, between sciences proper and technologies, or "arts."

Bacon acknowledges something he calls first philosophy, which is secular but not confined to nature or to society. It is concerned with the principles, such as they are, that are common to all the sciences. Natural philosophy divides into natural science as theory on the one hand and the practical discipline of applying natural science's findings to "the relief of man's estate" on the other, which he misleadingly describes as natural magic. The former is "the inquisition of causes," the latter, "the production of effects."

To subdivide still further, natural science is made up of physics and metaphysics, as Bacon understands it. Physics, in his interpretation, is the science of observable correlations; metaphysics is the more theoretical science of the underlying structural factors that explains observable regularities. Each has its practical, or technological, partner; that of physics is mechanics, that of metaphysics, natural magic. It is to the latter that one must look for the real transformation of the human condition through scientific progress. Mechanics is just levers and pulleys.

The function of mathematics in science

Mathematics is seen by Bacon as an auxiliary to natural science. Many subsequent philosophers of science would agree, understanding it to be a logical means of expressing the content of scientific propositions or of extracting part of that content. But Bacon is not clear about how mathematics was to be of service to science and does not realize that the Galilean physics developing in his own lifetime was entirely mathematical in form. Although one of his three inductive tables is concerned with correlated variations in degree (while the others concern likenesses and differences in kind), he really has no conception of the role, already established in science, of exact numerical measurement.

Bacon is fairly cursory about "human philosophy." Four somewhat quaint sciences of body are sketched—medicine, cosmetic, athletic, and "the voluptuary arts." The sciences of mind—logic and ethics—are practical, consisting of sets of rules for the correct management of reasoning or conduct, with no suggested theoretical counterpart. Bacon is unreflectively conventional about moral truth, content to rely on the deliverances of the long historical sequence of moralists, undisturbed by their disagreements with one another.

Bacon represents civil philosophy in the same uninquiringly practical way. It comprises not only the art of government but also "conversation," or the art of persuasion, and "negotiation," or prudence, the topic of proverbs and, to a considerable extent, of his own *Essayes.*

In principle, Bacon is committed to the view that human beings and society are as well fitted for inductive, and, in 20th-century terms, scientific study as the natural world. Yet he depicts human and social studies as the field of nothing more refined than common sense. It was, of course, an achievement to extricate them from religion, and to do so without unnecessary provocation. But in his conception they remain practical arts with no sustaining body of scientific theory to ratify them. It was left to Thomas Hobbes, for a time Bacon's amanuensis, to develop complete systems of human and social science. Bacon's practice, however, was better than his program. In his writings on history and law he went beyond the commonplaces of chronicle and precedent and engaged in explanation and theory.

Novum Organum

The new method. The core of Bacon's philosophy of science is the account of inductive reasoning given in Book II of *Novum Organum.* The defect of all previous systems of beliefs about nature, he argued, lay in the inadequate treatment of the general propositions from which the deductions were made. Either they were the result of precipitate generalization from one or two cases, or they were uncritically assumed to be self-evident on the basis of their familiarity and general acceptance.

In order to avoid hasty generalization Bacon urges a technique of "gradual ascent," that is, the patient accumulation of well-founded generalizations of steadily increasing degrees of generality. This method would have the beneficial effect of loosening the hold on men's minds of ill-constructed everyday concepts that obliterate important differences and fail to register important similarities.

The crucial point, Bacon realized, is that induction must work by elimination not, as it does in common life and the defective scientific tradition, by simple enumeration. Thus he stressed "the greater force of the negative instance"—the fact that while "all A are B" is only very weakly confirmed by "this A is B," it is shown conclusively to be false by "this A is not B." He devised tables, or formal devices for the presentation of singular pieces of evidence, in order to facilitate the rapid discovery of false generalizations. What survives this eliminative screening, Bacon assumes, may be taken to be true.

Bacon's tables

Bacon presents tables of presence, of absence, and of degree. Tables of presence contain a collection of cases in which one specified property is found. They are then compared to each other to see what other properties are always present. Any property not present in just one case in such a collection cannot be a necessary condition of the property being investigated. Second, there are tables of absence, which list cases that are as alike as possible to the cases in the tables of presence except for the property under investigation. Any property that is found in the second case cannot be a sufficient condition of the original property. Finally, in tables of degree proportionate variations of two properties are compared to see if the proportion is maintained.

Bacon rightly showed some hesitation in arriving at the goal he had prescribed for himself, namely constructing a method that would yield general propositions about substantial matters of natural fact that were certain and beyond reasonable doubt. But he hesitated for an insufficient, secondary reason. The application of his tables to a mass of singular evidence, he said, would give only a "first vintage," a provisional approximation to the truth, because of the defects of natural history, that is to say, the defects inherent in the formulation of the evidence.

There are, however, more serious difficulties. An obvious one is that Bacon assumed both that every property natural science can investigate actually has some other property which is both its necessary and sufficient condition (a very strong version of determinism) and also that the conditioning property in each case is readily discoverable. What he had himself laid down as the task of metaphysics in his sense (theoretical natural science in 20th-century terms), namely the discovery of the hidden "forms" that explain what is observed, ensured that the tables could not serve for that task since they are confined to the perceptible accompaniments of what is to be explained. This point is implied by critics who have accused Bacon of failing to recognize the indispensable role of hypotheses in science. In general he adopted a naive and unreflective view about the nature of causes, ignoring their possible complexity and plurality (pointed out by John Stuart Mill) as well as the possibility that they could be at some distance in space and time from their effects.

Another weakness, not sufficiently emphasized, is Bacon's preoccupation with the static. The science that came to glorious maturity in his own century was concerned with change, and, in particular, with motion, as is the natural science of the 20th century. It was with this aspect of the natural world that mathematics, whose role Bacon did not see, came so fruitfully to grips.

In the end it may be that the conception of a scientific research establishment, which Bacon developed in his utopia, *The New Atlantis,* was a more important contribution to science than his theory of induction. Here the idea of science as a collaborative undertaking, conducted in an impersonally methodical fashion and animated by the intention to give material benefits to mankind, is set out with literary force.

Human philosophy. Although, as was pointed out above, Bacon's programmatic account of "human and civic philosophy" (*i.e.,* human and social science) treats it as a matter of practical art, or technique, his own ventures into history and jurisprudence, at any rate, were of a strongly theoretical cast. His *Historie of the Raigne of King Henry the Seventh* is explanatory, interpretative history, making sense of the King's policies by tracing them to his cautious, economical, and secretive character. Similarly his reflections on law, in *De Augmentis Scientiarum*

and in *Maxims of the Law* (Part I of *The Elements of the Common Lawes of England*), are genuine jurisprudence, not the type of commentary informed by precedent with which most jurists of his time were content. In politics Bacon was as anxious to detach the state from religion as he was to disentangle science from it—both concerns being indicative of very little positive enthusiasm for religion, despite the formal professions of profound respect convention extracted from him. He endorsed the Tudor monarchy and defended it against Coke's legal obstruction because it was rational and efficient. He had no patience with the inanities of divine right with which James I was infatuated. Bacon wrote little about education, but his memorable assault on the Scholastic obsession with words—an obsession largely carried over, if to different words, by the humanists—bore fruit in the educational theory of Comenius, who acknowledged Bacon's influence in his argument that children should study actual things as well as books.

Political views

Assessment and influence. Bacon's personality has usually been regarded as unattractive: he was cold-hearted, cringed to the powerful, and took bribes, and then had the impudence to say he had not been influenced by them. There is no reason to question this assessment in its fundamentals. It was a hard world for someone in his situation to cut a good figure in, and he did not try to do so. The grimly practical style of his personality is reflected in the particular service he was able to provide of showing a purely secular mind of the highest intellectual power at work. No one who wrote so well could have been insensitive to art. But no one before him had ever quite so uncompromisingly excluded art from the cognitive domain.

Bacon was a hero to Robert Hooke and Robert Boyle, founders of the Royal Society. Jean d'Alembert, classifying the sciences in the *Encyclopédie,* saluted him. Kant, rather surprisingly for one so concerned to limit science in order to make room for faith, dedicated the *Critique of Pure Reason* to him. He was attacked by Joseph de Maistre for setting man's miserable reason up against God but glorified by Auguste Comte.

It has been suggested that Bacon's thought received proper recognition only with 19th-century biology, which, unlike mathematical physics, really is Baconian in procedure. Darwin undoubtedly thought so. Bacon's belief that a new science could contribute to the relief of man's estate also had to await its time. In the 17th century the chief inventions that flowed from science were of instruments that enabled science to progress further. Today Bacon is best known among philosophers as the symbol of the idea, widely held to be mistaken, that science is inductive. Although there is more to his thought than that, it is, indeed, central; but even if it is wrong, it is as well to have it so boldly and magnificently presented. (A.M.Q.)

MAJOR WORKS

PHILOSOPHICAL WORKS: *The Twoo Bookes of Francis Bacon. Of the Proficience and Advancement of Learning Divine and Humane* (1605); *Instauratio Magna* (1620), also known as *Novum Organum; Historia Naturalis et Experimentalis ad Condendam Philosophiam: Sive Phaenomena Universi* (1622), also known as *Historia Ventorum; Historia Vitae & Mortis* (1623); *De Dignitate et Augmentis Scientiarum* (1623).

LITERARY AND HISTORICAL WORKS: *Essayes* (1597), 10 essays enlarged to 38 as *The Essaies of Sr Francis Bacon Knight* (1612), and to 58 as *The Essayes or Counsels, Civill and Morall* (1625); *Francisci Baconi De Sapientia Veterum Liber* (1609); *The Historie of the Raigne of King Henry the Seventh* (1622).

POLITICAL WORKS: *A Declaration of the Practices & Treasons Attempted and Committed by Robert, Late Earle of Essex* (1601); *Certain Considerations Touching the Better Pacification, and Edification of the Church of England* (1604); *Sir Francis Bacon His Apologie, in Certaine Imputations Concerning the Late Earle of Essex* (1604).

LEGAL WORKS: *The Elements of the Common Lawes of England* (1630); *Cases of Treason* (1641); *The Learned Reading of Sir Francis Bacon, One of Her Majesties Learned Counsell at Law, upon the Statute of Uses* (1642).

POSTHUMOUSLY PUBLISHED WORKS: *Sylva Sylvarum: Or A Naturall Historie* (1627, with the unfinished *The New Atlantis*).

The standard edition, containing most of Bacon's writings, is James Spedding, Robert Leslie Ellis, and Douglas Denon Heath (eds.), *The Works of Francis Bacon,* 14 vol. (1857–74), which also contains English translations of many of the works and learned commentaries. The purely philosophical works were extracted from this in John M. Robertson (ed.), *The Philosophical Works of Francis Bacon* (1905, reprinted 1970). Benjamin Farrington, *The Philosophy of Francis Bacon: An Essay on Its Development from 1603 to 1609* (1964, reprinted 1966), contains English translations of several of Bacon's lesser known essays that, nonetheless, shed important light on his philosophy. A comprehensive, critical edition of Bacon's essays is Michael Kiernan (ed.), *The Essayes or Counsels, Civill and Morall* (1985).

BIBLIOGRAPHY. JAMES SPEDDING, *An Account of the Life and Times of Francis Bacon,* 2 vol. (1878, reprinted 1880), is a comprehensive biography based on the second half of the above-mentioned *Works of Francis Bacon,* titled *The Letters and the Life of Francis Bacon,* 7 vol. (1861–74). Spedding also produced *Evenings with a Reviewer: or, A Free and Particular Examination of Mr. Macaulay's Article on Lord Bacon: In a Series of Dialogues,* 2 vol. (1848, reissued 1881 as *Evenings with a Reviewer; or, Macaulay and Bacon*), which is a detailed and illuminating rebuttal of a famous attack made on the moral character of Bacon. There is a valuable section devoted to Bacon in JOHN AUBREY, *Brief Lives,* ed. by OLIVER LAWSON DICK (1949, reissued 1982). Also valuable is a memoir by Bacon's chaplain, WILLIAM RAWLEY, "The Life of the Right Honourable Francis Bacon, Baron of Verulam, Viscount St. Alban," which can be found in vol. 1 of the Spedding, Ellis, and Heath edition of *The Works of Francis Bacon* mentioned above. Other more recent works include FULTON H. ANDERSON, *Francis Bacon: His Career and His Thought* (1962, reprinted 1978); and JOEL J. EPSTEIN, *Francis Bacon: A Political Biography* (1977), a survey of his public career.

Critical studies (philosophy): THOMAS FOWLER, *Bacon* (1881); C.D. BROAD, *The Philosophy of Francis Bacon* (1926, reprinted 1976), an excellent lecture; FULTON H. ANDERSON, *The Philosophy of Francis Bacon* (1948, reissued 1971), a major, influential study; BENJAMIN FARRINGTON, *Francis Bacon: Philosopher of Industrial Science* (1949, reprinted 1979), which interprets Bacon's aims as primarily practical and industrial; LISA JARDINE, *Francis Bacon: Discovery and the Art of Discourse* (1974), an attempt to unify Bacon's method in scientific, political, and literary areas; ANTHONY QUINTON, *Francis Bacon* (1980), a short and readable exposition of Bacon's philosophy; PETER URBACH, *Francis Bacon's Philosophy of Science: An Account and a Reappraisal* (1987); RICHARD FOSTER JONES, *Ancients and Moderns,* 2nd ed. (1961, reprinted 1982), a study of the rise of the scientific movement in 17th-century England, seen largely as a movement inspired by Bacon's writings; MARY B. HESSE, "Francis Bacon," in D.J. O'CONNOR (ed.), *A Critical History of Western Philosophy* (1964, reissued 1985), pp. 141–153, an excellent and influential article on Bacon's scientific method; MARY HORTON, "In Defence of Francis Bacon: A Criticism of the Critics of the Inductive Method," *Studies in History and Philosophy of Science,* 4(2):241–278 (August 1973), a good exposition of some of Bacon's experimental principles; JOSHUA C. GREGORY, "Chemistry and Alchemy in the Natural Philosophy of Sir Francis Bacon, 1561–1626," *Ambix,* 2(2):93–111 (September 1938), a good account of some of Bacon's cosmological views; DOUGLAS BUSH, *English Literature in the Earlier Seventeenth Century, 1600–1660,* 2nd ed. rev. (1962, reprinted 1976), which contains a good critique of Bacon's historical, political, and legal writings; and ROBERT LESLIE ELLIS, "General Preface to Bacon's Philosophical Works," in vol. 1 of the Spedding, Ellis, and Heath edition of *The Works of Francis Bacon,* a notable and influential interpretation of Bacon's philosophy. For further research, consult REGINALD WALTER GIBSON, *Francis Bacon: A Bibliography of His Works and Baconiana to the Year 1750* (1950), and *Francis Bacon: Supplement* (1959).

Miscellaneous critical studies: CHARLES W. LEMMI, *The Classic Deities in Bacon: A Study in Mythological Symbolism* (1933, reprinted 1978), an important study of *De Sapientia Veterum.* See also GEOFFREY BULLOUGH, "Bacon and the Defense of Learning," pp.1–20, and RUDOLPH METZ, "Bacon's Part in the Intellectual Movement of His Time," pp. 21–32, both essays in *Seventeenth Century Studies Presented to Sir Herbert Grierson* (1938, reprinted 1967); KARL R. WALLACE, *Francis Bacon on Communication & Rhetoric; or, The Art of Applying Reason to Imagination for the Better Moving of the Will* (1943); VIRGIL K. WHITAKER, *Francis Bacon's Intellectual Milieu* (1962); BRIAN VICKERS, *Francis Bacon and Renaissance Prose* (1968); JAMES STEPHENS, *Francis Bacon and the Style of Science* (1975), a study of his rhetorical methods of communicating scientific knowledge; and BRIAN VICKERS (comp.), *Essential Articles for the Study of Francis Bacon* (1968).

(K.M.L./P.M.U.)

Bacteria

Bacteria are microscopic organisms present in almost all natural environments, often in extremely large numbers: billions in a gram of rich garden soil and millions in one drop of saliva. They constitute the class Schizomycetes, of the division Schizomycophyta. The influence of bacteria in the biosphere is incalculable. Without them, the soil would not be fertile and thus could not sustain plants, on which animals ultimately depend for food. Although some bacteria cause disease, most species are benign, and many are involved in processes of direct benefit to man. For the history and methodology of the study of bacteria, see BIOLOGICAL SCIENCES: *Microbiology.* See also DISEASE and INFECTIOUS DISEASES for coverage of bacterial diseases.

This article is divided into the following sections:

GENERAL FEATURES

Diversity of structure. Bacteria are unicellular microorganisms, among the smallest living creatures known (only the related rickettsias and the viruses are smaller). One of the periods on this page would cover about 250,000 average-sized bacteria, which are measured in microns (one micron, μ, equals $^1/_{1,000}$ millimetre, or $^1/_{25,000}$ inch). There are three bacterial cell types on the basis of shape (see Figure 1): spherical (coccus), rodlike (bacillus), and spiral (spirillum). Each type retains its character under standard conditions of laboratory cultivation but may show changes in appearance under different environments. When conditions are favourable, bacteria grow rapidly (some can reproduce every 15 minutes), forming visible colonies on culture plates in the laboratory. The colonies are distinctive in colour, shape, and texture for certain species and in some cases for the different varieties, or strains, within a single species.

Cell shapes

Distribution and abundance. Bacteria are ubiquitous, occurring in virtually every conceivable environment: from polar ice to hot springs; from mountain tops to the ocean deeps; from plant and animal bodies to forest soils. Most bacteria are active in environments in which the temperature is above 5° C (about 40° F); some marine and soil types are exceptional in being active at temperatures near or slightly below 0° C (32° F). The upper limit is around 30° C (86° F) for soil bacteria and 37° C (99° F) for animal parasites; the maximum temperature, above which growth does not occur, is about 70° C (160° F). Beyond these limits bacteria become inactive. Some survive in a dormant (or spore) state, reviving when conditions become more favourable. This capability has allowed bacteria to become perhaps the most widespread organisms on Earth.

Ecological significance

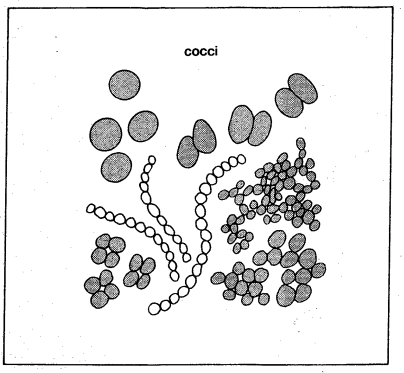
Importance. Bacteria are instrumental in performing numerous critical biochemical transformations of substances in nature, changing them from complex to simple compounds that can be used by plants, man, and other animals. Bacteria in the rumen (largest compartment of the stomach) of the cow digest the cellulose in grasses and animal feed, thus making this matter available as a nutrient for the animal. The organic waste substances in sewage are degraded by bacteria and transformed into compounds that are suitable nutrients for plant growth. In fact, all of the remains of animals and plants are eventually converted to soil through the activities of bacteria and other microorganisms and thus made available again to growing plants.

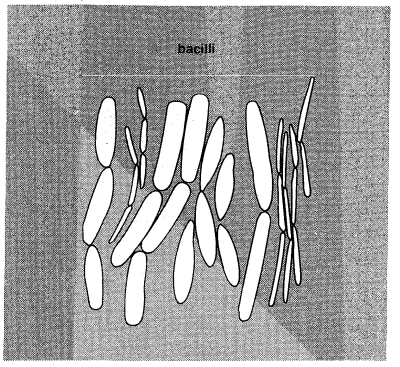
It can be assumed—until evidence disproves it—that any naturally occurring substance can be degraded (metabolized) by some species of bacteria. In some instances, these biochemical transformations are judged by man to be beneficial: as when *Streptomyces griseus* produces the antibiotic streptomycin as a by-product of its metabolism; or when certain strains of bacteria produce cheese and other dairy products by metabolizing constituents of milk; or when certain nitrogen-accumulating bacteria add needed nitrogen to the soil. Other transformations, however, may be detrimental to man's interests, as when *Clostridium botulinum* excretes a toxin responsible for botulism (a type of food poisoning), or when certain strains of bacteria excrete substances that cause disease. Less detrimental are the effects of bacterial deterioration: spoilage of food, corrosion of metals, decay of wood, and other undesirable alterations of substances. Certain strains of bacteria that can attack and partially decompose oil, tar, and related petroleum products may someday be useful as controls for these potential pollutants.

Higher animals, including man, live in constant intimacy with large numbers and a great variety of bacteria. The oral cavities, intestinal tracts, and skin are inhabited by bacteria that, under normal circumstances, create no problems (Figure 2). There are occasions, however, when bacteria break through the normal body barriers and cause an infection. Certain bacteria are more prone to this behaviour than others and are called pathogens, or disease producers (Figure 2). Some pathogens have an affinity for specific parts of the body: meningococcal bacteria infect the meninges, or brain membranes; tubercle bacteria invade the lungs; and diphtheria-causing bacteria establish themselves in the throat. Other bacterial pathogens exhibit less specificity: staphylococcal bacteria, for example, may infect the skin, causing boils or furuncles; the bloodstream, causing septicemia (blood poisoning); or the bones, establishing a condition known as osteomyelitis.

In some cases the biochemical transformations of bacteria are actually exploited industrially to convert raw materials into products that are economically valuable. When the substance is abundant and cheap and the product is valuable and in demand, a bacterial industrial process is likely to be established. Selected examples of products manufactured on an industrial scale through the utilization of bacteria are shown in Table 1. Other industrial processes are developed by man to combat the deterioration of materials by microorganisms (see FOOD PROCESSING).

Bacteria represent a form of life that can be conveniently studied under laboratory conditions. Furthermore, and more importantly, many of the biological processes that take place in the bacterial cell are closely related, if not





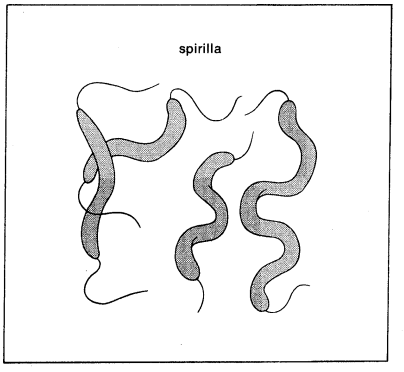


Figure 1: The three major bacterial cell types.

Bacteria as experimental subjects

identical, to processes that take place in higher organisms, including man. Thus, the bacterial cell provides an extremely useful model for the study of intricate biological, physiological, and biochemical processes. Indeed, much of the knowledge that has been acquired since the end of World War II in biochemical genetics, enzymes, and the synthesis of vitamins and their functions has resulted from investigations of bacteria.

NATURAL HISTORY

In view of the widespread occurrence of bacteria, it is not surprising that measures are taken by man to reduce or control their numbers in habitats that are exploited for human welfare. In the following sections, therefore, emphasis is on those particular environments of bacteria that impinge on man's needs.

Growth and reproduction. The life cycle of bacteria involves growth and reproduction, as in other organisms; certain features, however, are unique to microorganisms.

The term growth, in the microbiological sense, refers to increase in a given population rather than to increase in the size of an individual microorganism, or bacterium, in this context. There are, to be sure, changes in the size of an individual bacterial cell at certain stages of the multiplication process, as is shown below.

Generation time

Reproduction. Bacteria characteristically reproduce by an asexual process called binary fission, in which one cell divides into two new cells (Figure 3). A single bacterial cell, under optimum physical conditions for growth, performs metabolic functions including synthesis of intracellular substances. The cell elongates, and the cell wall becomes pinched in at the midpoint: finally, a transverse cell wall separates the parent cell into two new cells (daughter cells) that separate, and the process commences again. The reproduction and multiplication process is by geometric progression: one cell forms two, two cells form four, four form eight, eight form 16, 16 form 32, and so forth. The time required for the populations to double—*i.e.*, for one cell to divide in two—is the generation time (G); it can be calculated from the following formula:

$$G = \frac{t}{n} = \frac{t}{3.3 \log b/B}.$$

In the formula, B is the number of bacteria at the start of the investigation; b is the number of bacteria at the end of the time period; t is the time period; and n is the number of generations. The experimental data—that is, the values for B, b, and t—must be obtained during the period of the total growth cycle known as the log phase of growth (described below). The generation times of bacterial species vary over a wide range. *Escherichia coli,* one of the most rapidly growing bacteria, has a generation time of approximately 15 minutes; *Mycobacterium tuberculosis,* a slow-growing bacterium, can have a generation time as long as 16 hours.

Although binary fission is the characteristic and typical mode of reproduction for the true bacteria (order Eubacteriales), among some bacterial species, particularly those of the higher orders of bacteria, other processes of reproduction occur. *Rhodomicrobium vannielii* (order Hyphomicrobiales), for example, exhibits a budding type of reproduction; *Streptomyces* (order Actinomycetales) species produce chains of spores; *Mycoplasma* (order Mycoplasmatales) reproduce by the segmentation of elementary units within a body surrounded by a membrane.

Sexual reproduction

There are also instances of sexual reproduction (conjugation) among bacterial species. It occurs at low frequency among bacteria found in the intestine (enteric, or coliform, bacteria—*Escherichia, Shigella,* and *Salmonella*). In this process, conjugal pairs, or mating types, of bacteria make transient physical contact. Conjugal pairs consist of a donor (male) and a recipient (female) cell. In conjugation, a piece of chromosome from the donor cell is transferred into the recipient cell, in which it becomes a part of the recipient's chromosome. This is one way by which genetic material is exchanged in bacteria.

The bacterial growth curve. When bacterial cells are placed in a medium providing all of the nutrients necessary for growth (growth medium), the population increases according to a pattern identified as the bacterial growth curve (Figure 4). The four stages are the lag phase, log phase, stationary phase, and decline phase.

After their introduction into a medium, bacterial cells do not immediately reproduce according to their characteristic generation time; instead, the population remains constant for a period longer than the generation time. During

Table 1: Examples of Compounds Produced by Bacteria on an Industrial Scale

product	bacterium	substrate
Lactic acid	*Lactobacillus delbrueckii*	acid-hydrolyzed cornstarch, or whey, plus nutrient and $CaCO_3$
Bacterial amylase	*Bacillus subtilis*	vegetable protein plus sugar for surface cultivation; starch, cereal, grain, and protein for subsurface cultivation
Bacterial protease	*B. subtilis*	protein, carbohydrate, salts
Dextran	*Leuconostoc mesenteroides*	sucrose plus nutrients
Cobalamin (vitamin B_{12})	*Streptomyces olivaceus*	distiller's solubles, dextrose, $CaCO_3$
	Propionibacterium freudenreichii	$CoCl_2$
Vinegar	*Acetobacter* species	alcohol
Streptomycin	*Streptomyces griseus*	hydrolyzed protein and sugar
Monosodium glutamate	*Micrococcus* species	sugar

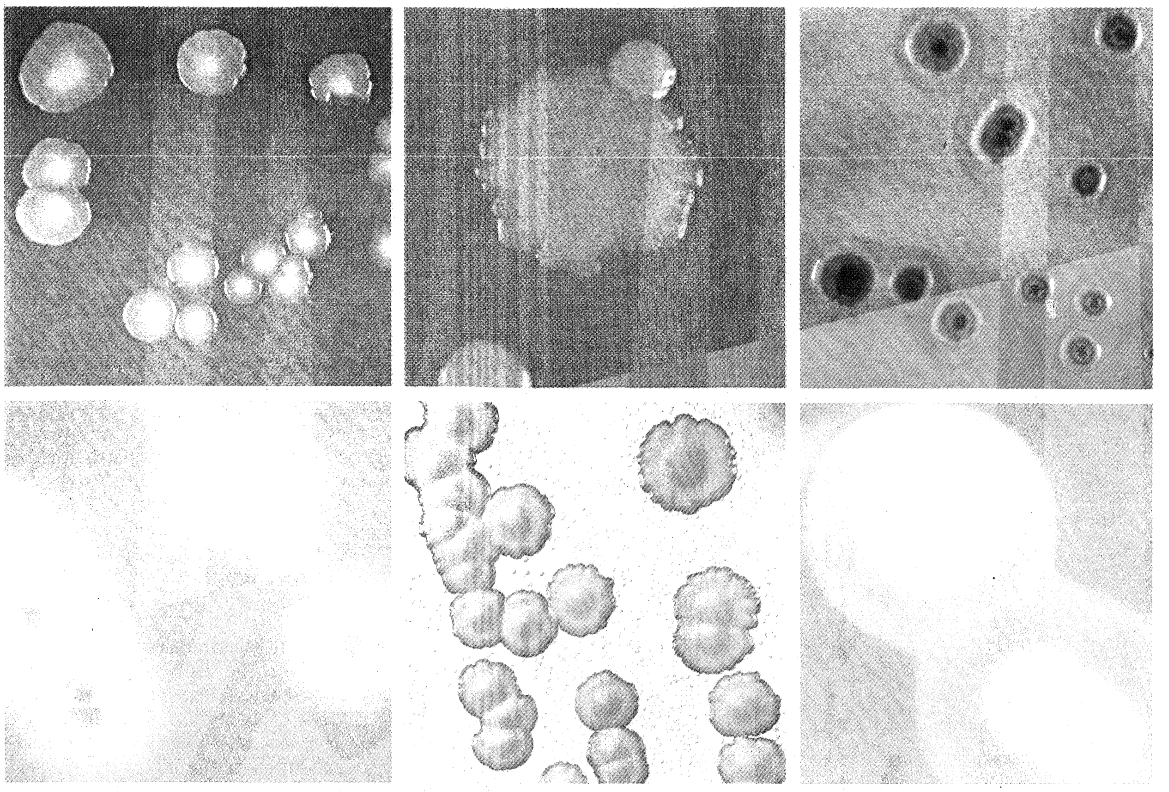


Figure 2: *Typical body flora (top row) and bacteria from infections (bottom row).* (Top left) *Neisseria flava* from the human nasal passage (magnified about 7 ×). (Top centre) Aerobic *Lactobacillus* (large colony) and *Staphylococcus albus* (small colony) from the human vagina (magnified 11 ×). (Top right) *Escherichia coli,* grown on eosin methylene blue dye, from a normal stool (magnified about 7 ×). (Bottom left) Colonies of beta-hemolytic streptococci on blood agar show zones cleared of blood cells surrounding each colony; these bacteria were isolated from the sore throat of a child who later developed rheumatic fever (magnified about 19 ×). (Bottom centre) *Hemophilus influenzae* (small colonies) isolated from a child suffering from spinal meningitis. Position of the *H. influenzae* colonies around larger *Staphylococcus* colonies, which provide a factor necessary for their growth, illustrates the "satellite phenomenon" (magnified about 14 ×). (Bottom right) *Clostridium perfringens,* isolated from infected human gallbladder, as grown under strict anaerobic conditions on blood agar. Zones of complete and partial clearing are characteristic of this species of *Clostridium* (magnified 8.5 ×).

A.W. Rakosy—EB Inc.

this period individual cells actively metabolize (synthesize new cytoplasmic material) and increase in size. After the cells have undergone a rigorous physiological adjustment to the new medium, they divide.

The first cell division initiates the log phase, during which repeated division occurs at a rate consistent with the generation time. Theoretically, this stage describes a logarithmic progression, the population doubling with each new generation of cells. In theory, also, all cells grow at the same rate and reproduce at the same time. This synchronous growth period eventually terminates as cells phase out of the regular reproduction pattern, and some actually die.

Under optimum conditions, in terms of the nutrients available and the physical environment, the maximum viable population is attained at the end of the log phase, with some vigorous bacterial species achieving a density of 10,000,000,000 cells per millilitre.

For a period of time, the length of which depends on the species, the population is stationary. Among the factors responsible for this levelling off of the population are the following, which generally occur in some combination: production of inhibitory substances, depletion of nutrients, and death of cells.

The stationary phase terminates as the death rate of the population exceeds that of formation of new cells. The population steadily decreases until, eventually, all of the cells die.

The coliform bacteria

Ecology. *Bacteria in water.* Good quality drinking water contains very few bacteria per millilitre and no coliform bacteria. The coliform bacteria, including *Escherichia coli* and *Aerobacter aerogenes,* are found in the intestinal tract of man and other animals. Their presence in water indicates that the water has been polluted with fecal material and hence may contain pathogens. The usual procedures employed in municipal water-purification plants—settling, filtration, and chlorination—are designed to remove or destroy these and other microorganisms.

Sewage, defined as the used water supply of a community, contains wastes produced by domestic and industrial sources. Bacteria in sewage are of importance for two reasons. First, as pollutants: human excrement in sewage may contain pathogenic bacteria, which, if not removed or killed, may enter domestic water sources or supplies (Figure 5). Second, as cleansers: the treatment of sewage, particularly the breakdown (dissimilation) of organic material (*e.g.,* proteins, carbohydrates, and fats) requires the activity of certain types of bacteria. Sewage treatment facilities, including residential septic tanks, municipal sludge digesters, activated sludge digesters, and trickling filter and sand filter processes, are all designed to utilize bacteria to break down organic matter in sewage.

The breakdown of organic matter, however, imposes a biochemical oxygen demand (BOD) on the environment into which sewage is dumped (*e.g.,* a body of water). The greater the amount of organic matter, the greater is the amount of oxygen required for its oxidation and, hence, the greater the BOD. This process can be very disruptive to aquatic life in natural streams and lakes. One of the objectives in sewage treatment is to oxidize organic matter as completely as possible and thereby reduce the BOD prior to the discharge of the sewage (effluent) into natural bodies of water. Sewage digestion tanks and aeration devices are designed to exploit the metabolic capacity of bacteria to accomplish this objective.

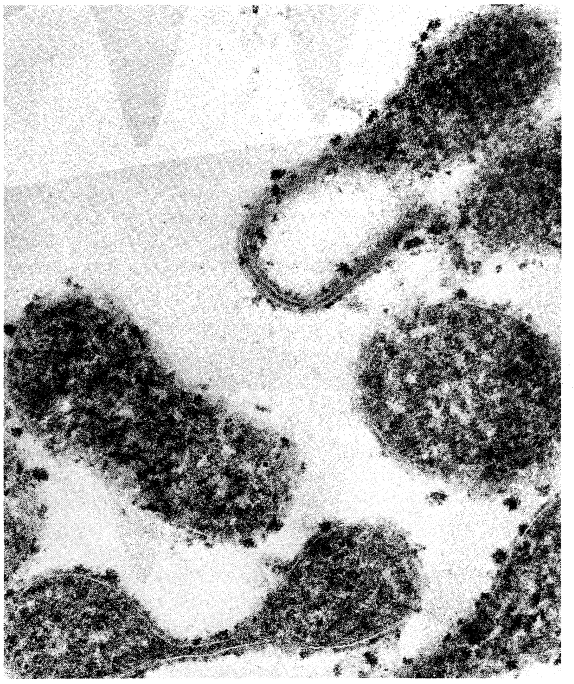


Figure 3: *An electron micrograph of cells of Mycoplasma hominis in several stages of reproduction.*
(Centre left) The elongated cell about to undergo binary fission. (Bottom) The dividing cell connected by a tubule, which, at its thinnest, consists only of the two membranes joined back to back. (Centre right) The coccoid-shaped daughter cell (greatly magnified).
From Jack Maniloff, "Electron Microscopy of Small Cells: *Mycoplasma hominis*," *Journal of Bacteriology* (December 1969)

The microbial population of the sea consists of bacteria, algae, protozoans, and fungi. Bacteria of all physiological and metabolic types inhabit the various regions, extending from the surface layer of the sea to the bottom mud. They are responsible for transformations of both organic and inorganic compounds that serve as nutrients for marine life. The dissimilation of organic compounds, under aerobic (oxygenated) conditions, yields ammonia, carbon dioxide, and sulfate and phosphate salts. These products serve as the nutrients for algae and other planktonic life, which, in turn, synthesize organic compounds that may eventually serve as food for mollusks and fishes.

Bacteria in air. Air contains bacteria and other microorganisms that are suspended and circulated for varying periods of time, depending upon atmospheric conditions and the size of the particles that carry the microorganisms.

Extremes of bacterial concentration

Generalizations about the microbial component of air can be made only with reference to a particular environment and the circumstances that prevail at a given time: for example, a hospital ward during bed-making time (agitation of bed linens and movements of personnel stir dust that may bear large numbers of bacteria into the air); a city street following a heavy rain (the air is washed and relatively free from bacteria). Air at an altitude of 10,000 feet (3,000 metres) usually has relatively few dust particles and, therefore, considerably fewer bacteria than are common in air at lower altitudes.

Tremendous numbers of bacteria are ejected into the atmosphere by a sneeze or cough. They remain suspended in air on particles referred to as droplet nuclei, which may consist of a single bacterium—in which case the particle remains airborne for a long period of time. When the particles consist of aggregates of bacteria coated with mucus or affixed to other cells, they settle out on surfaces in a short time. Airborne bacteria from the respiratory tract of man are potentially hazardous; they include *Mycobacterium tuberculosis,* which causes tuberculosis; *Neisseria meningitidis,* which causes meningitis; *Streptococcus pyogenes,* which causes strep infections; and *Diplococcus pneumoniae,* which causes pneumonia.

Bacteria occur on plant surfaces and will grow if conditions are favourable. They also invade and infect plant tissues, from which they are dispersed periodically into the atmosphere. Soil, particularly rich garden soil, contains countless bacteria per gram; soil dust dispersed into the air contributes a multitude of bacteria and resistant bodies formed by them (spores).

Airborne bacteria, as already mentioned, are important in the dissemination of diseases of man, of other animals, and of plants. They are also important as the source of contamination of many materials—*e.g.,* pharmaceutical preparations, surgical devices, and foods—thus necessitating precautionary procedures when "bacteria-free" conditions are desired (see below *Control*).

Bacteria in soil. Soil bacteria are extremely active in effecting biochemical changes in soil, which is the repository of the remains of plant and animal life; through microbial attack, materials are eventually transformed into the very substances that characterize soil—humus and minerals.

It is customary to view the chemical changes performed by bacteria in the soil as cyclic processes of the various elements; *e.g.,* the nitrogen, carbon, and sulfur cycles. At some stage of each cycle the element exists in its elemental form; *i.e.,* uncombined with any other elements. Certain species of bacteria are capable of converting the element into an inorganic compound that can be utilized as a plant nutrient and thereby transformed into an organic compound. The plant is consumed by animals, and the organic compound is incorporated into animal tissue. Eventually, animal and plant tissues return to the soil, where bacteria decompose them and thereby again release the elements as well as other products (see BIOSPHERE).

Nitrogen fixers

The nitrogen cycle serves to illustrate the role of bacteria in performing various chemical changes. Nitrogen fixation, one of the key transformations in the nitrogen cycle, is performed by bacteria of several physiological types. Nitrogen-fixing bacteria are capable of transforming raw nitrogen into a form suitable for use by plants. They live in intimate association with a leguminous plant, in which case they are designated symbiotic nitrogen fixers; this type lives within lumps (nodules) on the plant's root system. Species of bacteria that can fix nitrogen in soil that is free from the plant-root system are referred to as nonsymbiotic

(Top right, bottom left, bottom right) A.W. Rakosy—EB Inc.

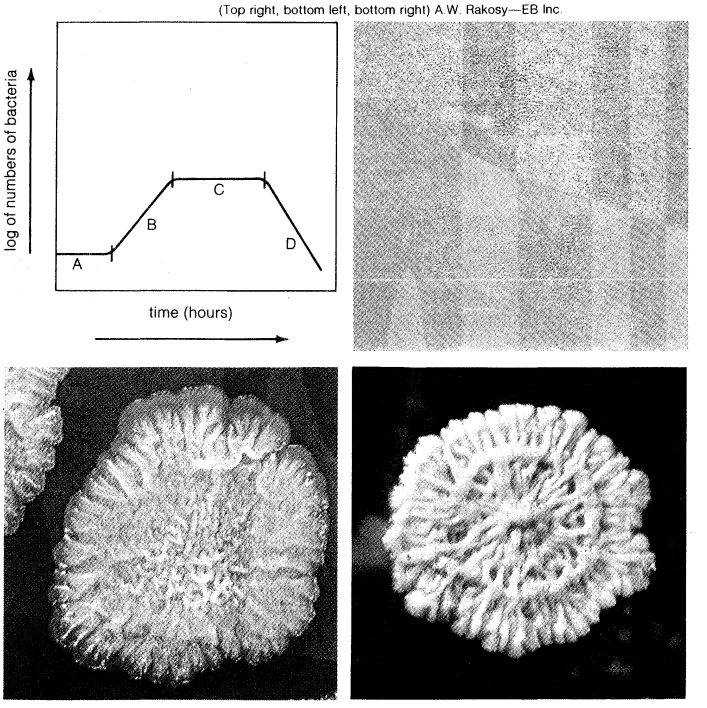


Figure 4: *Bacterial growth.*
(Top left) Generalized bacterial growth curve showing: (A) lag phase, (B) log phase (period of logarithmic or exponential growth), (C) stationary phase, and (D) death, or decline, phase. Sequence of bacterial colony growth in *Bacillus subtilis* grown at 37° C at: (top right) 18–24 hours (magnified about 6 ×), (bottom left) 48 hours (magnified about 9 ×), (bottom right) 96 hours (magnified about 9 ×).

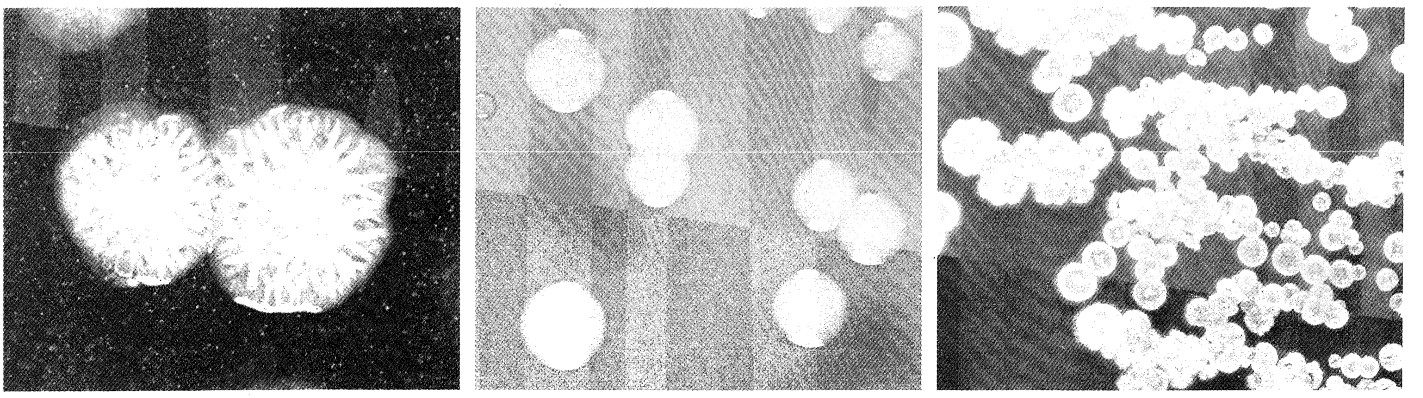


Figure 5: *Colonies of bacteria from water, dust, and soil.*
(Left) Rough colonies of *Pseudomonas pseudomallei* isolated from surface waters in Southeast Asia. These bacteria cause systemic meliodosis, a lethal infection in man. (Centre) Smooth colonies of *Micrococcus luteus*, on blood agar, isolated from dust particles from a clean room. (Right) Rough colonies of *Streptomyces griseus* isolated from soil (magnified 13 ×).
A.W. Rakosy—EB Inc.

nitrogen fixers. The best known genera are *Azotobacter* and *Clostridium,* although many others are now known to be capable of nonsymbiotic nitrogen fixation.

Bacteria in food. Milk drawn from a healthy cow contains relatively few bacteria per millilitre, but it is not sterile (free of living bacteria). Furthermore, procedures for handling the milk may add additional bacteria through contamination. Because milk is an excellent medium for the growth of bacteria, an initial small population (inoculum) can increase rapidly if the milk is not properly processed. Bacteria may merely cause spoilage or may present a serious health hazard if they are pathogenic. Bacterial pathogens transmitted through milk may originate in the cow or in man. A cow infected with the tubercle bacillus may transmit this organism, through milk, to man. Brucellosis, or undulant fever, also can be transmitted from cows to man via milk. Alternatively, an infected milk handler can contaminate the milk; outbreaks of typhoid fever, scarlet fever, and diphtheria have been known to occur in this fashion. Proper treatment of milk by the process of pasteurization—either the low-temperature, holding method (63° C [145° F] for 30 minutes) or the high-temperature, shorttime method (72° C [161° F] for 15 seconds)—destroys all pathogens.

Bacterial conversion of milk

On the credit side, however, selected species and strains of bacteria convert milk and casein (milk protein) into such desirable products as buttermilk, yogurt, and cheese. Commercial cultured buttermilk is prepared from skim milk that has been inoculated with a starter culture and allowed to incubate until the desired changes occur. The starter culture consists of *Streptococcus lactis* or *Streptococcus cremoris,* together with *Leuconostoc citrovorum* or *Leuconostoc dextranicum.* Yogurt and other fermented milk products are produced in a similar manner but through the activities of different selected cultures of bacteria.

The formation of cheeses is likewise dependent upon the activity of microorganisms. The curd from which cheese is made is precipitated (made to settle out) from milk by an acid-producing bacterium, such as *Streptococcus lactis.* Following removal of moisture and the addition of salt, the curd is allowed to ripen by the action of selected bacteria. Lactobacilli, streptococci, and propionibacteria are important for the ripening of Swiss cheese; *Brevibacterium linens* is responsible for the flavour of Limburger cheese; and molds (*Penicillium* species) are used in the manufacture of Roquefort and Camembert cheeses.

Bacteria in nondairy foods are as significant as those in milk and dairy products. The variety of bacteria that contaminate foods and the diversity of foods can result in a wide array of types of food spoilage. When allowed to grow in food, certain bacteria can cause food poisoning; they secrete a toxin that, when ingested by humans, can cause either a severe gastrointestinal upset—as in *Staphylococcus aureus* food poisoning—or death—as in botulism (caused by the toxin of *Clostridium botulinum*). It follows that one of the major concerns of food microbiology is the development and assessment of techniques to preserve foods from spoilage and contamination.

Food may be the carrier of pathogenic bacteria and thus be responsible for food-borne infections, among which the more frequently occurring are typhoid fever (*Salmonella typhosa*); salmonellosis (*Salmonella* species other than *S. typhosa*); and shigellosis, or dysentery (*Shigella dysenteriae*).

Notwithstanding the detrimental effects of food contamination, other bacterial populations are responsible for a variety of special foods, produced through bacterial fermentation; these include pickles and other pickled products, sauerkraut, and olives.

Degrees of control

Control. Many materials and products, as well as certain environmental areas, require either the reduction or destruction of microbial populations. In a hospital operating room, for example, procedures are observed that reduce the microbial contents of the air to a very low level; on the other hand, the glucose solution used for intravenous injection is processed to be sterile—absolutely free of any form of life.

Physical and chemical means are available to accomplish sterilization. The method of choice depends upon several considerations, not the least of which is the effect of the sterilizing procedure on the object being sterilized. If the material being sterilized is to be discarded following sterilization, such as used bacteriological media from a laboratory, then there is no problem except effectiveness of the sterilization procedure; however, if the material to be sterilized is a vitamin solution, for example, the procedure must be effective enough to sterilize without affecting the quality of the vitamin product.

Heat. High temperature, applied in a variety of ways, is one of the most effective sterilizing agents. Heat may be applied in the form of incineration, steam under pressure (autoclave), or dry heat (hot-air oven).

Incineration procedures range from the passing of an object through a bunsen burner flame in the laboratory to the burning of infected animal carcasses or contaminated bedding in large furnaces.

Steam under pressure, which is the principle of operation of the autoclave (an elaborate pressure cooker), is perhaps the most widely used sterilization procedure. The autoclave is a standard item of equipment in laboratories, hospitals, industries involved in food processing, and enterprises concerned with sterilization procedures and the manufacture of sterile products.

Dry heat, as in hot-air ovens, also accomplishes sterilization, but, in contrast to steam heat, higher temperatures and longer exposure times are required. Equipment that can be sterilized in an autoclave at 121° C (250° F) within ten to 15 minutes requires 160° C (320° F) for a period of two hours in a hot-air oven. Hot-air sterilization is used

for materials that might be adversely affected by moist heat or that should not be directly exposed to moist heat because of necessary packaging.

The heat afforded by boiling water and pasteurization processes markedly reduces the bacterial flora but does not truly sterilize. Pasteurization, as previously mentioned, is designed to kill only the serious pathogens that might occur in milk; some bacteria and, in particular, bacterial spores are not killed. Similarly, boiling water kills the vegetative (active) bacterial cells but not necessarily the spores.

Anti-bacterial action of ultra-violet radiation

Radiation. Ultraviolet radiation (in the 2650-angstrom region [one angstrom = 10^{-8} centimetre]) is extremely bactericidal; that is, capable of killing bacteria. When properly used—namely, under conditions that allow direct exposure of organisms to the radiation—the microbial population can be effectively reduced. Ultraviolet rays, however, have a very low order of penetration; a thin film of glass filters out most of the rays.

Gamma radiations, which are emitted from radioactive isotopes, have great penetrating power as well as a high lethal effect. This combination of characteristics—high penetration and high bactericidal activity—makes them extremely effective as sterilizing agents. But several technical problems are associated with practical applications of gamma radiations: the development of an adequate supply of radiation sources and the availability of equipment designed to guarantee safety to its operators.

Certain materials cannot be exposed to physical agents without being adversely altered in some manner. Many medicines, including antibiotics, and other biological solutions may be destroyed or inactivated by any of the methods described so far. In such cases filtration is the appropriate method for sterilization. A wide variety of filters are available that are porous enough to allow fluids to pass through but not microorganisms.

Ethylene oxide as a sterilizing agent

Chemicals. Sterilization can also be accomplished with such chemicals as the gas ethylene oxide. With the appropriate concentration, humidity, and temperature, ethylene oxide is a powerful sterilizing agent. In addition, it has the ability to penetrate considerably through materials. Ethylene oxide is widely used for the sterilization of many materials, including plastic devices, that could not undergo the other procedures.

Many chemicals for the control of microorganisms are available for application to environmental surfaces or materials; they are not intended to effect sterilization, but they do reduce the microbial population or eliminate certain types of microorganisms. These chemical agents are variously termed antiseptics, germicides, disinfectants, and sanitizers on the basis of their action.

FORM AND FUNCTION

In normal environments bacteria exist not only in large numbers but also in great diversity. In order to understand the characteristics of the individual species comprising these mixed populations, it is necessary to study each species as a pure culture, by isolating bacterial cells from colonies on a specific nutrient referred to as the medium. This isolation from the colony can be maintained as a pure culture by periodic transfer to fresh medium. Alternatively, a pure culture can be lyophilized (dried while frozen and sealed under vacuum) and kept viable (capable of dividing) in this condition for many years. Characterization of a pure culture requires a study of cell morphology, cultural and physiological characteristics, biochemical (or metabolic) processes, antigenic characteristics, and pathogenicity.

Morphological features. *The bacterial cell.* The typical cell of a species of Eubacteriales, or "true" bacteria, is a bacillus approximately one micron in diameter and a few microns in length. Coccoid cells may occur in characteristic arrangements: *e.g., Diplococcus pneumoniae* occurs in pairs; *Staphylococcus aureus* occurs in grapelike clusters; *Streptococcus pyogenes* occurs in chains; *Sarcina lutea* occurs as a cuboidal arrangement of cells; and *Micrococcus tetragenus,* as the name suggests, occurs in tetrads, or groups of four cells.

In contrast to the true bacteria are the "higher" bacteria, whose appearance suggests some primitive or abortive attempt toward cellular differentiation. They are usually much larger than the true bacteria and often bear some resemblance to yeasts, molds, algae, or protozoans.

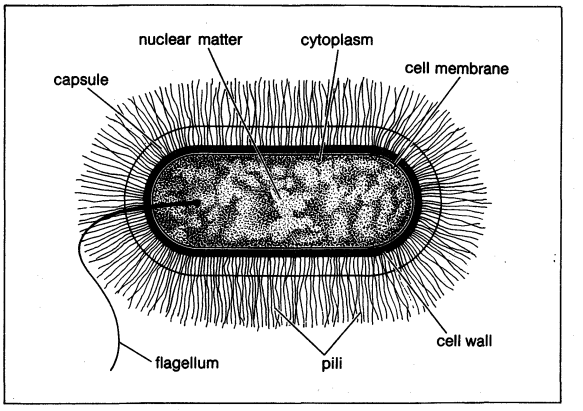


Figure 6: Schematic drawing of structure of a typical bacterial cell of the bacillus type.

The typical bacterium

A typical bacterial cell is shown in Figure 6; not all bacterial species possess all of the structures shown. All bacterial cells contain nuclear substance, but it is not organized into a discrete nuclear structure as in higher organisms; there is no nuclear membrane, and mitotic division, the mechanism by which the cells of plants and animals divide, does not occur. The blue-green algae are akin to bacteria in this respect; together they are characterized as procaryotes, in contrast to eucaryotes, the cells of which possess a discrete nuclear structure.

All bacterial cells possess a membrane and a cell wall. There are distinct differences among species in terms of the chemical composition of these structures, however. The cells of some bacterial species possess whiplike structures called flagella, the number and arrangement of which are typical for a species. Shorter, rigid appearing, spikelike projections known as pili, or fimbriae, appear on some cells. Certain cells are surrounded with a gelatinous or slimy material, the capsule (Figure 7). Many bacteria can assume a dormant state as a spore; in fact, during sporulation the entire vegetative cell, in essence, is transformed into the spore body.

From W.H. Taylor and E. Juni, "Pathways for Biosynthesis of a Bacterial Capsular Polysaccharide," *Journal of Bacteriology* (May 1961)

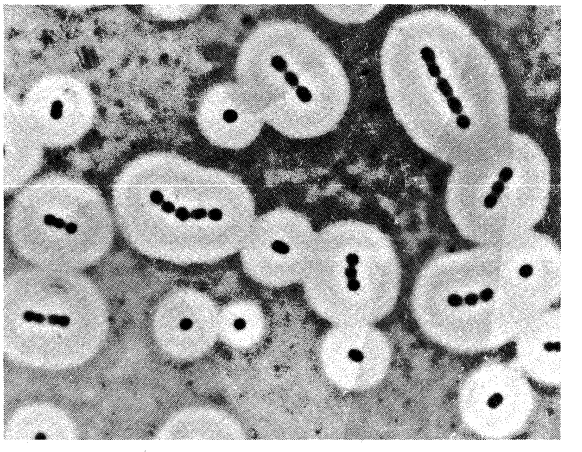


Figure 7: Capsular material surrounding these bacteria (*Acinetobacter calcoaceticus*) is revealed in a suspension of India ink and viewed through the light microscope (magnified about 2500 ×).

Staining techniques are used to demonstrate the various bacterial cell structures. One of the most important staining procedures involves the gram stain (named after its inventor, H.C.J. Gram, a Danish physician). This is a differential stain; *i.e.,* bacteria are said to be gram positive or gram negative depending upon whether they retain the purple colour of the original stain (crystal violet) at the end of the procedure, or whether it is washed out and the

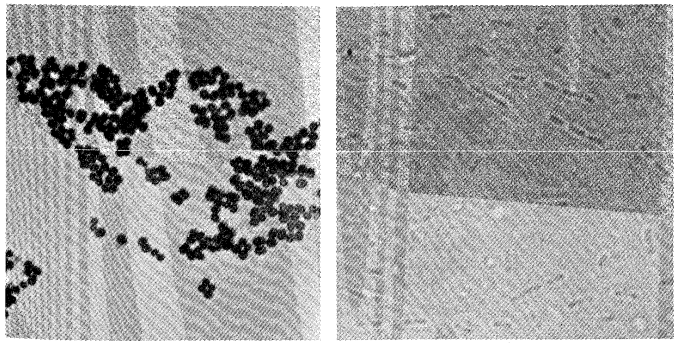


Figure 8: *Bacteria isolated and coloured with gram stain.*
(Left) Gram positive cocci, *Staphylococcus aureus,* from a laboratory culture. (Right) Gram negative bacilli with a capsule, *Klebsiella pneumoniae,* from a pneumonia lung abscess (magnified 1000 ×).
A.W. Rakosy—EB Inc.

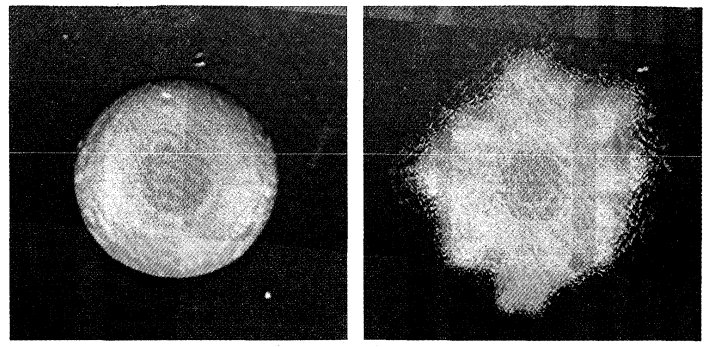


Figure 9: *Colonial morphology.*
Citrobacter freundii displaying two forms: (left) smooth colony, (right) rough colony.
Rhodes Scherer, National Animal Disease Laboratory, Ames, Iowa

red colour of the counterstain (safranin) shows (Figure 8).

More refined characterization of the anatomy of a bacterial cell can be accomplished by electron microscopy, in which bacterial cells are sliced into very thin sections and then viewed under very high magnification; electron micrographs reveal a great deal of complex detail: layers in the cell surface, internal structures, as well as connections or continuity between certain structures.

The bacterial colony. The gross appearance of bacterial growth in or on media defines the cultural characteristics of a species. When a specimen containing bacteria is inoculated onto an agar medium (a standard preparation of nutrients with a gel-like consistency), colonies develop, each from an isolated bacterial cell. Such colonies (Figure 9) vary in characteristics depending upon the species. They may be pinpoint in size or several millimetres across; flat or raised and convex; smooth or broken edged; stringy, brittle, or buttery in consistency; coloured internally or excreting substances that colour the surrounding medium; and opaque, transparent, or translucent.

Bacteria cultivated as agar slant cultures (slanted medium in test tubes) exhibit differences in characteristics much like those described for colonial appearance (Figure 10). In liquid media (broth), growth may be confined to the surface as a film (pellicle); uniformly distributed throughout the liquid; or particulate, with a tendency to form a sediment.

Physiological features. *Physical requirements.* On the basis of temperature requirements for growth, bacteria are grouped into three categories: psychrophiles, 0° to 30° C (32° to 86° F); mesophiles, 15° to 45° C (59° to 113° F), with best growth in the range 25° to 40° C (77° to 104° F); and thermophiles, 45° to 60° C (113° to 140° F) and above. Bacteria within each group exhibit specific minimum, maximum, and optimum temperatures for growth; for example, *Pseudomonas delphinii* grows at 1° C and 30° C (34° F and 86° F), with optimum growth occurring at 25° C (77° F), and *Bacillus thermoliquefaciens* grows at 37° C and 70° C (99° F and 158° F), with optimum growth occurring at 60° C (140° F).

Temperature categories of bacteria

Atmospheric oxygen is required by some, but not all, bacteria; others are inhibited by its presence. Bacteria are classified as aerobes when they require oxygen to grow and as anaerobes when they cannot grow in the presence of oxygen; facultative anaerobes do not require oxygen and can grow in its presence.

Some bacteria grow in a wide range of salt concentrations; others, such as marine bacteria, require salt levels of 10 to 15 percent for optimal growth. Most bacteria grow best in an environment near neutrality (neither acidic nor alkaline); some, however, grow under strongly acid conditions (*Thiobacillus thiooxidans*) and others under strongly alkaline conditions (*Nitrobacter* species). Most bacteria, especially those intimately associated with man, grow best without light. Photosynthetic bacteria, however, require light.

A.W. Rakosy—EB Inc.

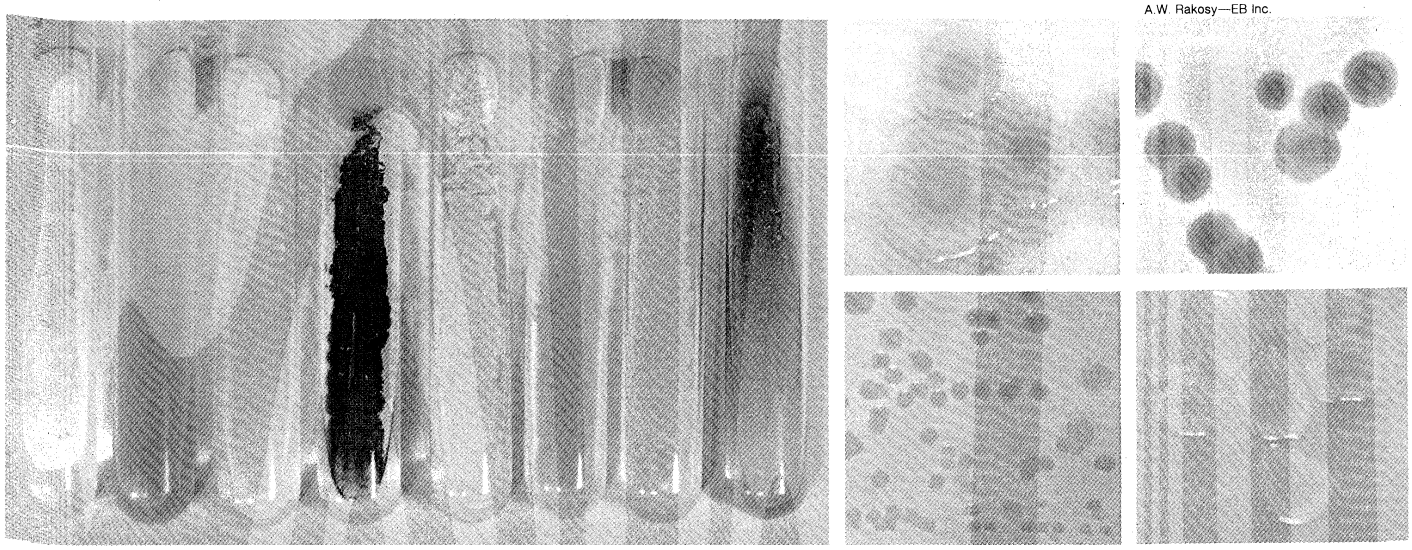


Figure 10: *Chromogenic bacteria.*
(Left) Agar slant cultures, left to right: *Sarcina lutea; Pseudomonas aeruginosa,* from water in a hospital humidifier; *Pseudomonas fluorescens; Chromobacterium violaceum; Sarcina aurantiaca; Serratia marcescens,* from soil; *Staphylococcus albus; Pseudomonas aeruginosa* from the urine of a person with cystitis. Agar colonies: (top centre) *Pseudomonas aeruginosa* isolated from sputum, (top right) *Chromobacterium violaceum,* from soil, (bottom centre) *Serratia marcescens,* from inadequately cleaned eating utensil, (bottom right) *Pseudomonas aeruginosa* from urine.

Nutritional requirements. The establishment of the specific nutritional requirements for various bacterial species required the development of media consisting of known chemical compounds in prescribed amounts, on which the bacteria could be cultivated in the laboratory. In a broad sense, bacteria, on the basis of nutritional requirements, can be divided into two groups: autotrophs and heterotrophs. Autotrophs are capable of growing entirely on inorganic ingredients, with atmospheric carbon dioxide as the exclusive source of carbon. Heterotrophs require an organic form of carbon and, in addition, may require other organic substances, such as amino acids and vitamins.

The autotrophic sulfur-oxidizing bacteria can grow on powdered sulfur, ammonium sulfate, potassium phosphate, calcium chloride, magnesium sulfate, ferric sulfate, water, and carbon dioxide. From these relatively simple chemical substances the autotrophic bacterium can synthesize and organize the vast array of complex organic substances of which it is made.

Heterotrophs exhibit great diversity in their nutritional requirements. *Escherichia coli,* the most frequently studied heterotrophic species, can grow on the ingredients listed above for autotrophic sulfur-oxidizing bacteria, except that carbon from an organic source (such as the sugar glucose) is also essential. *Salmonella typhosa* has a requirement slightly different from that of *E. coli* in that not only glucose but also an amino acid (tryptophan) is required. *Staphylococcus aureus,* another typical heterotroph, requires several amino acids and at least one vitamin, thiamin. *Lactobacillus* species are considerably more fastidious; they require many amino acids, several vitamins, and purine and pyrimidine compounds.

Some bacteria, especially certain pathogens, such as the syphilis bacterium (*Treponema pallidum*), require a complex medium that is provided only by living animal tissue. Such bacteria have so far not been cultivated even in a complex bacteriological medium of peptone (a digest of animal protein) and meat extract (or meat broth), with supplemental substances such as blood, serum, and other animal fluids.

Biochemical activities. Bacteria are capable of carrying out many chemical changes, both in the breakdown of complex substances and in the synthesis of new products. A compound may be degraded to different end products by different bacterial species, just as the same nutrients may be synthesized into different substances by different species.

Decomposition of carbohydrates

Dissimilation. Complex organic carbon compounds, such as pectin, cellulose, and starch, are readily degraded by many bacteria. Pectins are complex carbohydrates that occur in plants. Enzymes produced by species of *Erwinia, Bacillus,* and *Clostridium* are capable of converting pectin to galacturonic acid and further to sugar and other products. Cellulose, the major constituent of plant tissue, is transformed to dextrins and then to glucose by species of *Cellulomonas, Cytophaga, Streptomyces,* and *Clostridium.* Several *Bacillus* and *Clostridium* species produce enzymes that convert starch to sugar.

After the complex carbohydrates have been broken down into their constituent units (*e.g.,* glucose), these units are utilized in a variety of ways—*e.g.,* to produce a variety of other compounds. Bacterial species are characterized in part on the basis of the characteristic products they form from glucose. Table 2 lists several genera of bacteria, together with representative products of their glucose dissimilation.

Proteinases

The process of protein dissimilation (proteolysis) is accomplished by enzymes known as proteinases, which are produced by species of such genera as *Clostridium, Bacillus, Proteus,* and *Pseudomonas.* The proteinases decompose the protein molecules, hydrolyzing (breakdown involving water) the linkages (peptide bonds) between the amino acids that constitute the molecule. The result is the formation of peptides, chains of amino acids. Enzymes called peptidases then break down the peptides into individual amino acids. In turn, the amino acids may be degraded through several processes; deamination, the removal of the amino group ($-NH_2$) by enzymes called deaminases, the end products being ammonia (NH_3) and a fatty acid; decarboxylation, the removal of the carboxyl group ($-COOH$) by enzymes called decarboxylases, the end products being carbon dioxide and an amine; or amino acid fermentation, a breakdown more extensive than either deamination or decarboxylation, and resulting in more complex and varied end products, depending on the amino acid being degraded.

Table 2: Bacteria Grouped According to Major Products of Glucose Dissimilation

groups (with examples of some genera)	representative products
Lactic acid bacteria	
Streptococcus *Lactobacillus* *Leuconostoc*	lactic acid only or lactic acid plus acetic acid, formic acid, and ethyl alcohol; those species producing only lactic acid are *homofermentative;* those producing lactic acid plus other compounds are *heterofermentative*
Propionic acid bacteria	
Propionibacterium *Veillonella*	propionic acid plus acetic acid and carbon dioxide
Coli–aerogenes–typhoid bacteria	
Escherichia *Aerobacter* *Salmonella*	formic acid, acetic acid, lactic acid, succinic acid, ethyl alcohol, carbon dioxide, hydrogen, 2,3-butylene glycol (produced in various combinations and amounts depending on genus and species)
Acetone, butyl-alcohol bacteria	
Clostridium *Butyribacterium* *Bacillus*	butyric acid, butyl alcohol, acetone, isopropyl alcohol, acetic acid, formic acid, ethyl alcohol, hydrogen, and carbon dioxide (produced in various combinations and amounts depending on species)
Acetic acid bacteria	
Acetobacter	acetic acid, gluconic acid, kojic acid

Source: M.J. Pelczar, Jr., and R.D. Reid, *Microbiology* (1972).

Fats, or lipids, composed of fatty acids and glycerol (*i.e.,* triglycerides), can be degraded by such bacteria as *Pseudomonas, Proteus, Achromobacter, Alcaligenes, Bacillus, Micrococcus,* and *Clostridium.* The end products of the dissimilation of fats are glycerol and fatty acids, which can be broken down further.

Synthesis. The extensive capacity of a bacterium to synthesize complex molecules is shown when the simple chemical substances upon which an autotroph grows are compared to the complex chemical composition of the bacterial cells that are produced. In addition to the large, complex molecules that constitute cell structure, other substances of significance are elaborated by bacterial cells, including antibiotics, pigments, toxins, and polysaccharides (complex carbohydrates).

Virulence

Pathogenicity. It was implied above that a pathogen has the potential to produce a disease in a given host species—a plant, an animal, or even another bacterium species. The power of a bacterium to cause a disease, termed virulence, varies within a species. Some strains are highly virulent; *i.e.,* a small number of cells can establish the infection. Other strains have a low degree of virulence and produce infection only when transmitted in massive numbers. Furthermore, some strains of a pathogen may lose their virulence. A quantitative expression of virulence can be determined by performing laboratory experiments to establish the number of bacterial cells required to infect or kill a standard experimental animal. The expression lethal dose$_{50}$ (LD_{50}) refers to the number of bacteria necessary to kill 50 percent of the organisms inoculated.

Factors that contribute to the virulence of a bacterial pathogen are not yet completely understood, except in those instances in which the pathogen causes damage by secreting a potent toxin (an exotoxin). Diphtheria bacteria (*Corynebacterium diphtheriae*), for example, establish themselves in the mucous membrane of the upper respiratory tract of man; as they grow, they produce an exotoxin that is absorbed by the mucous membranes and causes the death of cells in the host. It is the exotoxin that does the damage to the host; a strain of *C. diphtheriae* that does not produce the exotoxin cannot cause diphtheria.

Not all virulent pathogens produce an exotoxin; most,

in fact, do not. *Salmonella typhosa,* the typhoid fever bacterium, produces an endotoxin, a complex substance associated with, or bound to, the surface structures of the bacterial cell. Endotoxins damage host tissues and host metabolism in ways not clearly understood. Other substances that contribute to the virulence of various pathogens are described in Table 3. (M.J.P.)

Table 3: Substances Contributing to the Virulence of Pathogenic Bacteria

substance	action
Hyaluronidase	increases permeability of tissue spaces to bacterial cells
Coagulase	increases resistance of bacteria to phagocytosis (engulfment by defense cells, or phagocytes)
Hemolysins	destroy red blood cells
Collagenase	dissolves collagen, a connective tissue protein
Leucocidin	kills white blood cells (specifically leucocytes) and hence decreases phagocytic action
Exotoxins and endotoxins	interfere with normal metabolic processes

Antigenic features. When bacterial cells enter the tissues of an animal, it is likely that they will act as antigens, agents that evoke the production of substances called antibodies. Antibodies are produced by an animal as part of its immunological defense against any foreign substance that can threaten its welfare.

Antigen–antibody reactions are highly specific and highly sensitive. A bacterial cell consists of many different antigens (*e.g.,* flagellar, capsular, and cell wall antigens). The antigen pattern in one bacterial species differs from that in another. Although it is not uncommon for related species (*e.g., Salmonella typhosa* and *Salmonella paratyphi*) to share certain antigens, each also contains other antigens that are unique.

Variability. The characteristics of one bacterial species are sufficiently definitive and constant to delineate it from other species. This distinctiveness does not mean, however, that each characteristic is evident and manifest by all strains of all species under all conditions.

nviron-
ientally
nduced
ariation

Reversible changes. Bacteria, under uniform conditions, will manifest a constancy of characteristics. If the same species of bacteria is placed in different environments—different physical conditions or different chemical composition of medium—the resultant growth may differ. In fact, the morphological and physiological characteristics are not identical at all stages of the growth curve of a bacterium. The morphology of cells from an "old" culture differs from that of "young" cells. In addition, the formation of a capsule is significantly influenced by the composition of the medium; bacteria that, in a nutrient broth, exhibit no capsules may produce large capsules when grown in milk. Some of the enzymes produced by bacteria are produced only when the compound on which the enzyme acts (substrate) is present; they are called adaptive enzymes, in contrast to constitutive enzymes, which are produced irrespective of the substrate.

Changes of the types just described are transient; they reflect what occurs during a stage of growth or in response to a change in the environment. The genetic endowment (genotype) of the organism remains the same, regardless of the different expressions (phenotype) that are seen under different environmental circumstances.

Permanent changes. Daughter cells that contain genetic information different from that of the parent cell constitute a new genotype. Such permanent-type genotypic changes may occur through four different processes: mutation, conjugation, transformation, and transduction.

Mutation involves a sudden alteration of a gene that is inherited by subsequent generations. Some bacterial cells in the process of normal growth undergo mutation; however, the ratio of mutant cells to unchanged cells is very small. The number of mutants in a population can be greatly increased by exposing the bacteria to certain physical agents (for example, ultraviolet rays and X-rays) or to chemicals (for example, mustard gas and organic peroxides). Bacterial mutants exhibit alterations in nutrition, drug resistance, pigmentation, and colonial form, among other characteristics.

Conjugation (sexual reproduction), a rarity in bacteria, results in recombinations of genes in the cells that pair (see above *Reproduction*).

Transformation also involves the transfer of genetic information from one cell to another, but pairing does not occur. Deoxyribonucleic acid (DNA) that is released from one cell (the donor) is taken up by another cell (the recipient) and incorporated into the genetic apparatus of the latter.

Changes caused by viruses

Transduction involves the transfer of genetic substance from one cell to another via a bacteriophage, a virus that infects bacteria. The bacteriophages produced in a host cell are released when the cell is destroyed by the infection. Some virus particles may carry fragments of DNA from the host bacterium, which, under certain circumstances, may become incorporated into the DNA of the recipient cell, thus changing the genetic constitution of the latter.

EVOLUTION

Bacteria have existed from very early periods in the history of life on Earth. They have been detected as fossils in rocks dating from at least Devonian times (as early as 395,000,000 years ago). On the basis of their indistinct nuclear matter, bacteria are assumed to be closely related to the blue-green algae. It has been speculated that a photosynthetic ancestor may have given rise to both the bacteria and the blue-green algae.

Relationship to other microorganisms

Several groups of bacteria show features that suggest relationships to other classes of organisms: these features do not necessarily indicate a close relationship, however. Actinomycetes form branching units and reproductive stages similar to the fungi. Mycoplasmas (formerly called PPLO and L forms) are variable organisms that seem to lack in organization. Myxobacteria resemble, in some ways, the slime molds.

CLASSIFICATION

Distinguishing taxonomic features. Shape and size of the bacterial cell are aids in classifying bacteria, as is the appearance of the colonies that are formed. Other characteristics, however, assume even greater significance because they are more consistent under different environmental conditions. These include the kinds of foods utilized, the products of metabolism, reactions to specific chemicals, antigenic composition, and degree of tolerance to environmental change.

The 10 orders listed in this classification are distinguished primarily on the shape and rigidity of the bacterial cell and on locomotor ability; on the capacity of individual bacteria to aggregate in chains or clusters of special shape; and on special physiological characteristics.

Annotated classification. The following classification is based on an edition of *Bergey's Manual of Determinative Bacteriology.*

CLASS SCHIZOMYCETES

Unicellular microorganisms generally ranging from 1 to 5 microns in size; variable in shape and in nutritional needs; lacking a distinct nucleus; most species without chlorophyll; occur singly or in chains or clusters and form distinctive colonies; about 1,500 species; worldwide distribution.

Order Pseudomonadales

Rigid-walled cells of variable shape, in some species forming chains; photosynthetic pigment present in certain species; cells usually motile by means of a single flagellum. Species in soil and in freshwater and saltwater. Examples of genera: *Vibrio comma* (cholera bacteria), *Pseudomonas, Nitrosomonas, Thiobacillus.*

Order Chlamydobacteriales

Rigid-walled cells in many-celled filaments (trichomes), frequently ensheathed; occasionally produce motile spores; trichomes often attached to a surface; species in freshwater and marine habitats. *Sphaerotilus natans* common in polluted water.

Order Hyphomicrobiales

Rigid-walled cells often attached to surface by a stalk; reproduction by budding (as in yeasts) rather than by ordinary division. Genera include *Rhodomicrobium* and *Hyphomicrobium.*

Order Eubacteriales

Rigid-walled cells, coccoid or bacilloid, sometimes in chains; motile forms move by means of laterally emergent flagella; not acid-fast (*i.e.,* retaining a bacterial dye when treated with an

acidic solution); includes the largest number of genera of concern to man—*e.g., Escherichia, Diplococcus, Staphylococcus, Streptococcus, Bacillus, Lactobacillus.*

Order Caryophanales
Rigid-walled cells in trichomes; motile by means of lateral flagella; very large cells (up to 30μ long and 3μ across); occur in water and decomposing matter as well as in the intestines of arthropods and vertebrates. *Caryophanon latum,* common in cow dung, and *Simonsiella muelleri,* found in the mouths of humans and domestic animals.

Order Actinomycetales
Rigid-walled cells that may grow out in a branching system, resembling mold colonies; includes *Mycobacterium tuberculosis* (tuberculosis bacterium), *Streptomyces.*

Order Beggiatoales
Rigid-walled cells, usually large and often in trichomes that move by gliding motion as do some blue-green algae; genera include *Beggiatoa, Thiothrix.*

Order Myxobacteriales
Flexible-walled cells that creep on surfaces. Stalked fruiting bodies usually develop from a spreading colony, like slime molds. Found in soil, compost, manure, and rotting wood; genera include *Myxococcus, Chondrococcus, Sorangium.*

Order Spirochaetales
Spiral cells that swim by flexion; found in water and in the bodies of vertebrates; genera include *Borrelia, Treponema,* and *Leptospira,* all parasites of man and other animals.

Order Mycoplasmatales
Flexible-walled cells, nonmotile, highly variable in shape at different life stages; includes *Mycoplasma* and forms once known as pleuropneumonia-like organisms (PPLO) and L forms, which are apparently intermediate between true bacteria and rickettsias.

Critical appraisal. The classification of bacteria, as given above, encompasses 10 orders. The question of higher categories, however, remains a subject of study and debate. Some biologists favour a taxonomic system in which the bacteria (class Schizomycetes) are grouped with the rickettsias and viruses (class Microtatobiotes) as the division Schizomycophyta, which, in turn, is grouped with the Phylum Cyanophyta (blue-green algae) as a subkingdom, Monera, of the kingdom Protista (of equal rank with the kingdoms of Plantae and of Animalia). Such a scheme obviates the need for assigning these procaryotic organisms (*i.e.,* without distinct nuclei) to the kingdom Plantae, which is not quite suitable.

For many years the actinomycetes, myxobacteria, and mycoplasmas were listed as bacteria even though they have unusual life cycles and variable structure. It is now fairly well established, however, that they do belong in the class Schizomycetes. The rickettsias, which were occasionally grouped with the bacteria, are now considered an order, Rickettsiales, of the class Microtatobiotes. A completely revised taxonomic scheme appears in the latest edition of *Bergey's Manual of Determinative Bacteriology* (see *Bibliography* below).

BIBLIOGRAPHY. R.S. BREED *et al.* (eds.), *Bergey's Manual of Determinative Bacteriology,* 8th ed. (1974), a reference and sourcebook accepted as standard throughout the world for classification of bacteria and related microorganisms; J.E. BLAIR, E.H. LENNETTE, and J.P. TRUANT (eds.), *Manual of Clinical Microbiology* (1970), a reference work describing methods and techniques for the isolation and identification of pathogenic bacteria and other disease-producing microorganisms; C.J. CORUM (ed.), *Developments in Industrial Microbiology,* vol. 10 (1970), a documentation of the Proceedings of the 25th General Meeting of the Society for Industrial Microbiology on Low Level Microbiological Assays, Industrial Microbiology, and World Food Problems, with other contributed papers; H.W. DOELLE, *Bacterial Metabolism* (1969), details of the biochemical reactions and processes of particular microorganisms; W.C. FRAZIER, *Food Microbiology,* 2nd ed. (1967), a textbook on the activities of microorganisms important in foods, as well as methods for their detection and control; T.R.G. GRAY and DONALD PARKINSON (eds.), *The Ecology of Soil Bacteria* (1968), reports given at the International Symposium on Soil Bacteriology, University of Toronto; A.E. KRISS *et al., Microbial Population of Oceans and Seas* (1967; orig. pub. in Russian, 1964), a synoptic picture of the distribution and range of marine microorganisms; C.A. LAWRENCE and S.S. BLOCK (eds.), *Disinfection, Sterilization, and Preservation* (1968), a handbook on the principles and practical aspects of the control of microorganisms by chemical and physical methods; JOEL MANDELSTAM and K. MCQUILLEN (eds.), *Biochemistry of Bacterial Growth* (1968), a description of the way in which the simple organic and inorganic constituents of the medium are transformed into bacterial cell material; M.J. PELCZAR, JR. and R.D. REID, *Microbiology,* 3rd ed. (1972), a textbook presenting the major areas of study in the field of microbiology; ALAN RHODES and D.L. FLETCHER, *Principles of Industrial Microbiology* (1966), a general survey of industrial microbiological processes; A.J. SALLE, *Fundamental Principles of Bacteriology,* 6th ed. (1967), a textbook concerned with the fundamental concepts of basic and applied microbiology; R.Y. STANIER, M. DOUDOROFF, and E.A. ADELBERG, *The Microbial World,* 3rd ed. (1970), an advanced textbook covering the major characteristics of microorganisms—*i.e.,* morphology, physiology, and biochemistry; F.S. THATCHER and D.S. CLARK (eds.), *Microorganisms in Foods: Their Significance and Methods of Enumeration* (1968), reports of the meeting of the International Committee on Microbiological Specifications for Foods, including information on occurrence, methods of detection, and technical procedures; and A.H. WALTERS and J.J. ELPHICK (eds.), *Biodeterioration of Materials* (1968), a collection of scientific papers given at the First International Biodeterioration Symposium.

(M.J.P./Ed.)

Baghdad

Baghdad (also spelled Bagdad, Arabic Baghdād) is the capital of Iraq and one of the largest cities in the Middle East. Located near the centre of Iraq, about 330 miles (530 kilometres) from the head of the Persian Gulf, Baghdad is famous as the capital of the ʿAbbāsid caliphs and the setting of many of the stories in *The Thousand and One Nights.* Baghdad exhibits marked contrasts in architecture and life-styles, combining Oriental bazaars, shrines, and mosques with riverfront cafés, Western-style luxury hotels, and modern high-rise apartments. With almost a third of the country's population, Baghdad is the centre of Iraq's political, economic, and cultural life.

This article is divided into the following sections:

Physical and human geography

THE LANDSCAPE

The city site. Baghdad is situated on the Tigris River at the river's closest point to the Euphrates, 25 miles to the west. The Diyālā River joins the Tigris just southeast of the city and borders its eastern suburbs. The terrain surrounding Baghdad is a flat alluvial plain 112 feet (34 metres) above sea level. Historically the city has been inundated by periodic floods from the Tigris' tributaries to the north and east. These ended in 1956 with the completion of a dam on the Tigris at Sāmarrā', north of Baghdad, and the ending of the floods has permitted extensive expansion of the city to the east and west. To the north, urban expansion has absorbed the medieval townships of al-Aʿẓamīyah on the east bank and al-Kāẓimīyah on the west bank.

Climate. The climate is hot and dry in summer, cool and damp in winter. Spring and fall are brief but pleasant. Between May and September the average daily maximum temperature is 105° F (41° C), and the high may reach 120° F (49° C) at midday in July and August. Intense daytime heat is mitigated by low relative humidity (10 to 50 percent) and a temperature decline of 30° F (17° C) or more at night. In winter the average daytime temperature is about 55° F (13° C), and the temperature occasionally drops below freezing. Rainfall is sparse (six inches, or 150 millimetres, annually) and mainly occurs between December and April. There is no rain in summer. In spring and early summer the prevailing northwesterly winds (shamals) bring sandstorms that frequently bathe the city in a dusty mist.

The city layout. *The districts.* The city extends along both banks of the Tigris. The east-bank settlement is known as Ruṣāfah, the west-bank as al-Karkh. A series of modern bridges, including one railroad trestle, links the two banks. From a built-up area of about four square miles (10 square kilometres) at the beginning of the 20th century, Baghdad has expanded into a bustling metropolis with suburbs spreading north and south along the river and east and west onto the surrounding plains.

'he city entre

The older core of the city consists of a rectangle about two miles long and one mile wide located on the east bank. Its length extends between two former city gates, al-Muʿaẓẓam Gate, now al-Muʿaẓẓam Square, in the north and ash-Sharqī Gate, now Taḥrīr Square, in the south. From the Tigris the rectangle runs eastward to the inner bund, or dike, built by the Ottoman governor Nāẓim Pasha in 1910. Rashīd Street in downtown Baghdad is the heart of this area and contains the city's financial district, many government buildings, and the copper, textile, and gold bazaars. South of Rashīd Street a newer commercial area with shops, cinemas, and business offices has spread along Saʿdūn Street. Parallel to Saʿdūn, Abū Nuwās Street on the riverfront is the city's showpiece and its entertainment centre, featuring cafés, restaurants, luxury hotels, and, along its southern reaches, rows of ultramodern townhouses.

Adjacent to these commercial districts are older, middle-class residential areas, such as as-Sulaykh to the north, al-Wizarīyah to the west, and al-Karrādah to the south, now densely settled. Baghdad University and a fashionable new residential area are located on al-Jādrīyah, a peninsula formed by a bend in the Tigris.

'lanned eighbour- oods

Since the late 1950s the city has expanded eastward beyond the bund. Planned middle-class neighbourhoods are located between the bund and the Army Canal, which connects the Tigris and Diyālā rivers. Beyond the canal, at the eastern edge of the city, is a sprawling low-income housing development inhabited by more than 1,000,000 urban migrants.

On the west bank are a number of residential quarters, including al-Karkh (an older quarter) and several upper middle-class districts with walled villas and green gardens. Chief among these is al-Manṣūr, surrounding the racetrack, which provides boutiques, fast-food restaurants, and sidewalk cafés that appeal to its affluent professional residents.

Architecture and monuments. The architecture of the city ranges from traditional two- or three-story brick houses to modern steel, glass, and concrete structures. The traditional Baghdad house, usually located in a crowded narrow street, has latticed windows and an open inner courtyard; a few fine specimens from the late Ottoman period are tucked away in traditional quarters of al-Karkh, Ruṣāfah, and al-Kāẓimīyah. The typical modern middle-class dwelling is built of brick and mortar and has a garden and wall.

Traditional and modern houses

While no monuments survive from the early ʿAbbāsid period, examples of late ʿAbbāsid architecture include the ʿAbbāsid Palace (late 12th or early 13th century) and the Mustanṣirīyah (an Islāmic law college built by the caliph al-Mustanṣir in 1233), both restored as museums, and the Sahrāwardī Mosque (1234). The Wasṭānī Gate, the only remnant of the medieval wall, has been converted into the Arms Museum.

Another group of buildings dates from the late 13th and 14th centuries (the Il-Khanid and Jalāyirid periods). These include the minaret of the caliph's mosque (1289); the ʿAqūlī Mosque (1328); and two superb buildings constructed by the Jalāyirid governor Marjān ibn ʿAbd Allah: the Marjān Mosque (1356), partly demolished in 1946, and the Marjān Khān (1359), a restored caravansary (inn). A number of mosques, bazaars, and public baths survive from the Ottoman period.

A cultural revival in the post-1958 period has produced many modern monuments, the work of contemporary artists and sculptors. Among the best known are Jawād Salīm's Liberation Monument in Taḥrīr ("Liberation") Square, depicting the struggle of the Iraqi people to achieve liberty before the 1958 revolution, and Muḥammad Ghānī's "Murjāna Monument," which depicts Murjāna, Ali Baba's housekeeper in *The Thousand and One Nights*, pouring boiling oil on the 40 thieves. Two monuments are dedicated to the war dead. A large, modernistic shield, built by Khālid ar-Raḥḥāl in 1982, commemorates the Unknown Soldier. The Martyr's Monument, a 150-foot split dome built in 1983, commemorates the casualties of the Iraqi-Iranian war.

Commemorative sculpture

THE PEOPLE

The population of greater Baghdad has grown tremendously since World War II, exceeding 4,000,000 by 1987. The vast majority of the population is Muslim and Arab. The Muslims are divided, however, between the two main sects of Islām, the Sunnites and the Shīʿites. Other ethnic and linguistic groups include Kurds, Armenians, and people of Indian, Afghan, or Turkish origin. A substantial Persian-speaking population departed for Iran in the 1970s and '80s in the wake of troubles between Iran and Iraq. There are several Eastern-rite Christian communities, notably the Chaldeans and Assyrians, and a small Jewish community with ancient roots in Mesopotamia; most Jews left the country for Israel at the beginning of the 1950s.

Baghdad has a large community of foreign Arabs, including hundreds of thousands of Egyptian workers and a sizable number of Palestinians, many of whom are the second generation to live in the city. The Western community, once substantial, has been reduced since 1958 and is limited mainly to businessmen, members of the diplomatic corps, and executives of foreign companies.

Changes in neighbourhood patterns

Traditionally, people of the same sect, ethnic group, or craft lived together in separate quarters or neighbourhoods, but oil wealth and massive migration from rural areas to the city have resulted in distribution based on socioeconomic stratification. Some patterns persist, however. As the city expanded physically, the government offered parcels of land for a minimal fee to various professional associations. Thus doctors, lawyers, army officers, and those of other occupational groups have tended to concentrate in new neighbourhoods, each with its own mosques, shops, and schools, creating a pattern of cities within the city. In the 1970s the government attempted to curb "horizontal" expansion, and a new phenomenon, high-rise apartments, appeared.

Baghdadis have an affinity for gardens and family recreation. On weekends the city's restaurants, cafés, and public parks are filled with people, particularly along Abū Nuwās Street, where restaurants serve the local delicacy *masgūf,*

Tigris fish roasted over an open fire. Other recreational centres include two islands in the Tigris that have swimming pools and cafés, the Lunar Amusement Park, and az-Zawrā' Public Park and Zoo.

THE ECONOMY

Industry. Most of Iraq's industry, finance, and commerce is concentrated in and around Baghdad. At least half of the country's large-scale industry and much of its smaller industry is located in the Baghdad governorate. The exception is heavy industry (petroleum, iron, steel, and petrochemicals), which is situated near the oil fields in the north (Kirkūk) and the south (in Baṣrah and az-Zubayr). Most economic activities are owned or controlled by the government, which both stimulates and monopolizes the country's economic activities.

Baghdad's modern industry

Modern industry began in the interwar period, spurred by the Law for the Encouragement of Industry in 1929. Early factory production centred on textiles (cotton ginning, spinning, and weaving), food processing, brick making, and cigarettes. Beginning in the 1950s, the government used increased oil revenues to develop industries. The city now produces a wide variety of consumer and industrial goods, including processed foods and beverages, tobacco, textiles, clothes, leather goods, wood products, furniture, paper and printed material, bricks and cement, chemicals, plastics, electrical equipment, and metal and nonmetallic products. Despite the growth of modern industry, however, a large percentage of Baghdad's labour force still works in traditional economic activities, such as retail trade, production of handmade consumer goods, auto and mechanical repairs, and personal services.

The most important industry in Baghdad is the government, the city's principal employer. Hundreds of thousands of citizens work for the government, directly or indirectly, in the civil service, in government-run educational institutions, and in government-owned industrial and commercial enterprises.

Commerce and finance. The main offices of the Central Bank of Iraq, which has the sole right to issue currency, and the commercial Rafidayn Bank are in Baghdad. No foreign banks are allowed. The main offices of the government companies for commerce, trade, and industry are located in Baghdad, as are the branches of foreign companies operating in Iraq.

Transportation. Baghdad is the hub of the country's transportation system. Saddam International Airport, west of Baghdad, serves numerous international airlines, including Iraqi Airways. The major railway lines of the state-owned railway meet at Baghdad. These connect Baghdad with Baṣrah and Umm Qaṣr near the Persian Gulf, with Kirkūk and Irbīl in the northeast, with Mosul in the north, and with al-Qā'im near the Syrian border in the northwest.

Baghdad is also the centre of a regional road network, connecting the city by overland routes with Turkey, Syria, Jordan, Iran, Kuwait, and Saudi Arabia. Baghdad is also connected by road with Europe. Within the city, a network of expressways completed in the 1980s relieves traffic congestion and links the city centre with its suburbs. The main means of public transportation are the red double-deck bus (introduced by the British) and the public taxi.

ADMINISTRATION AND SOCIAL CONDITIONS

Government. Baghdad is both a national and a provincial capital. The governor (*muḥāfiẓ*) of the Baghdad province is nominated by presidential decree but is responsible to the minister of interior. The city is governed by a mayor, who is appointed by the president. As the seat of the national government, Baghdad contains the offices of the president, the Council of Ministers, the National Assembly, and the headquarters of the governing party.

Public services. Since the 1950s the government has greatly expanded public services in Baghdad, providing low-income housing for poor and middle-income families, as well as electricity, water, sewage, and medical facilities. Baghdad has numerous hospitals and clinics, many of them specialized, and a major medical complex, Madīnat aṭ-Ṭub ("Medical City").

Education. Public school facilities have expanded rapidly since the 1950s. Education is compulsory through primary school, and statistics show nearly total compliance in Baghdad. The Baghdad governorate has more than 1,000 primary schools, several hundred intermediate and secondary schools, and a number of vocational schools, as well as numerous technical institutes and teachers' training schools. Baghdad is the centre of higher education in Iraq. The University of Baghdad was established in 1957, although some of its faculties were founded much earlier. There are, in addition, three other institutions of higher learning: al-Mustanṣirīyah University, the University of Technology, and al-Bakr Military Academy. Education is free up to and including the university level.

CULTURAL LIFE

Baghdad has become an active cultural centre for the Arab world, producing some of the most prominent modern sculptors, painters, poets, and writers. Iraqi poets, for example, pioneered the free-verse movement in Arabic.

Among the most important of Baghdad's museums are the Iraqi Museum, containing important archaeological treasures from ancient Mesopotamian history; the National Museum of Modern Art, containing a permanent collection of painting, sculpture, and ceramics by Iraqi artists; and the Museum of Iraqi Art Pioneers, holding the works of Iraqi artists who laid the foundation of the modern Iraqi art movement.

The city's Muslim shrines

Several of the most important mosques and shrines in the Islāmic world are found in Baghdad, including the shrine of the Shī'ite imams Mūsā al-Kāẓim and Muḥammad al-Jawād, in al-Kāẓimīyah; the shrine of the Sunnite jurist Abū Ḥanīfah, in al-A'ẓamīyah; and the shrine of 'Abd al-Qādir al-Jīlānī, founder of the Qādirīyah Ṣūfī order, in Ruṣāfah. All contain libraries and are centres of Muslim pilgrimages.

All mass media are controlled by the government. Two major daily newspapers are published in Arabic, and a variety of political, cultural, and professional journals are published. English is the most widely used foreign language, but publications in European and Asian languages can be found. Radio Baghdad broadcasts to the entire country over several frequencies and in several languages. Baghdad's television station began operation in 1956.

The National Theatre, one of the best equipped in the Arab world, has a regular schedule of plays, concerts, musical productions, and cinema. The National Troupe for Popular Arts presents Iraqi dance and folklore and tours world capitals. Cinema plays an important role as a source of popular entertainment in Baghdad. The Baghdad International Fair, held annually in October, includes industrial displays, theatrical productions, and other cultural activities.

History

FOUNDATION AND EARLY GROWTH

Archaeological evidence shows that the site of Baghdad was occupied by various peoples long before the Arab conquest of Mesopotamia in AD 637, and several ancient empires had capitals located in the vicinity. The true founding of the city, however, dates from 762, when the site, then occupied by a Persian village called Baghdad, was selected by al-Manṣūr, the second caliph of the new 'Abbāsid dynasty, for his capital. His city, built within circular walls and called Madīnat as-Salām ("City of Peace")—and known as the Round City—was located between present-day al-Kāẓimīyah and al-Karkh. More a government complex than a residential city, it was about 3,000 yards (2,700 metres) in diameter and had three concentric walls. Its four equal quarters were used mainly to house the caliph's retinue. Four main roads led from the caliph's palace and the grand mosque at the centre to various parts of the empire.

Baghdad's founding in 762

The limited size of this city resulted in rapid extramural expansion. Merchants built bazaars and houses around the southern gate and formed a district called al-Karkh. From the northeast gate the Khurāsān road was joined by a bridge of boats to the east bank of the Tigris. There, around the palace of al-Manṣūr's heir apparent, al-Mahdī,

grew up the three suburbs of Ruṣāfah, ash-Shammāsīyah, and al-Mukharrim, the forerunners of the modern city. By 946 the seat of the caliphate was fully established on the east bank, and Ruṣāfah grew to rival the Round City.

Baghdad reached the zenith of its economic prosperity and intellectual life in the 8th and early 9th centuries, under al-Mahdī, who reigned from 775 to 785, and his successor, Hārūn ar-Rashīd (786–809). It was then considered the richest city in the world. Its wharves were lined with ships from China, India, and East Africa. The caliph al-Ma'mūn (813–833) encouraged the translation of ancient Greek works into Arabic, founded hospitals and an observatory, and attracted poets and artisans to his capital. The glory of Baghdad in this period is reflected in stories in *The Thousand and One Nights.*

From the mid-9th century onward the 'Abbāsid caliphate was gradually weakened by internal strife, by failure of crops caused by neglect of the irrigation system, and finally, in the 10th century, by the intrusion of nomadic elements. A civil war between Hārūn ar-Rashīd's two sons resulted in destruction of much of the Round City. Between 836 and 892 the caliphs abandoned Baghdad for Sāmarrā' in the north, and the city was taken over by the unruly Turks they had imported as bodyguards. When the caliphs returned to Baghdad they made their capital on the east bank. Invasions and rule by alien elements (the Būyids from 945 to 1055 and the Turkish Seljuqs from 1055 to 1152) left parts of the city in ruins.

CENTURIES OF DECLINE

This long, slow decline was merely a prelude to the devastating attacks from which Baghdad would not recover until the 20th century. In 1258 Hülegü, the Mongol conqueror, overran Mesopotamia, sacked Baghdad, killed the Caliph, and massacred hundreds of thousands of residents. He destroyed the surrounding dikes and headworks, making restoration of the irrigation system impossible and thereby destroying Baghdad's potential for future prosperity.

The Mongol sack of Baghdad

Thereafter Baghdad became a provincial capital, first of the Mongol emperors of Iran, the Il-Khanids (1258–1339), and then of their vassals, the Jalāyirids (1339–1410). In 1401 the city underwent yet another Mongol sack by Timur (Tamerlane), after which it fell under the sway of two successive Turkmen dynasties (1410–1508), both of which did little to restore its fortunes.

In 1508 Baghdad was temporarily incorporated into the new Persian empire created by the Ṣafavid shah Ismā'īl I. The city was not to remain under the Persians, however. In 1534 the Sunnite Ottoman sultan Süleyman I retook the city. Despite repeated Persian attacks, it remained under Ottoman rule until World War I, except for a brief period (1623–38) when it was taken and held by the Persians.

BEGINNINGS OF MODERNIZATION

In the 19th century European influence grew in Baghdad with the establishment of French religious orders and increased European trade. In 1798 a permanent British diplomatic residency was established there, and the British residents soon acquired a power and prestige second only to that of the governor.

Prosperity began to be restored to Baghdad with the opening of steamship travel on the Tigris in the 1860s. Between 1860 and 1914 several energetic, reforming Ottoman governors improved the city, especially Midhat Paşa. During his tenure (1869–72), he destroyed the city walls, reformed the administration, started a newspaper, and set up a modern printing press. The telegraph, military factories, and modern hospitals and schools were also established, along with a municipal council.

THE MODERN CITY

In 1920 Baghdad became the capital of the newly created state of Iraq. Recognizing British conquest of the state in World War I, the League of Nations granted Great Britain a mandate to govern Iraq, and it did so until 1932. British influence remained dominant until 1958, when the Hashemite monarchy that Britain had helped to establish was overthrown in a military coup d'état. For a decade after 1958 Baghdad underwent a period of political turbulence, with a succession of coups and military regimes. In 1968 the Arab Socialist Ba'th Party came to power. The Ba'thist government achieved relative stability and internal development, particularly after 1973, when rises in oil prices greatly increased revenues to the government and the populace. It was in this period that Baghdad saw its greatest expansion and development. Both were curtailed, however, by the eruption in 1980 of a bitter war with neighbouring Iran.

British influence

BIBLIOGRAPHY. GUY LE STRANGE, *Baghdad During the Abbasid Caliphate* (1900, reprinted 1983), remains the standard work on the city's history to 1258. GASTON WIET, *Baghdad: Metropolis of the Abbasid Caliphate,* trans. from French (1971), is a general, more anecdotal account. A.A. DURI, "Baghdad," in *Encyclopaedia of Islam,* new ed., vol. 1 (1960), pp. 894–908, brings the history to the middle of the 20th century and includes a bibliography of original sources. A collection of scholarly articles on the history and culture of the city (in French) can be found in a special issue of *Arabica,* vol. 9 (1962). JACOB LASSNER, *The Topography of Baghdad in the Early Middle Ages* (1970), offers a detailed analysis of the city's early geography and development. ROBERT M. ADAMS, *Land Behind Baghdad: A History of Settlement on the Diyala Plains* (1965), studies the area around the city. A discussion of the architectural monuments of Baghdad, with beautiful photographs, is presented in IHSAN FATHI, *The Architectural Heritage of Baghdad* (1964); and JOHN WARREN and IHSAN FATHI, *Traditional Houses in Baghdad* (1982), is an account of domestic architecture. Modern Baghdad is sparsely covered. FREYA STARK, *Baghdad Sketches* (1937), is a personal account of life and customs, now somewhat dated. Later impressions and good photographs are found in the chapter "Baghdad" in GAVIN YOUNG, *Iraq, Land of Two Rivers* (1980), pp. 25–67; and WILLIAM ELLIS, "The New Face of Baghdad," *National Geographic,* 167 (1):80–109 (January 1985). Useful information and detailed city maps are offered in the guidebook prepared by the BAGHDAD WRITERS GROUP, *Baghdad and Beyond* (1985).

(P.A.Ma./L.Y.B.)

Balkans

Since the early 19th century, the name Balkan, a Turkish word meaning mountain, has been applied to the easternmost of the three great southern peninsulas of Europe. The peninsula shades gradually into the European mainland, and it is therefore difficult to assign its exact geographical boundaries; but for the purpose of this article the Balkans are taken to mean the territory of the modern states of Greece, Albania, Yugoslavia, Bulgaria, and Romania. (For further treatment of the physical and human geography and the history of these states, see the articles ALBANIA; BULGARIA; GREECE; ROMANIA; and YUGOSLAVIA.) Western Turkey is often included because of historical considerations and will be discussed here where appropriate. The area possesses certain basal resemblances to the Iberian and Italian peninsulas, particularly in its relation to the folded mountain chains of southern Europe and in its structural elements. During the 19th century, when geographers and geologists were acquiring new knowledge of the interior of the region and were coming to regard it as an entity, not merely as a background to Greece and Byzantium, great political changes were taking place within it. The people, submerged by the Turkish advance, began to organize themselves into national states and, as the Turkish empire contracted, new names appeared on the map. The growth of the new states was accompanied by much turmoil, which had reflex ef-

fects outside the peninsula limits, however these be drawn; but the essential point is that it drew general attention to the region. It became increasingly clear that all the older European states, if in varying degree, were interested in the delimiting of Balkan boundaries, and thus the facts disclosed by detailed geographical study had more than purely technical importance.

No mountain barrier separates the peninsula from the continental mainland. There is thus no sharp break of continuity such as is experienced when the Alps are crossed and a new world is disclosed in Italy. This physical continuity is accentuated by a notable increase in the width of the peninsula toward the north. The line of the Danube and its tributary the Sava has sometimes been chosen as a convenient northern limit (east to west). This limit has a certain justification. The Danube–Sava line, easily recognized on a map, served for a period as a boundary to the Turkish empire and thus as the frontier of Christendom. However, a geographically satisfactory frontier in this northwestern section is difficult to draw. Even the Danube–Sava line, at least to the west of the Danube's Iron Gate, has never been a limit so far as people are concerned; it bears no relation to political frontiers, and the post-1918 state of Yugoslavia extends well beyond it.

The second outstanding feature is the peculiar structure that causes the peninsula to fall into two very unequal and very dissimilar parts. To the south is Greece, a secondary peninsula, with an average width of only about 125 miles (200 kilometres). Although both sections are highly mountainous, not only is the Greek section much narrower but also it has a peculiarly dissected coastline that brings sea influences within easy reach of nearly every part. The broad, continental northern section, on the other hand, is largely removed from the surrounding seas because of its width, the nature of its shorelines and, in part, of the direction of its mountains. In climate, in vegetation, and in possible crops, it differs profoundly from Greece. No less profound was the effect of the actual remoteness from the sea routes so freely open to the people of the south. It is this division into two parts—one sharing in the full life of the Mediterranean peoples, the other cut off from it—rather than the absence of a definite northern limit that made the European world so slow to recognize the existence of a Balkan peninsula. Until the peoples of the continental segment awoke, the whole northern area tended to be regarded only as a broader equivalent of the Alps or Pyrenees: the real peninsula was the Grecian one.

The article is divided into the following sections:

Physical and human geography

THE LAND

Relief. It is the presence within the peninsula of young fold mountains that makes it essentially similar to the Iberian and Italian peninsulas. Two separate series of these can be recognized, one, of transverse direction, lying to the east, and the other, which is longitudinal, in the west. The Transylvanian Alps swing around in a great curve, the Danube breaking through at the Iron Gate at the western apex of the curve, and are continued in the Balkan Mountains (Stara Planina), which have a roughly parallel direction. Fingering out eastward into several separate ranges and breaking off steeply on the shores of the Black Sea, these rise to a maximum height of nearly 8,000 feet (2,440 metres), and the most noted of their passes, the Shipka Pass, has a summit level well above 4,000 feet. Northward these mountains sink gradually to a limestone tableland, presenting a marked contrast to the alluvial Walachian plain (Cîmpia Romînă) beyond the river, and the presence of this tableland means that the northern slopes as far as the passes, relatively high though these are, are gentle.

Just as the Balkans are a continuation of the Carpathian branch of the Alpine chain, so the main chain itself bends down the western side of the peninsula. From the Julian Alps north of Trieste a series of mountains runs in a southeasterly direction close to the coast and parallel with it. These, to which the general name of Dinaric Alps may be given, rise to 8,274 feet in the peak Bolotov Kuk of Durmitor; but their significance as a barrier does not depend upon their height. They are characterized by the great development of massive limestones, particularly extensive in the area lying behind the peninsula of Istria. These limestone areas, called karst, display to a very marked extent certain peculiar relief features, dependent on the effect of rainwater on their constituent rocks. Thus the surface soil is very thin, bare rock being frequently exposed; running water is usually absent at the surface, most of the rivers sinking, after a short course, into cavities of the rocks; and caves and sinks are common, as well as elongated depressions called locally *polje*, or fields, because only in them as a rule is there sufficient depth of soil to permit cultivation.

In places the limestone mountains rise steeply from the shore of the Adriatic, but a certain amount of subsidence has occurred, with the result that numerous islands fringe the coast. Because the mountain folds run parallel to the shore, the islands tend to be elongated in the direction of the coastline, and the straits and inlets tend to have the same direction.

The Dinaric Alps may be said to extend to the neighbourhood of the mouth of the Drin in Albania. There the coast changes in direction, trending almost north–south, and the mountain belt thins out and draws back from the coast, so that the Albanian lowland invervenes between it and the sea. This triangular area, with its inland apex at Elbasan, extends southward to Vlorë (formerly Valona); it shows another contrast with the Dalmatian area farther north in that several large and permanent rivers flow from the mountains across the lowland to the sea; there, then, access to the interior becomes at least relatively easy.

Vlorë, with the adjacent sheltering peninsula ending in Cape Linguetta (Albanian Kep i Gjuhëzës), marks the beginning of a new change. The coast resumes a southeasterly direction and the fold mountains become more conspicuous as the Pindus range, which extends through peninsular Greece. Beyond the down-faulted Gulf of Corinth, in the Peloponnese, the ranges diverge like fingers, leaving narrow triangular plains in between. In the eastern part of the Greek peninsula several ranges diverge eastward or southeastward from the Pindus, and between them lie the well-watered open plains which, like those of the Peloponnese, nourished early centres of civilization.

The Balkan Peninsula's central core

The central core of the Balkan Peninsula is an old crust block, roughly triangular in shape, with its apex pointing

toward Belgrade and its broad base approaching the shore of the Aegean Sea. That sea is believed to overlie a former extension of the crust block that has sunk beneath its waters. The numerous Greek islands represent fragments of the surface of this lost land that remained above sea level when the remainder sank. The narrow staits of the Bosporus and Dardanelles are also regarded as flooded parts of the courses of rivers which crossed the old land.

The central core of the peninsula, though it has retained a position above sea level, has been greatly modified as a result of the formation of the fold mountains on its margins. The narrowed northern region, constricted between the Dinaric Alps and the curve of the Balkans, is a broken hilly country traversed by a continuous longitudinal depression through which the Morava River flows on its way to the Danube. South of Niš (Nish), Yugoslavia, however, and extending to the Gulf of Salonika is a region of great structural complexity which seems to have received the full force of the thrust. Faults are innumerable, and closed basins alternate with short and steep highland belts. The basins tend to be elongated in a longitudinal direction, and the major river courses—for example the Vardar—consist of alternating basins and gorges. Many of the basins have formerly been lakes, and since they are floored with fertile soil, they are fitted to become centres of population; but their isolation from each other has had very important human effects.

In contrast with this fractured and much subdivided region, the southeastern part of the triangle, that lying between the Balkans and the Aegean, shows relative simplicity. There the core reaches its greatest height (9,596 feet in the Rila Mountains), and there also is the broadest unbroken mass of elevated ground. The general name of Rhodope (Rodopi) may be given to the whole block, though the separate parts have local names. The Rhodope upland is separated from the Balkan Mountains by a considerable lowland, the Rumelian plain. This is one of the most considerable tracts of lowland within the peninsula and continues to the shores of the Sea of Marmara. There is also an interrupted belt of plain in Thrace, between the Rhodope and the Aegean, the total result being to make the southeastern part of the peninsula much less continuously hilly than the northwest, where lowlands are virtually absent.

Settlement patterns. The build of the peninsula has influenced the routes and the areas of settlement within and the zones of effective contact with adjacent lands outside. The Balkans have indeed proved in practice much less of a barrier to human movement than would appear from a map.

Yugoslavia. An orographical map shows an almost continuous area of high ground on the west, continued into the Grecian peninsula, which is mostly mountainous. In the northwest the way in which the high ground within the peninsula passes into the Alps proper means that the Danubian plains have no natural, easy exit to the Adriatic. But the mountain belt thins out in the karst of the Slovenian Julijske (Julian) Alps. Although this belt is neither wide or lofty—it does not rise much above 5,000 feet—because of its karstic nature it forms a very effective barrier. The chief elements of the belt are the Velika Kapela and Velebit mountains, both remarkably waterless and barren. Continuous river valleys to serve as natural routes are generally absent and, where they do occur, the stream tends to flow in steepsided canyonlike gorges that form a great obstacle to transverse movement. Further, not only do these lands form a barrier between the sea and the interior but also they can as a rule support only a scanty and scattered population, for the local resources are small.

Beyond the karst areas lies a tract of undulating country. This area is a continuation of the plains of the Danube and, where the Drava and Sava converge toward one another in Slavonija, it includes a considerable area of true plain. This whole area of mountain and plain was once the kingdom of Croatia-Slavonia; from a physical standpoint it is a transition region between central Europe and the Balkan Peninsula. Until the creation of Yugoslavia it had little direct relation to the Adriatic, though economic and political causes led to Hungary's making great efforts to develop Rijeka (Fiume) as a grain port during the late 19th century. In the past, Croatia-Slavonia was politically, economically, and culturally attached to central Europe, but ethnically it belongs to the Balkan Peninsula.

Dalmatian coast islands

The islands off the Dalmatian coast to the south are usually fertile and there is often a strip of productive land fringing the inlets. Water is also easy to obtain, for the streams which were lost on the heights above emerge as full-grown rivers where the rocky hills descend in cliffs to the sea margin, or springs even bubble up on the seafloor itself. In contrast with the dry and barren lands above, therefore, there is possibility of cultivation and settlement on the shore. But the coastal areas are too narrow and the difficulties of communicating with the interior too great to have allowed for the rise of indigenous civilizations there. The scattered towns on this Dalmatian coast represent islets of ancient but alien (Venetian) culture and, until relatively recent times, scarcely influenced the interior of the peninsula. They arose as offshoots of areas enjoying much greater advantages.

To the south of the Velebit mountains a narrow strip of coastline, with the mountain crest behind, forms Dalmatia. But the historic Dalmatia is really an interrupted series of maritime towns, Zadar, Šibenik, Trogir, Split, and Dubrovnik (respectively the former Zara, Sebenico, Trau, Spalato, and Ragusa) being among the most important. Until the creation of Yugoslavia these towns had little connection with the interior.

The actual mountain belt, from the borders of Croatia-Slavonia to the confines of Greece, extends through the Yugoslav republics of Bosnia-Hercegovina and Montenegro, to Albania. In Bosnia the limestone rocks of the coastal area give place to others, including sandstones, which allow for the development at once of deeper soils and of a more normal drainage system. Numerous rivers drain into the Sava, and Bosnia can be reached from that river, and thus from the Danubian plains, with relative ease. Other route lines connect it with the interior. Hercegovina is a karstic area with only one important river, the Neretva, which flows to the Adriatic. Montenegro is essentially a mountain aerie, a refuge that withstood invasion in the past as much perhaps because of its worthlessness as of the difficulty of conquest.

Eastern Yugoslavia comprises the western part of the crust block, with its marked contrast between the northern section, draining into the Danube mainly by the Morava, and its complex southern section, draining into the Gulf of Salonika by the Vardar. West of the Morava the country is undulating, and lowlands fringe the southern bank of the Sava. The Western Morava, a tributary entering the main stream from the west, also helps to define a block of land which affords possibilities of settlement. This was the nucleus of 19th-century Serbia, with Belgrade, at the junction of the Sava and Danube, as its capital. The complex southern region is Macedonia, with its jumble of people and its long history of turmoil and disorder.

Albania. Albania, with its coastal lowland, formerly malarial, and its mountainous hinterland, is a region of much interest. As already seen, it affords the possibilities of routes to the interior by the Drin, Shkumbin, and Vijosë valleys.

Greece. Peninsular Greece comprises three relief elements: (1) the Pindus ranges in the centre and in the west, continuing southward beyond the Gulf of Corinth into the four-fingered Peloponnese; (2) the western coastal belt, which is more broken than that of Dalmatia, while the offlying Ionian Islands are larger and exhibit less parallelism to the coast; (3) the eastern ranges and plains, which in the north belong (including Mt. Olympus, 9,570 feet) to the old crust block and farther south are offshoots of the Pindus. The continental or truly Balkan portion of Greece comprises the uplands and plains of southern Macedonia and Thrace, which provide an east–west passageway between the Rhodope and the sea. The plains are largely spread around the lower courses of the major rivers—the Vardar (Axiós), Struma (Strimón), Néstes, and Maritsa (Évros). They were formerly swampy, but many drainage works were executed between World Wars I and II.

Bulgaria. Bulgaria has as its northern frontier the Dan-

ube, except where the river takes its great bend to the north; there the frontier leaves it and runs slightly south of east to the Black Sea, much of the steppelike Dobruja being included in Romania. Southward Bulgaria extends to the Rhodope crest, and it is thus nearly bisected by the Balkan Mountains. Sofia, the capital, lies in a small basin between the Balkan Mountains and a northwesterly prolongation of the Rhodope, the basin being drained by the Iskŭr River, which breaks through the Balkans to enter the Danube. South of the range, and separated from it by a longitudinal depression, lies a parallel upland, the Sredna Gora (Middle Forest). The invervening depression (known, from its principal product, as the "Valley of Roses") is watered by the Tundzha River, which seems to be making for the Black Sea near the port of Burgas, but turns instead sharply southward, breaks through the western end of the Istranca and joins the Maritsa at Edirne.

Bulgaria's "Valley of Roses"

The upper Maritsa, on which stands Plovdiv, flows through the wider depression which has been called the Rumelian plain. These two fertile lowlands, with their bounding uplands, form Eastern Rumelia, which was not united politically to northern Bulgaria until 1885. This region, with southern Macedonia, formed the granary of Turkey in Europe, as it did of the earlier Eastern Empire. The Maritsa depression shows certain analogies to the valley of Andalusia in Spain, both in its position between fold mountains and a crust block and in its value to an invader.

Transportation. Greece within its peninsula, Serbia in the Morava region, and Bulgaria astride the Balkan Mountains all became independent states while Turkey still held Thrace, Macedonia, Albania, and, at least nominally, a large part of the northwest. That the progressive contraction of Turkish territory led to such bitter and prolonged conflict was largely the result of the nature of the routes within the continental section of the peninsula, and particularly of the difficult access of both Bulgaria and Serbia to open water. Thus the natural route lines demand careful consideration. But the routeways which were important in the historic or even the recent past are not necessarily those of most value, for the current political subdivision of the peninsula and in particular the interposition of iron curtain frontiers across the routeways materially lessen their significance.

On the northern, southern, and eastern margins of the peninsula, respectively, are situated the three nodal points of Belgrade, Thessaloníki, and Istanbul, all owing their importance to the land and water routes which converge upon them and all linked together by rail.

Belgrade is situated at the junction of four great routes: from the north and the east, the Danube valley; from the south, the Morava–Vardar corridor; from the west, the Sava valley. The Danube is continuously navigable, despite the partial interruption of the Iron Gate, downstream to the Black Sea as well as upstream. The Sava, though not a first-class waterway, can be used by steamers as far as the Kupa confluence at Sisak. Apart from the waterways, no fewer than seven railway routes converge on the city; the most important are the Orient Express routes from western and central Europe to the Aegean and the Bosporus—from France and Italy via Zagreb and the Sava valley, from Belgium and Germany via Budapest and the Danube plain; to Athens via the Morava and Vardar valleys, to Istanbul via the Morava valley and Sofia, and to Bucharest via Vršac. The last-named route lost its international significance after Romania became a Soviet satellite. An important narrow-gauge railway leads southward across the mountains to the Adriatic coast at Dubrovnik.

The port at Thessaloníki

Thessaloníki is the only good port on the northern coast of the Aegean and is the best exit to open water for much of the interior. Three major land routes and some minor ones converge upon it. The first group consists of the meridional furrow indicated by the direction of the Morava and Vardar rivers, the route from Istanbul by the Thracian lowland and the route to Athens.

The Morava–Vardar furrow is followed by the railway from Belgrade to Thessaloníki by way of Niš and Skopje. Although the headstreams of the two rivers, despite their contrary direction, actually interjoin in wet weather, it must not be assumed that a continuous valley line extends from the Danube to the Thermaíkós Kólpos (the Gulf of Thérmai). The Morava, upstream from Niš, and the Vardar, downstream from Skopje, both pass through gorges which offered considerable resistance to through communication in prerailway days. The Morava gorge at Vranje was actually the Serbian frontier at one stage in the development of that state and at the time the railway was built (1887)—a clear indication of the break in the furrow there. The Belgrade–Skopje section of this route was effectively duplicated in 1931 by construction of a railway which leaves the Morava valley at Lapovo and proceeds via Kragujevac, the Ibar valley and the Kosovo Polje.

The second major route entering Thessaloníki runs through the Thracian plains and hills from Istanbul. It dates from 1895 and was essentially strategic in character—it was built at least 12 miles from the coast to be beyond what was then the range of naval bombardment, and the two major ports which it served, Thessaloníki and Dedeağaç (modern Alexandroúpolis), were both provided with bypass lines. It has had relatively little international significance, though it remains a possible route from central Europe to Istanbul, avoiding the crossing of the iron curtain frontier of Bulgaria. The third route is the southward continuation of the Orient Express route, from Thessaloníki to Athens. This, sometimes called the Greek Longitudinal railway, was completed after the Balkan Wars, in 1916, but its functioning as an international route was delayed by World War I and its aftermath until 1920. It is a difficult route, alternately crossing plains and mountain ranges; its ruling gradient is 1:50.

The westward communications of Thessaloníki are less important. A railway runs to the Florina basin and then turns north, crossing the Yugoslav frontier to Bitola and eventually linking with the Vardar valley line at Titov Veles. The road to the Adriatic, though of no significance at all in modern times, was an important Roman route—the Via Egnatia—leading from Bitola, Lake Ohrid, and the Shkumbin valley to Durrës (Italian Durazzo, Roman Dyrrhachium).

The fact that Thessaloníki is politically Greek and not attached either to Yugoslavia, despite the presence of the Morava–Vardar furrow, or to Bulgaria, despite the relative nearness of Sofia, is explicable rather by the strength of a cultural and historical tradition than by purely physical facts. It is an Aegean port of much importance in the modern world, and to the Greeks the idea that control of Aegean trade is their national right is one that admits of no argument. Part of its basis is of course the geographical fact that their somewhat barren land could not support them unless supplemented by the sea trade for which they have always shown natural aptitude.

Istanbul, with a superb natural position where the waterway from the Black to the Aegean sea crosses the land route from Asia Minor to the Balkan peninsula, is connected to Belgrade by a diagonal furrow, certainly as important as the north-to-south one from Belgrade to Thessaloníki. The route follows the Morava valley to Niš, ascends the Nišava tributary and crosses the Dragoman pass (2,550 feet) to the basin of Sofia, then by another pass it reaches the Maritsa valley and follows this past Plovdiv and Edirne until the valley of the Ergene enables it to turn east toward Istanbul.

The main international railway routes through the Balkan peninsula were conceived in central Europe about 1869, when Turkey still held a large part of the peninsula. The wars of 1876–78, which materially altered the political map, upset the plans, but the main lines, Vienna–Belgrade–Sofia–Constantinople and Niš–Thessaloníki, which in any case were to be international and not local avenues of transport, were opened in 1888. Subsequently, the political boundaries changed several times, more notably after the Balkan Wars of 1912–13 and after World War I, and the creation of a railway network thus responded to different national and local stimuli at different times. For its size and economic development, Bulgaria has a fair system, with important lines linking Sofia with the Danube and with the Black Sea ports of Varna and

The opening of international railway routes

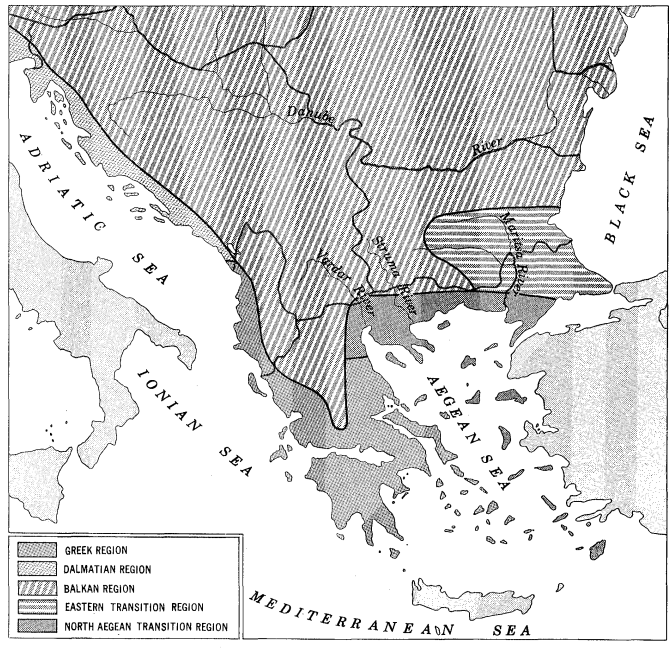
Burgas, an important though difficult route across the Balkan Mountains and many branch lines. After World War II the construction of a bridge across the Danube at Ruse made possible a through route to Moscow. Yugoslavia, assembled like a jigsaw from Serbia, parts of Austria-Hungary, and what remained of Turkey, had to create a national railway system from a series of fragments—and largely succeeded in doing so between World Wars I and II. Its main needs were for better communication between the capital and the Adriatic ports of Split and Dubrovnik, for more railways in the south, and for bridges across the Danube and Tisza rivers.

Climate. A central European climate prevails through the greater part of the broad northern section of the peninsula and even extends into the centre of the northern part of the Greek section, and elsewhere the climate is Mediterranean. No notable difference in climate separates the peninsula proper from the adjacent Danubian lands.

The Mediterranean climate is of peculiar interest because it is so closely linked to a particular type of culture showing a very delicate adaptation to local conditions. Three essential features differentiate the Mediterranean climate: the winters are warm in relation to latitude and the greater part of the rain falls during that period; the summers are hot and dry and one or more months may be practically rainless; throughout the year the skies are clear and there is abundant sunshine, for the winter rains come in heavy showers of short duration. Two regions within the peninsula show typical Mediterranean climate.

One of these is made up by the Greek islands, the Peloponnese, the margins of the northern part of the Greek peninsula, and a narrow strip on the coast of Albania. Even within the region so defined, however, elevation above sea level, as in the mountains, or an upland girdle, as in the plain of Thessaly, may produce local modifications. Athens and Corfu may be taken as representative places within this region. Athens, on the eastern side of a peninsula, largely sheltered from rain-bearing winds, has a much smaller total rainfall than the island of Corfu, which faces the wet winds of winter; but, despite the difference in the amount of rain in the two places, its distribution is closely similar. At both places winter temperatures, as far as monthly averages are concerned, do not fall below 48° to 50° F (9° to 10° C). Summer temperatures range from 78° to 81° F (25° to 27° C; July averages), which permits the ripening of subtropical fruits. In Athens, July and August are practically rainless and only about one-fourth of the total fall occurs in the period April to September, so that more than three-quarters of the total falls in the cooler half of the year. More than one-third falls in the two rainy months of November and December. Corfu has a rainfall four times as great as that of Athens, and, although only July can be described as practically rainless, only about 20 percent of the total rainfall occurs in the period April to September.

A narrow strip along the coast of Dalmatia is regarded as forming a second major region. Dubrovnik may be taken as typical, bearing in mind, however, that it is the most southerly of the important towns of Dalmatia. The winter temperatures there are closely similar to those experienced in the Greek region, much of Dalmatia being remarkable for its mild winter climate. The summers are not quite as hot as in Athens or Corfu, but the real distinguishing feature is the less marked periodicity of the rainfall. There is no rainless month, although July continues to be the driest period of the year. More than 30 percent of the total rainfall occurs during the period April to September, so that summer drought is far less marked. Oranges can be grown without irrigation, and their coexistence with the olive is a distinguishing feature. Not only, however, is the strip with this typical Mediterranean climate very narrow, but changes occur with considerable rapidity toward the north. Except where, as in Split, the form of the coastline gives shelter, the more northerly places are often exposed to the blast of the bora, or cold northerly wind of winter, which at once is a danger to shipping and excludes the more delicate Mediterranean fruit trees. When Trieste is reached on this eastern Adriatic coast, the somewhat cold winters (January average below 40° F) and the fact that



Five major climatic regions of the Balkan peninsula.
From W.G. Kendrew, "Climates of the Continents," by permission of the Clarendon Press

the period April to September shows rather more than half the total amount of rain, with a rainfall maximum in October and a secondary one in June, mark the transition from the Mediterranean climate to the central European one. The northerly Dalmatian towns and the coast of Croatia show stages of the transition.

Southern and eastern Macedonia and western Thrace fall into a third region, characterized by notable modifications of the Mediterranean climate. Thessaloníki may be selected as a representative locality. The winters are cold for the latitude (average January temperature 41° F [5° C]) because of the bitterly cold northerly winds which blow down the Vardar valley, and this feature is accentuated as the coast is left. But the summers are hot and the range of temperature between summer and winter is greater than at any of the places already discussed. There is no rainless month, and, though July is still the driest period and November and December are the wettest months, there is a much more even distribution of rain throughout the year, the six colder months having very little more rain than the six warmer ones. The colder winters, especially where there is exposure to wind, again limit the distribution of the more delicate Mediterranean fruit trees, and growth even of hardy plants is checked during winter. On the other hand the high summer temperatures and the fairly heavy summer rainfall, with the possibility of irrigation from the mountain snows, make it possible to grow crops demanding both heat and moisture but being indifferent to winter cold because of their annual nature. Among such are cotton, rice, tobacco, corn (maize), and, historically, the opium poppy.

As one passes northward from Thessaloníki into the interior of the peninsula, the change in this fourth region, from the modified Mediterranean to the true central European climate, is rapid and is accentuated by the relief. The general features of that climate are the cold winters, the temperatures then showing little relation to latitude, the warm summers and a rainfall well distributed throughout the year, but with a tendency toward an early summer maximum. Skopje, in a latitude somewhat lower than that of Dubrovnik, though in a more elevated position, has mean January temperatures well below the freezing point and fully 18° F (10° C) lower than those of Dubrovnik. This means that there is a definite winter stoppage of agricultural activities, December, January, and February all being too cold even for the growth of grass. On the other hand, despite its elevation, the summer temperatures at Skopje are not more than a few degrees lower than those of Dubrovnik. The rainfall is fairly well distributed

thoughout the year but May, June, and October are the rainiest months, the maximum fall coming in May when temperatures are already fairly high (more than 62° F [17° C]). Belgrade, considerably farther north but less elevated than Skopje, has a very similar temperature range, but November is already a winter month. June instead of May is the rainiest month and the summer rainy period lasts for the three months of May, June, and July, again with a second rainfall maximum in October. Broadly speaking, this climate, with summer heat and summer moisture coinciding, is one well suited to corn among cereals. With local variations caused by height above sea level and by degree of exposure, this type of climate prevails throughout the greater part of the interior of the peninsula, which thus falls into a fourth or central major climatic region. The total rainfall shows a tendency to increase toward the northwest, in Bosnia and northwestern Serbia; it diminishes toward the east.

As a fifth major region, eastern Thrace and the plains of Bulgaria may be included. Over much of this region, especially in Thrace, the total rainfall is very small, giving the landscape a steppelike appearance, and the winters are very cold as a result of exposure to winds from the Russian plain. The Maritsa valley allows Mediterranean influences to penetrate into southern Bulgaria, where also the Balkans give a certain amount of shelter from the cold winds of winter. Parts of the valley plains of Bulgaria have in consequence much milder winters than northern Macedonia or Serbia, and, as there is a tendency for the winters to be wetter than the summers, the climate is sometimes described as modified Mediterranean. The dry, sunny summers favour wheat rather than corn, and the vine is grown in sheltered places. Cotton and rice are grown with irrigation.

To this general division of the whole area into five climatic regions, of which the central one covers by far the greatest area, it may be added that the fact that so much of the surface is elevated introduces numerous modifications in detail. Because of the cold winters of the central area, much of the winter precipitation falls as snow. No mountain within the peninsula rises in the strict sense above the snow line, but the higher peaks of the Rhodope are only snow-free for about one month in the year, and even the Balkan Mountains show some snow until July.

Plant life. The flora of the Balkan peninsula is one of the richest in Europe and consists of about 6,600 species of vascular plants. The largest families, on the basis of numbers of species, are the Compositae, Leguminosae, Caryophyllaceae, Labiatae, Gramineae, and Cruciferae. The genera with most species are *Centaurea* (knapweeds), *Silene* (campions), *Dianthus* (pinks), *Trifolium* (clovers), *Campanula* (bellflowers), *Carex* (sedges), *Verbascum* (Mulleins), and *Thymus* (thymes). Floristically, Greece is the richest area. Apart from the actual number of genera and species the flora of the Balkan peninsula is botanically noteworthy for the number of its endemics (plants found nowhere else in the world). The central portion of the Balkan peninsula is geologically old and has been a land surface at least since the advent of the flowering plants. Moreover, the effects of the Ice Age were much less drastic than in northern and central Europe and many plants found a refuge in the peninsula. There are, therefore, a considerable number of old relict types in the flora. Examples are *Ramonda* and *Haberlea* of the Gesneriaceae, related to western Chinese genera, the Serbian spruce (*Picea omorika*), the Balkan yam (*Dioscorea balcanica*), and *Forsythia europaea*, the only native European species of this well-known garden genus. In contrast, there are many species that are most probably of relatively recent origin and may have partly originated by hybridization following the clearance of woody communities by man. Thus many of the pinks, campions, mulleins, and thymes are closely related one to another and appear still to be evolving new microspecies.

Given the old Tertiary basis of the flora, the chief migration routes have been from the north and the east. From the north there was an extension of central European species, especially with the oncoming of the glacial period. The Aegean Sea is of recent geological origin, and plants from Asia Minor undoubtedly contributed to the floristic wealth of the Greek mainland and islands.

The vegetation of the Balkan peninsula clearly demonstrates the controlling influence of climate on plant life. In the northern and central parts the climax communities are deciduous forests up to about 4,900 feet (1,500 metres), then coniferous forests to about 5,900 feet (1,800 metres) and above these are high mountain scrub and herb and mat associations. In the southern and western coastal districts evergreen Mediterranean woods and brushwoods naturally dominate, with oak and conifer forests at the higher altitudes. Throughout the peninsula there has been great destruction of the woody vegetation by man.

Animal life. The distribution of the animal life of the Balkan peninsula is a fusion of three major elements. Mediterranean forms dominate along the southern and western borders of Greece, Albania, and Yugoslavia; eastern steppe species in the lowlands of the east; and central European forms in the interior. Some of them are widely distributed in Europe, mainly in the mountains, whose peaks still have an Alpine element.

Among Mediterranean mammals are wild goat (bezoar), jackal, porcupine, and several kinds of bat. Eastern elements include the ground squirrel (*Citellus*), some forms of hamster, the mole rat (*Spalax*), etc. The mole, lynx and wildcat, marten, wolf and fox, bear, hare, boar, and red and roe deer are some of the central European mammals, while the chamois is an example of the Alpine element. The same types of distribution are found among the breeding birds. In the south and in Dalmatia, Mediterranean forms predominate, with the great cuckoo, the melodious warbler (*Hippolais*), etc.; and these also penetrate some of the great river valleys (*e.g.*, the Vardar) far to the north. Central European forms include chaffinch, creeper, and some thrushes; among eastern types are the little bustard, the lesser gray shrike, and the red-footed falcon; among Alpines are the mountain lark (*Eremophila alpestris*) and the Alpine ptarmigan. The big birds of prey include the golden imperial eagles and several kinds of vulture. The red-legged partridge (*Alectoris*) and the stork are common.

Among reptiles various forms of lizard are known, especially in seaside regions, and a certain number of species of gecko, turtle, and snake, among which are three kinds of poisonous viper. There are amphibians belonging to the Palearctic fauna; *e.g.*, the true toads, while central European fauna include true frogs, spadefoot toads, and salamanders. There are also some Mediterranean forms. The fishes of lakes and rivers include carp, barbel, small salmon, wels or sheatfish (*Silurus*), etc.

The same zoogeographical pattern is found in the invertebrates. Among harmful plant insects are tussock moths and Moroccan, Italian, and migratory locusts. After World War I the San José scale and the Colorado beetle were brought to Europe from America. The Colorado beetle, which did not reach the Balkans until after World War II, spread quickly.

In low-lying land near water, domestic buffaloes are not uncommon. The important meat animal is the pig and the chief milk and meat animals are cattle and sheep. Hens and turkeys are common among domestic birds.

Very special and localized faunas, not discussed here, live in the delta of the Danube, in the islands of the Aegean Sea, and in Crete.

THE PEOPLE

Ethnic distribution. Five major ethnic groups or races are to be found in the Balkan peninsula. These are the Albanians, the Greeks, the Bulgarians, the Southern Slavs (Serbs, Croats, Slovenes, Montenegrins), and the Turks. All of them, with the possible exception of the Southern Slavs, have at one time been spread over wider areas than those they now occupy. This is particularly true of the Turks, who for several centuries were primarily the governing occupiers of most of the peninsula. They also fairly effectively colonized certain parts of it, particularly in Thrace, Macedonia, and in northeastern Bulgaria, and were finally ousted after World War II. This is also true of the Bulgarians, who at one time extended throughout much of Macedonia and Thrace. World Wars I and

Frontier adjustments

II, and the resultant frontier adjustments and population exchanges, have given the political map a much closer resemblance to the map of nationalities than it has ever had before. Outside of the northwestern fringes of the peninsula, where some Romanians, Hungarians, Austrians, and Italians are included in Yugoslavia, the Albanians of the Kosovo Autonomous Region are probably the most sizeable minority left inside the territory of another country. The most mixed region is undoubtedly Macedonia. There, ethnic differences have been confused by differences of religion, and adherents of the Greek Orthodox Church were long regarded as Greek, whatever their ethnic heritage; similarly, all Muslims were formerly regarded as Turks, whether they were of Bulgarian, Albanian, or Southern Slav stock.

Ethnology. The ethnology of the Balkan peninsula is more involved than that of either the Italian or the Iberian; first, because this peninsula has a more fragmented geography; and second, because it is open to entry from more directions. On the northwest the plains of the middle Danube afford a way to central Europe; on the northeast there is a wide corridor through Moldavia from the steppes of the Ukraine; on the east, an easy passage is possible from Asia Minor across the Bosporus and through Thrace; while the coasts of Greece are accessible to all the eastern Mediterranean basin, and those of Dalmatia, though cut off by mountains from the interior, are easily approached by water from across the Adriatic. People entered the Balkans from all these directions and left vestiges of racial types and customs derived not only from many parts of southern and central Europe but also from wide areas of central Asia and the near east.

The main ethnic divisions of the Balkans are based on linguistic and religious differences rather than on true racial distinctions. Yet it is remarkable that in many cases the cultural group has clearly recognizable physical features which it may have acquired through long isolation and inbreeding. Moreover, despite repeated immigrations into the peninsula from the heart of the continent, certain racial types have persisted there and absorbed the new blood from outside. Thus there is little doubt that the Slav-speaking folk who poured into the Balkans from the 3rd century AD, and particularly after the 6th century, were predominantly of Nordic race; yet the present day Southern Slavs belong principally to the Dinaric race—tall, dark, and with high and broad head and prominent nose—a type that is probably very ancient there and that is also found farther west in a highland zone extending through northern Italy, southern Germany as far as Austria, and Switzerland.

The Southern Slavs. These people, separated by the intrusive Magyars from their linguistic kinsfolk to the north, are united by ties of language and political allegiance. They are, however, subdivided by differences of religion, the Croats and Slovenes being Roman Catholics, the Serbs proper and Montenegrins mostly Greek Orthodox. Moreover, among the Dalmatians, Bosnians, and Hercegovinians, who are normally classed with the Serbs, there are in addition to the Greek Orthodox Christians considerable Roman Catholic minorities, and in the last two cases also a fair proportion of Muslims.

The Slovenes of Carniola are physically very similar to their Austrian neighbours. They are of low height, with small round heads of Alpine type, and include a good number of fair individuals. Among the Croats to the south, however, there are on the average more tall individuals and more broad heads, dark colouring, and prominent noses; and farther south still, among the Serbs, these Dinaric features become pronounced. Among the Serbs there is a preponderance of high broad heads, giving a mean cephalic index of about 85; the face tends to be long, the colouring of eyes and hair to be dark, and the nose to be high-rooted and straight or convex, with a downward-turning tip.

The Bosnians and Hercegovinians are, like the Serbs, on the average tall and broad-headed; but there is an interesting physical difference between the Roman Catholic and Muslim communities, doubtless the result of protracted inbreeding within each group. The former includes a larger proportion of tall, fair, and very broad-headed individuals.

Along the Dalmatian coast, the extreme broad-headedness of the population of the mountainous interior is not found, and among the inhabitants are more long-headed individuals of the Mediterranean race. Farther southeastward along the coast, from Istria to the borders of Albania, the population becomes on the average progressively taller and more dark in complexion.

In physique and culture, the Montenegrins, who live in rugged and isolated limestone mountains, are the most distinctive of the Southern Slavs. Although they speak a Slav Serbian language, they resemble the Albanians in their loyalty to their exogamous clans. They are probably the tallest people in Europe, and at the same time they are thickset and very heavily built. Their heads tend to be broad and high-backed and their noses large and beaked. In these respects they resemble the Dinaric type, which seems to be very old-established in the western Balkans. But they differ from the standard Dinaric type in having broad rather than high faces, a high proportion of light eyes and, in particular, a preponderance of red hair, red beards, and freckling.

Albanians. About half of this group lives outside the boundaries of Albania, for the most part in Yugoslavia. They are distinguished from the Slavs by language in the first place, the Albanian tongue being a mixture of elements from Illyrian, Thracian, Latin, Slavonic, and Turkic. They are also separated from their neighbours by religion and general culture, and there is no doubt that they represent the survivors of the native inhabitants of the Balkans in classical times, the Thracians, Macedonians, and Illyrians, who were driven back into isolation by the invading Slavs. Later, the Albanians became even more alien to the Slavs by adopting Islām from their Turkish conquerors in the late 15th century. Nearly all the Albanians in Yugoslavia and most of those in Albania are Muslim.

Distinctions between Albanians and Slavs

There are two distinct groups of Albanians, each with its own dialect, costume, and customs. These are the Tosks in the south and the Ghegs in the north and on the plain of Kosovo. The Ghegs comprise 10 tribes, each of which is subdivided into political groups known as *bairaks* and on a quite separate system into exogamous patrilineal clans, or *fis*. This *fis* system often leads to inbreeding within small areas, through the institution of cross-cousin marriage, with the result that the physique of the Ghegs varies from district to district. In the north, near the borders of Montenegro, a tall thickset type, like the characteristic Montenegrin, is found, while to the east the population includes more individuals of the Atlanto-Mediterranean strain, tall but not very heavily built, with long head, dark hair, and dark brown eyes. But the most characteristic physical type among the Ghegs, which there as elsewhere is probably by origin a hybrid, is the Dinaric, commonest in the tribe of Dibra. This type is of medium height, with a characteristically narrow and convex nose, light brown eyes, and dark brown hair, and above all with a very broad and flat-backed head. This extreme occipital flattening may be partly artificial, resulting from the use of a hard cradle, but it appears that this is not the whole explanation and that it is also an inherited racial character.

The Tosks, like the Ghegs, are extremely broad-headed; in fact, the mean cephalic index measured in a Tosk district of southwestern Albania, 90.8, is the highest recorded in Europe. But the broad-headed character of the Tosks is of a different type from that of the Ghegs. The head is globular in shape, with high and bulbous forehead, and the back of the head is not flattened. In addition, the nose lacks the high bridge and depression of the tip that are often observed among the Ghegs. The Tosks, in other words, belong to the Alpine variety of the broad-headed people of Europe rather than to the more specialized Dinaric variety, and in this respect resemble closely the inhabitants of southern and central France.

Greeks. Inside Greece the population is by no means homogeneous. Whole districts, even as far south as Attica, still speak Albanian, and wide pastures on the Pindus mountains are still the preserve of Romanian shepherds. The common loyalty of the modern Greeks, therefore, is linguistic and religious rather than territorial.

Variation among Greek communities

Among the Greek-speaking peoples can be traced a variety of communities of different physical types. These physical distinctions are doubtless the result of segregation and inbreeding within the various small compartments into which the territory of Greece is naturally divided, and it is likely that these distinctions were equally noticeable in classical times. Differences of head form from region to region afford a good illustration. In the high western slopes of the Pindus mountains or the district of Epirus, as well as in Macedonia, the Greek population, like their Tosk neighbours in Albania, is extremely broad-headed, with mean cephalic indexes of about 85–88. In the Peloponnese, Attica, and the Ionian Islands, on the other hand, the populations have a mean cephalic index of 81 or 82, while in Thessaly the inhabitants are comparatively long-headed, with a mean cephalic index of 77. There is also a regional variation in the colouring of the complexion, the Macedonian Greeks, for example, being noticeably fairer than those of the Ionian Islands and the Peloponnese.

Briefly, three main types can be traced within the Greek population. The first, with heavy and dark beard, strong brow ridges, and eyebrows which run together, clearly belongs to the Alpine race. The second is taller, with a fairly long head, straight nose, and dark brown hair and eyes, and represents a tall variety (Atlanto-Mediterranean) of the Mediterranean race. Third, much less common than the others, is a Dinaric element, with very dark hair, flat back to the head, and narrow facial features. Very occasionally also a blond Nordic strain can be observed.

The Greek population, therefore, is of mixed ancestry. Yet to judge from measurements taken in a Greek community in Boston, there are certain special physical features which have become characteristic of these people and serve to distinguish them from their neighbours in the Balkans. Such are the broad jaws, high nasal bridge, wide cheeks, and broad nose with a tendency to turn up rather than down at the tip.

Bulgarians. These people were members of a heathen Ugrian tribe who entered the Balkans from eastern Russia in the wake of the Slavs and soon afterward adopted Greek Orthodox Christianity and abandoned their language in favour of the Slavic tongue of their predecessors. Later, the country fell under Turkish rule, and Tatar and Circassian settlers entered in considerable numbers. Despite these repeated incursions, however, the bulk of the population of modern Bulgaria still belongs to the very ancient tall variety of the Mediterranean stock, usually known as the Atlanto-Mediterranean type, which inhabited the territory in Neolithic times. The chief characteristics of this type are dark colouring, moderate to tall stature, narrow face, and, above all, a long head. The Bulgarians, on the average, are among the longest-headed folk of the Balkans, and only in the west, in the region of Macedonia where they come into contact with the Albanians, is there any significant broad-headed element in the population. The second chief racial constituent in the Bulgarian population is the Neo-Danubian, a moderately broad-headed race, generally of light complexion and with a characteristic concave snub-tipped nose. A slight flattening of the face suggests Mongoloid admixture, though the basic element of this type is undoubtedly Nordic. It is found also in central Finland and in the black earth region of Russia, and was probably introduced into Bulgaria during the Slavic and Ugrian invasions.

The Vlachs

Romanians. These people, who are locally called Vlachs, are the most scattered of all the people of the Balkans. They have a common language, of Latin derivation, and a common pastoral nomadic heritage, but their racial type varies widely from region to region. In Walachia, for example, the main element in the population resembles that of Bulgaria to the south, being of medium height, with dark hair and eyes, narrow forehead, and nose and head of medium breadth. These features indicate a tall variety of the Mediterranean race. Farther north in Moldavia, where the Romanian plain abuts on the black earth region, there is a higher proportion of individuals of the Neo-Danubian race, fairer and with flatter and broader faces and more snub noses. To the west, however, just over the crest of the Carpathians in Bukovina, the Vlach population belongs to the Dinaric race. Their heads are appreciably broader and larger than those of the plainsfolk, their stature taller, their faces longer and broader, their noses larger and more prominent, and the backs of their heads much flatter. The Vlachs of Macedonia and Istria appear also to belong substantially to the same tall, dark, prominent-nosed, and broad-headed Dinaric race.

In short, the Vlachs have no racial uniformity but represent the descendants of the aborigines who, during the 150 years of Roman rule in the province of Dacia, in the 2nd and 3rd centuries AD, adopted something of the language and civilization of their rulers. After the withdrawal of Roman rule they were scattered by the various incursions of Goths, Slavs, Bulgars, and Turkic peoples, but survived in isolated mountain districts in many parts of the Balkans, especially in Macedonia, northern Greece, and southern Albania, where they took to a pastoral, seminomadic economy. Their physical features no doubt varied from one district to another at the time when they first came under Roman influence and have since become even more localized as a result of intermixture of the various groups with their immediate neighbours.

The Slavic word *vlach* means "foreigner," and it appears to be related to the terms Welsh and Walloon in western Europe.

Not only are the Vlachs or Romanians proper, as defined by language and culture, widely scattered through the Balkans outside Romania, but the inhabitants of Romania itself are very mixed, and in the district of Dobruja (a small plateau enclosed between the Black Sea and the lower reaches of the Danube), for example, there are besides the Vlachs, representatives of the following peoples: Bulgars, Ottoman Turks, Gaguz, Armenians, Kurds, Circassians, Gypsies, and Jews. These last two groups form important minorities throughout the Balkans, though their numbers were greatly reduced in World War II and, afterward, by emigration. The surviving communities are particularly concentrated in Romania.

Balkan Jews. The Jewish communities in the Balkans are predominantly made up of Sefardim, that is, the descendants of those driven out of Spain in the persecution of 1492. Some still speak a form of Spanish known as Ladino, and until World War II they often preserved a dress and customs that recalled their origins in Spain. After the war, the much-reduced community was largely assimilated into the general urban population of Romania. Physically they must be classed, there as elsewhere in southern Europe and the Near East, as a variety of the small Mediterranean race. The chief type is slender and delicately built and is of moderate stature with fairly long head, long narrow face, straight or convex nose, and dark colouring. A second type has a shorter and broader face, well-developed chin, thicker lips, short and straight nose, heavy eyebrows, and deep-set eyes and is altogether of broader and more muscular build. The Jewish communities have generally preserved their racial characters, which are the outcome of a stable combination of several varieties of the brunette Mediterranean race. Although darker than most of the people of central Europe, they are fairer than their neighbours within Romania. Among the Jews of Thessaloníki a minority are of blond colouring, although in other physical features they are no different from their coreligionists.

Balkan Gypsies. The Gypsies of Europe are especially concentrated in the Balkans, with a majority living in Romania and Hungary, and many in Bulgaria, Greece, and Yugoslavia. A great many have now settled and have intermarried to a varying extent with their neighbours. But wherever they retain their nomadic habit, the original racial type can still be traced. This type is of short stature, with long and low-vaulted head, small and fairly long face, narrow nose, straight black hair, and very dark eyes and complexion. The Balkan Gypsies are characteristic of the dark-complexioned small Mediterranean racial type, which is at home in India, and this identification is confirmed by their language and traditions. Their speech is basically Indian, a derivative of Sanskrit, though it includes words borrowed from the languages of numerous countries, such as Iran, Armenia, and Greece, across

which they travelled on their way through the Near East or during their wanderings in Europe. They appear to be the descendants of pariah tribes of northern India who moved to the west early in the present millennium and entered Europe about AD 1300. In the Balkan countries some of them were made landed serfs, while others were allowed to wander and to practice their traditional trades as tinkers, wood-carvers, gold panners, minstrels, and sorceresses.

Turks. The communities of Turkish-speaking Muslims in the Balkans, relics of the long period of Ottoman control, have been rapidly depleting in the 20th century because of migration back to Turkey. There has been a particularly large exodus from Bulgaria. Away from the small area of European Turkey, it is unlikely that many of these Balkan Turks are racially akin to the Ottoman conquerers. This is shown clearly by the range of physical types that is found among these European Turks. For example, those of Bosnia-Hercegovina belong to the local tall and very high- and broad-headed Dinaric race; while a series of measures taken in Sliven (Slivno) in Bulgarian Rumelia showed, on the contrary, that the Turks there were long-headed and in this and other respects very akin to their Bulgarian neighbours. Manifestly, very many of the Balkan Turks are indigenous folk whose ancestors, from compulsion or self-interest, adopted the Turkish language and religion.

An exception must probably be made for the Yürüks, or seminomadic shepherds, such as inhabit the southern slopes of Mt. Ograzden in Yugoslavia. There, as in Turkey, these people preserve in a very pure form the language, economy, and customs, and probably also the physical type, of the old Turks who emigrated from central Asia.

The Bektāshī communities of Turks, however, are much more clearly of local origin, as is suggested by their doctrines and institutions, many of which are compromises with the practices of Greek Orthodox Christians. Such are their belief in the Trinity and their celebration of Christmas Eve. Since the dissolution of the dervish orders in Turkey in 1925, the Bektāshī communities of the Balkans, in particular of Albania and parts of Yugoslavia such as the district of Strumica, have become the main survivors of this enigmatic sect. A colony was also established in New York City by immigrants. The true Bektāshī are the members of an order of Muslim dervishes, who follow the rule of Ḥājjī Bektāsh Walī, their 13th-century founder; but the villages which take instruction from these dervishes also adopted this name. The chief ways in which the Bektāshī differ from orthodox Muslims—although they arrogate to themselves the title of Sunnī—are in their devotion to ʿAlī and to his children Ḥasan and Ḥusayn, in memory of whose martyrdom they fast for the first 12 days in the month of Muḥarram; in the privileges they accord to their women, who are even allowed to become dervishes; in their religious proceedings; in their custom of worshipping in tekkes (monasteries) and turbehs (mausoleums) rather than in mosques; and in their open compromises with the feasts and customs of their Christian neighbours.

Ethnic survivals in the Balkans. The Bektāshī villages provide an example of the very mixed cultural traditions of the Balkans. In this peninsula people with varied customs and beliefs have come together and though they often preserved their old racial and political differences and rivalries, they have at the same time interchanged ideas, beliefs, crafts, and artistic symbols. These countries may indeed be regarded as a somewhat disorderly ethnological museum, in which traces of numerous people are haphazardly preserved. Mary Durham's studies of Balkan jewelry and tattoo patterns, for example, demonstrate how there are often retained side by side designs from Muslim and Christian sources, and even from ancient Illyrian and Mithraic sun cults.

Demography. Due to the decreasing isolation of rural communities and to the rising standards of living, particularly among the ever-increasing proportion of the population which is urban-dwelling, both birth and death rates have tended to decline.

Population density within the four countries of the peninsula is relatively low, compared with most of central and western Europe. There are, of course, wide differences within the countries, from region to region, which result from differences in altitude, relief, and agricultural productivity. In most of the mountainous areas, but particularly in the Greek Pindus ranges, in northern Albania, and in Montenegro, densities are very low, while high densities are reached in the plains of northern Yugoslavia (particularly the former Serbian area) and in the plains and basins of central Bulgaria.

History

BALKANS TO 1815

The Balkan Peninsula has been inhabited since Lower Paleolithic (Old Stone Age) times (*c.* 200,000–100,000 BC), but how continuously is not yet known. The Mousterian industries of Neanderthal man from the early part of the last glaciation (*c.* 70,000–40,000 BC) are well represented, followed by those of the Upper Paleolithic Aurignacian and Gravettian, *c.* 40,000–*c.* 8000 BC, associated with the physical remains of modern man (*Homo sapiens*). Evidence of habitation layers of postglacial food-gatherers in caves includes the chipped-stone industry developed from the Gravettian, and it continues into the beginning of the food-producing era.

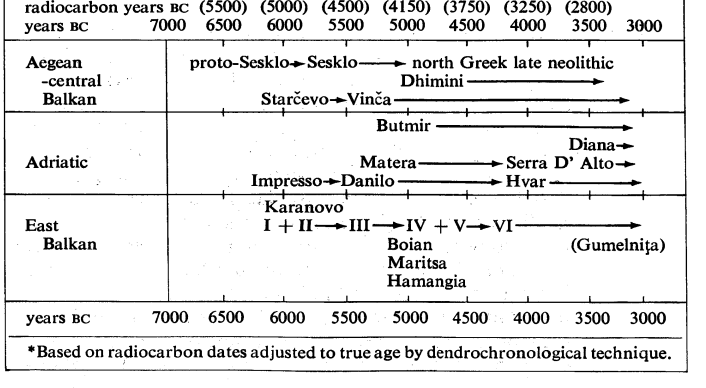
The development of the Neolithic Period around 7000 BC was characterized by increased sedentariness and reliance upon domesticated plants and animals, permitting the existence of larger demographic units. From this period onward, one can speak of the Balkan culture in more detail.

Old European civilization. With east central Europe, the Balkans formed the core of the earliest European civilization, Old Europe, dated to the period 7000–3500 BC. Old Europe is not merely a provincial reflection of the great civilization of the ancient Near East but rather a distinct, unique culture with roots of its own. The Old Europeans discovered the use of copper in the 6th millennium BC and formed population aggregates that often amounted to small townships, which inevitably involved craft specialization and religious and governmental institutions. The Old Europeans also evolved a written script.

Regional groupings

Old European civilization may be divided into five regional groups: (1) the Aegean and central Balkan, (2) the Adriatic, (3) the middle Danube, (4) the east Balkan, and (5) the Moldavian–western Ukrainian. The first region embraces the culture of the Neolithic Period in what is presently Greece, Yugoslavia (except the Adriatic seaboard), southeastern Hungary, and western Romania, succeeded by the Chalcolithic Vinča, Butmir, and Tisza cultures. The second area consists of the painted-pottery complex on either side of the Adriatic, including the Matera-Serra d'Alto in southern Italy and the Danilo in Dalmatia. The third variant includes the Lengyel culture of the middle Danube Basin, which displays similarities to the Adriatic complex. To the eastern bloc belong three cultural groups with individual traditions in art: the Hamangia on the Black Sea coast; the Karanovo in Thrace, and its Chalcolithic successors, and the Boian and Gumelnița complexes, which occupied most of modern Macedonia, Bulgaria, and Romania. In the fifth region, Moldavia and

Chronology of Balkan Neolithic and Chalcolithic Complexes*



*Based on radiocarbon dates adjusted to true age by dendrochronological technique.

Prehistoric Balkan sites.

the western Ukraine, local Neolithic culture was followed by the formation of the Cucuteni (Tripolye in Russian) civilization. On coastal and inland plains, mounds created by the accumulation of cultural debris attest to the permanence of the farming communities, whereas in the Danubian region there is evidence that a wandering agriculture was practiced. Houses were arranged either around a central structure or in parallel rows. The typical house was a rectangular structure of timber posts completed with wattle-and-daub walls, semisubterranean dwellings, trapezoidal structures; cultures living in dry Mediterranean conditions built characteristic houses with mud-brick walls, stone foundations, and internal buttresses.

The Neolithic complex left a remarkably uniform artifactual expression, including polished-stone axes, adzes, and small ornaments; clay stamp seals; bone spatulas; bird-shaped vases; and clay models of shrines. Clay and marble votive figurines reflect several types of goddesses, notably the Snake Goddess, the Bird Goddess, and the Great Goddess, and their epiphanies in the form of a bird, toad, bee, butterfly, deer, bear, and dog. A male god is also represented. Loom weights, spindle whorls, and woven-mat impressions on pot bases attest the production of textiles. Typical of the fine pottery is a highly burnished, hemispherical bowl, often on a ring base; the vessel is commonly coated with a red slip, with red-on-white or white-on-red painted decoration. Cemeteries were unknown and interments are found within the settlement. Infants were buried inside egg-shaped vessels.

Differentiation into regional groups progressed steadily. By the mid-6th millennium BC, there appeared the first traces of metallurgy, the Neolithic thus yielding to the Chalcolithic, the age of a mixed copper and stone technology.

Proto-Sesklo and Starčevo

Central Balkan region. The earliest Neolithic occupation of the Aegean and central Balkans is differently named in each of the modern European countries in which it is distributed—Proto-Sesklo, Starčevo, Körös, Criș—but these are regional variants of a widespread, comparatively uniform cultural complex.

Early Neolithic technology

The subsistence bases relied upon emmer, a primitive variety of wheat, and the domesticated sheep or goat. The identical bone structure of the sheep and goat makes it impossible to distinguish which animal is represented by the remains, but, because neither species was indigenous to Europe in the terminal Pleistocene, or Ice Age, the animals in question must have been introduced from their natural habitat in Anatolia and the Near East. Though their initial introduction was probably a function of ethnic movement across the Aegean, subsequent diffusion may have resulted from acculturation of the Mesolithic population under the stimulus of contact and trade with early farming communities. Radioactive-carbon dates and typological study show the Thessalian and Macedonian pre-pottery and early ceramic sites to be Europe's earliest Neolithic settlements.

Excavations conducted in the late 20th century on the Danube above the Iron Gate provide new insight into the process of economic transition to a Neolithic way of life. The early levels comprise permanent settlements without ceramics; their subsistence was based on fishing and on hunting with the dog, the only domesticated animal. The subsequent phases belong to the characteristic central Balkan Neolithic (Starčevo), with the domesticated sheep or goat and wheat. Starčevo farmers increasingly exploited domestic cattle and pigs and hunted and fished in response to the better forested environment. Two of the systematically excavated sites in this region are Lepenski Vir, excavated by D. Srejović, and Padina, excavated by B. Jovanović.

Sesklo

Sites containing Sesklo materials have been discovered throughout Greece; most of these were newly founded settlements, implying an increase in population density. The beginning of the Sesklo phase is defined by the appearance of fine white-slipped pottery, decorated with flame and stair designs, painted in dark red. Shapes included globular-necked jars, footed bowls, straight-walled bowls, and cups and jugs.

The site of Otzaki, in Thessaly, has disclosed a three-phase development of Sesklo ceramic art. Excavation has uncovered two close-standing rows of square houses constructed of rectangular mud bricks. These houses frequently rest upon a stone foundation and are occasionally strengthened by internal buttresses.

The Vinča sequence is best documented at the eponymous site, situated 14 kilometres (nine miles) east of Belgrade, which has yielded 23 feet of stratified Vinča deposit overlying the Starčevo levels excavated by the Yugoslav archaeologist Miloje Vasić. Fine Vinča ceramic wares are burnished in orange or black and decorated with a shallow linear channeling.

Vinča

The origin of the Vinča black burnished ware need be sought no farther afield than the Maritsa Valley of central Bulgaria. At its greatest extent, the Vinča complex occupied the central Balkans from the Rhodope Mountains north as far as the Banat; west to northeast Bosnia; and eastward to western Bulgaria, southwest Romania, and Transylvania. In architecture and subsistence economy, no marked difference has been discerned between Vinča and the Neolithic Period. Wattle-and-daub architecture continued to be employed, with evidence of the use of split planking in floor and wall construction. Floors were clay plastered. Increasing trade is evidenced by the widespread occurrence of Aegean Spondylus shell, which was used for beads and bracelets. The archaeological data indicate gradual social and economic change and an attendant population increase. Vinča sites occupy as much as 20 acres of river terrace, with houses organized into lined streets.

Concomitant with this expansion was a remarkable artistic development of the Vinča culture, marked by an increase in the quality and number of figurines and other objects that served a ritual function. These symbolic figures testify to an important intensification of spiritual life during the period. Vinča art remained distinct from that of neighbouring groups throughout a millennium or more; and its gradual evolution reflects a remarkably stable and well-organized social structure.

Linear writing

The 1961 discovery at Tărtăria in Transylvania of three clay tablets inscribed with pictographs and linear signs has encouraged the explanation of import from Mesopotamia at about 3000 BC, but this hypothesis has been invalidated by the mutually reinforcing evidence of stratigraphic typologies and radioactive-carbon dating, which locate the early Vinča period in the latter part of the 6th millennium

BC. The Tărtăria tablets were found in a secure Vinča context, and inscribed linear signs on figurines, vases, and spindle whorls of definitively local manufacture appear to confirm that the late 6th and 5th millennia witnessed the development of linear writing. Equivalent inscriptions occur in the east Balkan civilization.

Tisza

Tisza settlements, named after the River Tisza, contemporary to Vinča and distributed in eastern Hungary and the Yugoslav Banat, are smaller than Vinča, consisting of rectangular and semisubterranean dwellings with pitched roofs apparently resting at ground level. Local domestication of cattle on a scale unprecedented in temperate Europe was proved here by S. Bökönyi.

Butmir

The inland Bosnian Butmir variant was strongly influenced by the Vinča culture and was already producing copper. It is best known from the eponymous site at Butmir (Sarajevo), excavated between 1893 and 1896 by W. Radimsky and F. Fiala, which yielded a significant number of sculptures and fine ceramics, including black-polished ones with incised triangle and spiral decoration encrusted with red and white paint. Typical shapes were biconical bowls; high-pedestal, footed "wine" cups; and hole-mouth pyriform vases. At Obre II, a Butmir site in a forested upland environment near the upper Bosna, excavated by A. Benac and M. Gimbutas, faunal analysis revealed that more than half the domestic animals were cattle and that a considerable amount of hunting was practiced. Einkorn wheat was the chief cereal crop, supplemented by barley and lentils; at the Butmir site itself, pear and apple pips have been found.

At Obre II, wattle-and-daub houses of two rooms, about 49 feet long, containing beehive-shaped bread ovens, wooden platforms, ash pits, storage and ritual vases, spindle whorls, loom weights and tools, were revealed. The calibrated radioactive-carbon dates from this site place the three periods of the Butmir culture at 5000–4000 BC.

Impresso

Adriatic. The Early Neolithic culture of the Adriatic region is known as the Impresso complex; it is characterized by grit-tempered wares impressed with shells or with a stabbing tool. The simple ornamented bowls and the farming economy they served are believed to have developed as a result of rapid diffusion, a corollary to maritime movement and trade along the Adriatic littoral. The Impresso culture of western Yugoslavia represents only a part of a complex widely dispersed throughout the Mediterranean world. The initial stimulus to agriculture probably came through eastern seaborne contact with the indigenous Mesolithic peoples of the Mediterranean coastlands. No considerable ethnic migration need be invoked in explanation.

Impresso sites occupy both caves and open settlements. The Dalmatian open-air settlement of Smilčić near Zadar consisted of wattle-and-daub houses and was enclosed by a deep ditch.

Danilo

The beginning of the Danilo period is marked by the appearance of elaborate red-on-cream painted wares, with geometric motifs akin to Sesklo designs. This was a flourishing Neolithic culture of the end of the 6th and 5th millennia BC, with contacts with Greece and southern Italy.

Middle Danube. The local indigenous population of Cro-Magnon type prevailed in the middle Danube until the 5th millennium BC, when the periodically migrating farmers were supplanted by the Lengyel complex, named after a site in western Hungary that displays different architectural and artistic traditions. The Lengyel physical type contrasts markedly with the central European Cro-Magnon and is related, rather, to the people of the Adriatic region. Lengyel settlements are fortified with wide ditches and palisades, and the typical painted piriform vases and "fruit stands" have analogies in the Danilo complex. These similarities may reflect an ethnic infiltration from the Adriatic to the Sava basin and the region east of the Alps. The villagers cultivated wheat, barley, and Italian millet; and in addition to tending sheep or goats, they hunted and locally domesticated both cattle and pigs.

East Balkan area. The east Balkan civilization began around 6000 BC with the first appearance of Neolithic occupation along the Maritsa Valley in Bulgaria. The best known sites are Karanovo, Azmak, and Kazanlŭk. The Karanovo sequence, phases I–VI, has become universally adopted as a chronological yardstick for the development of east Balkan civilization. (Roman numerals are used by archaeologists to designate the successive layers of occupation found at a particular site, in this case phase I representing the earliest known settlement and higher numerals representing later settlements.)

Karanovo I–III

In the villages of the Karanovo I–II period, rectangular one-roomed houses of 20–23 feet to a side, with thick, wattle-and-daub walls, an aligned plank floor, and an internal oven, were arranged in parallel rows. The ceramics show affinities with central Balkan wares, and a rich tool kit includes numerous millstones and sickles of deer antler with flint blades inserted like sawteeth. The plentiful remains of einkorn and emmer wheat, lentils, and bones of domesticated animals confirm the role of mixed farming. Karanovo III pottery typically displays a dark, lustrous black surface, sometimes decorated with plastic bands and knobs or linear incision. The most diagnostic forms are cylindrical and pear-shaped vases with single handles. Karanovo III complex is also known as Veselinovo, after another site in eastern central Bulgaria. Karanovo III corresponds to a considerably increased population; during this phase, elements of the Karanovo III assemblage were carried, probably by a movement of peoples, northwest into the lower Danube region and south beyond the Rhodope Mountains to eastern Macedonia and Thrace. In the north, the intruders came in contact with the Hamangia group on the coast of the Black Sea and with the settlers of the central European Linear Pottery culture. This contact along the lower Danube resulted in the amalgamation of the individual ceramic traditions within the assimilating expansion of the Karanovo III population, bringing about the formation of the "Boian" tradition, a northern variant of east Balkan civilization (the name Boian is derived from an island settlement near the Danube south of Bucharest). The "Vădastra culture" in western Romania represents a sister branch of the same civilization.

Boian

The Boian farmers cultivated einkorn, millet, beans, and flax; they also domesticated and kept cattle, pigs, sheep, and goats, but cattle were by far the most important. Houses arranged in parallel rows in compact villages were fortified by ditches. The Boian pottery was ornamented with excised, white, encrusted spiral and meander designs. There are five subbases of Boian.

During its second phase, Boian material culture spread as far as Moldavia in the northeast, influencing the formation and development of the Cucuteni civilization; it was this vigorous phase that witnessed the first signs of metallurgy. A late Boian temple edifice with trichrome wall decorations and with two exquisitely painted pillars inside has been discovered in the island site of Căscioarele, southeast of Bucharest. Skeletons from a large cemetery at Cernica (near Bucharest) are predominantly of a small-statured Mediterranean type.

East Balkan culture

The advanced stage of east Balkan culture is the Gumelnița civilization, known from at least 500 tells (mounds of cultural debris) in Romania, Bulgaria, and eastern Macedonia. Gumelnița itself lies southeast of Bucharest. Other important sites are: Căscioarele; Tangîru and Hîrșova, in the lower Danube region of Romania; and Ruse and Hotnica, in northern Bulgaria. North of the Aegean, the most noteworthy sites are the mounds of Sitagroí and Dikili Tajh on the plain of Drama. Sedentary Gumelnița communities occupied compact villages or small towns for a millennium or more. Ceramic models of houses show gabled roofs, round windows, and plastered walls with designs painted in red, yellow, and white. These models all attest to the existence of two-story buildings. A model of a temple edifice discovered on the Danube island of Căscioarele shows four shrines standing on a large substructure.

Gumelnița

Gumelnița fine ceramic vessels are distinguished by their technological sophistication. Graphite painting required special kilns to prevent oxidation of the graphite.

Metal production and trade exhibited steady growth during the period: needles, awls, fishhooks, and pins were of copper, and, at the end of the period, axes and adzes were manufactured, a development also found in the

Vinča, Tisza, Lengyel, and Cucuteni civilizations. Workshops of flint, copper, gold, Spondylus shell, and pottery imply craft specialization and increasing division of labour. Gold was obtained from Transylvania and copper from the Carpathian sources. Evidence of linear writing is found on ceramics throughout the duration of Boian–Gumelnița civilization.

Hamangia

The Hamangia settlements and cemeteries are found along the coastal strip between northern Bulgaria and the western Ukraine. Most information comes from the cemetery at Cernavodă, where some 350 graves have been excavated by D. Berciu. Examination of the skeletons has revealed that the population was predominantly Mediterranean; but also found was a distinct local type with a short, broad skull.

The Hamangians practiced mixed farming; they cultivated wheat and vetch and herded sheep and goats, cattle and pigs. Deepsea fishing appears to have been important to early subsistence.

Figurine art characteristically shows a standing female, breasts and buttocks well-developed, with a columnar head and neck, lacking facial features. Less schematic are the exceptional pair of seated figures from Cernavodă—a male "Thinker" and a comfortably relaxed female, both masked. The earliest ceramics were impressed with cardium shell. Ornaments, found abundantly as grave goods, include huge bracelets and beads of Spondylus shell. The Hamangian complex was superseded by the east Balkan Gumelnița civilization.

From the Bronze Age to the coming of the Slavs. *Indo-Europeanization during Bronze Age.* Old Europe was developing into an urban culture, but its power was cut short by a steadily increasing infiltration of the seminomadic pastoralists from the Russian steppes in about 3500 BC.

Kurgan culture

Their culture is called Kurgan because of their burial in tumuli (*kurgan* means "barrow" in Turkic and Russian), covering graves in deep shafts. It reveals elements of the hypothetical mother culture of the Indo-European speakers as reconstructed with the help of common words. In the period *c.* 3500–2300 BC, their presence is traced in Danubian Europe, and, after 2300 BC, their arrival is documented in the Aegean and Adriatic regions. Changes in social structure, economy, and religion show that people of different background imposed their ways of life. Centres of the new power are witnessed by strongholds such as Vučedol (at Vukovar in northern Yugoslavia) and Nagyárpád (at Pécs in southern Hungary). Cultural uniformity is evidenced over the Danubian plain down to Macedonia before 3000 BC. The new culture with persisting elements of Old European substratum is known by the names of Cernavodă in Dobruja, Ezero in Bulgaria, Coțofeni in Transylvania, and Baden in the middle Danube region.

By 2000 BC, nuclear groups parent to Indo-European Illyrian-, Armenian-, Venetan-, Phrygian-, Mysian-, Dacian-, Thracian-, and Greek-speaking units were formed; their cultures were typified by military aristocracies, hill forts, small villages, horses, and vehicles. Their archaeological complexes are dubbed Otomani and Wietenberg in Transylvania; Monteoru in Moldavia; Tei in the lower Danube region and Thrace; and Incrusted Pottery in Pannonia and northwestern Yugoslavia.

The centres of military power and of bronze metallurgy in the mid-2nd millennium were in central Europe and Greece: the Unětice-Tumulus in the upper and middle Danube region; the Otomani-Wietenberg in the Carpathian Basin; and the Mycenaean culture in Greece. Trading of transcontinental Baltic amber and Transylvanian and Mycenaean gold and bronze is richly evidenced from the 17th through the 15th century BC.

Around 1400 BC the Tumulus people extended their influence over the whole middle Danube Basin and to the Adriatic coast. As Urnfield peoples, they caused an upheaval in the Balkans in Anatolia and in the east Mediterranean region before and around 1200 BC. Their graves and pots are found in Macedonia, while their weapons are found in Greece, Crete, Cyprus, Syria, Palestine, and Egypt. The destruction of the Mycenaean culture is attributed to the middle Danubian Urnfield invasion through the Balkan Peninsula and the Adriatic Sea. The Urnfield people were Indo-European speakers, parent to Venetan and perhaps Phrygian, Mysian, and Armenian. In past research they were assumed to be Illyrians, but this name can be used only as a broad cover name and should not be confused with the Illyrians proper, known since the time of Herodotus in Dalmatia, Bosnia, and Hercegovina. This period of raids and migrations, instigated by central European Urnfield peoples (misleadingly called an "Aegean migration"), resulted in a Phrygian, Mysian, and Armenian exodus to Anatolia, the destruction of the Hittite Empire, the Dorian infiltration into Greece, and the introduction of iron technology and a new ethnic configuration in the Balkans.

Illyrians and Thracians during 1st millennium BC. Illyrians proper occupied western Yugoslavia, Albania, and Epirus, with the Morava–Vardar line as their approximate eastern limit; in the northeast they extended into southwest Romania. Tribes north of the Sava River, Venetans and Pannonians, are usually treated as northern Illyrians. The general name for the period from 750 to 450 BC is "Illyrian Hallstatt." The exploitation of iron ores, particularly in northwestern Bosnia, increased commerce. From the 7th to the 5th century BC, strong tribal centres emerged, together with a powerful aristocracy, as witnessed by extravagantly furnished royal burials. Autochthonous traditions were maintained even after an exchange with Greek merchants and colonizers of the Adriatic coast started around 600 BC. Glasinac, an Illyrian centre near Sarajevo, displayed an uninterrupted development until the period of Celtic influence in the 4th century BC.

The Iron Age in Illyria

From the late 6th and early 5th centuries, the Greek influence is evidenced by masterpieces of Greek workmanship, which arrived through southern Italy. To this period belongs the treasure of Trebeništa at Lake Ohrid, which includes a large bronze krater, or mixing bowl. Greek influence increased markedly in the 4th century BC after the foundation of towns on the islands and coast of the Adriatic, among them Issa (Vis), Pharos (Hvar), Corcyra Melaina (Korčula), Salonae (Solin), Epidaurus (Cavtat), Iader (Zadar), and Tragurion (Trogir).

Written evidence of an Illyrian kingdom on the borders of the Greek world appeared at the end of the 5th century BC; its capital was at Skodra (Skoútari). Before Illyria was annexed to the Roman Empire in 168 BC, it was ruled by 15 monarchs carrying Illyrian names.

The Thracian culture formed of the Bronze Age substratum mixed with that of the seminomadic peoples from the Ukraine. The ethnic infiltration is shown by the appearance of an eastern archaeological complex called Noua, dated to the 12th century BC. The cultural melange that produced the Basarabi–Babadag complex of the Hallstatt era is held to be the basic component of the Daco-Thracian culture. Repeated incursion of groups of warriors from the northeast occurred at the end of the 8th century BC, bringing with them Caucasian elements in art and weaponry. This period is called Thraco-Cimmerian; it coincides with the conquest of the Cimmerians north of the Black Sea by the Scythians, but the actual ethnic infiltration of the Cimmerians into the Balkans is not evidenced. The Scythians proper appeared in the middle of the 6th and the beginning of the 5th centuries BC; but it was not until the beginning of the second half of the 4th century BC that the Scythians tried a violent and massive penetration into the territory. Contacts with the Scythian world contributed to the wider diffusion of iron metallurgy and the formation of a distinctive Thraco-Scythian style of art.

At the end of the 7th and in the 6th century BC, the Greeks founded several colonies along the western shore of the Black Sea: Apollonia Pontica (Sozopol), Mesembria (Nesebŭr), Odessus, Callatis (Mangalia), Tomis (Constanța), and Istrus (Hystria). The Greek cultural influx was of fundamental importance. It included the potter's wheel, which was in use by the local inhabitants at the end of the 6th century BC, and the striking of coins, which the Greeks themselves borrowed from the Lydians in Anatolia. The Greeks were interested in timber and wheat from Moldavia and the Ukraine and in silver and gold from the mines of southwestern Thrace. Greek influence steadily increased. At Panagyurishte, in central Bulgaria, a gold

Greek colonies on the Black Sea

service weighing more than 13 pounds (six kilograms) included vessels of superb workmanship by Greek artists of the late 4th century BC. Dionysian festivals and worship of Bendis (Artemis) and Ares are ascribed to Greek influence.

Dacians, the northern Thracians, are known from Greek sources of the 4th and 3rd centuries BC. A good deal about southern Thracian life and manners is revealed by the Greek authors Homer, Herodotus, Thucydides, and Xenophon. Thracian nobles are shown on Greek vases wearing caps of fur, decorated cloaks, and leather boots. The society was classified into chieftains, warriors, commoners, and slaves. In the mid-5th century BC, the first known Thracian state emerged in the Maritsa Valley, ruled over by the Odrysian king Teres. According to Thucydides this state enjoyed a large revenue and general prosperity. The Odrysian king Sitalces allied himself with Athens and in 429 overran Macedonia. The 4th century marks the final flourishing of the Thracian civilization. Before mid-century Philip II of Macedon brought unity to the area between southern Thrace and Albania and a new power—Macedonia—emerged. Philip's son Alexander III the Great crossed the lower Danube in 335 BC and attacked the Getae.

By about 300 BC, the Getae of Dobruja and the lower Danube, a people akin to the Dacians of Transylvania and Moldavia, combined to form what was virtually a state, an evolved type of military democracy having as its leader the *basileus* Dromichaites, who in 292 conquered Lysimachus. Greek cities on the coast of the Black Sea found protectors and allies in the Getian *basileis.*

Arrival of the Celts

About the beginning of the 3rd century BC, bands of Celtic warriors appeared east of the Tisza and advanced toward Oltenia and Transylvania. This movement can be traced by the evidence of their warriors' graves. The military superiority of the Celts, which had at its disposal a developed iron technology, insured their success. The Celtic element, which was present in Romania for several centuries, contributed to the development of iron metallurgy and to the adoption of the iron plowshare, which led to a great expansion in agriculture. Celtic influence left imprints in a novel Getian–Dacian silverwork and coinage. The southern wave of Celtic invasion—toward Thrace, Macedonia, Thessaly, and Greece—was broken at Delphi in the winter of 279–278 BC.

The Dacian state extended between the Tisza River and the lower Danube in the period between 200 and 31 BC, reaching its peak during the reign of Burebistas (60–44 BC).

Roman conquest and barbarian colonization. After Macedonia crumbled, Thracians, Illyrians, and Dacians were left to their own devices until a major Roman irruption at the end of the 3rd century BC. Thracians and Illyrians combined in self-defense, and there followed a struggle for 150 years. The southern Illyrians were conquered and annexed in 168 BC, but hostilities continued until AD 9, when all Illyrians were subjugated by the Roman general Tiberius (later emperor). The territory along the Adriatic, Dalmatia, Iapodes, and Liburnia, was united as the Roman province of Illyricum. In 29 BC the area between the Danube and the Balkan Mountains was conquered by Crassus and became the Roman province of Moesia, while the area further south was incorporated into the province of Thrace. The emperor Domitian's expedition against the Dacians in AD 85 consolidated his position north of the Danube. In 89 the king of Dacia, Decebalus, was forced into a collaboration with Rome against the Germanic Marcomanni, the Quadi, and the Iranian Sarmatians, the latter coming from the Pontic steppe. The complete conquest of Dacia in 106 followed two vigorous campaigns by the emperor Trajan in 101–103 and 105–106. The Dacian capital of Sarmizegethusa was captured and the province of Dacia was established.

The process of Romanization was swift. In Romania, Latin completely supplanted the Dacian and Thracian languages. Trade and handicrafts penetrated deep into the provinces. Trade routes connected the growing cities: the northern road led to Sirmium (Sremska Mitrovica on the Sava), Singidunum, Viminacium, and Ratiaria on the Danube; the southern led to Stobi (south of Titov Veles) and Scupi (now Skopje); and the eastern ran from Sardica (Sofia) via Naissus (Niš) and Margum to Viminacium on the Danube; in the west, the road crossed from Salonae and Narona on the Adriatic through Bosnia to Sirmium. Commercial expansion reached its culmination in the course of the 2nd century. Rome became decentralized and its art cosmopolitan. At the end of the 3rd century, Sirmium became a capital of the empire. Diocletian built his palace near Salonae (near Split), and official art was created in Dalmatia, Pannonia, and Moesia. From the time of Marcus Aurelius, the empire lived under the protection of an Illyrian–Thracian army. When Constantine I established Constantinople as a second Rome, in AD 330, the eastern part of the empire split in two: Thrace was included in the prefecture of the Orient, whereas the rest of the peninsula was in the prefecture of Illyrium.

In the 3rd century there began an infiltration of the Goths into the peninsula. About AD 214 the Goths clashed with the Romans at the Dacian frontier and conquered Dacia. The Goths then attacked Thrace continually through the middle and late 3rd century. The Gothic state flourished for nearly 200 years until the Huns invaded and swept across the country.

The coming of the Slavs

The devastation of the empire by the Huns, followed by that of the Bulgars at the end of the 5th century and that of the Avars in the late 6th century, prepared the ground for a wide Slavic dissemination from their homeland north of the Carpathians in the Ukraine.

In the Middle Ages. With the settlement of the Germanic Ostrogoths in Italy and the disintegration of the Western Roman Empire in the final decades of the 5th century, the Balkans again came under the direct control of the Eastern Empire at Constantinople, at least in theory. But the barbarian invasions of that century had left the Balkan Peninsula a sparsely populated wasteland destined to attract new and more permanent settlers, the Slavs. An Indo-European people, the Slavs had settled in the valley of the upper Vistula in central Poland during the great prehistoric migration that brought most of the ancestors of the modern Europeans from Asia. Basically a sedentary people with a highly developed tradition of agriculture and animal husbandry as well as of hunting and fishing, the Slavs were easy prey to the highly mobile and often strongly united nomadic hordes that swept across eastern Europe. These barbarian tribes easily superimposed their rule over the native Slavic political system of democratically run clans, owing only loose allegiance to a chief, or *župan.* It was as vassals of such nomadic tribes that the Slavs spread into south Russia and Pannonia, a former Roman province that occupied parts of modern Austria, Hungary, and northern Yugoslavia. From these two areas beyond imperial control, the Slavs began their steady movement into the Balkans. By AD 517 groups of Slavs were crossing the lower Danube to raid in Thrace, Macedonia, Thessaly, and even Epirus. Almost simultaneously, others were moving into the northwest corner of the Balkans toward the Dalmatian coast. The massive military endeavours of the Byzantine emperor Justinian I elsewhere in the empire left no troops free to man the fortresses protecting the Danube River boundary of his realm. Imperial officials in Constantinople attempted to neutralize the northeastern arm of the Slavic pincers encircling the Balkans by offering the Slavs on the lower Danube the status of *foederati,* paid border guards for the empire. The Slavs, however, became vassals of the Avars, a new Turkic horde of nomads who appeared on the south Russian steppe.

Slav invasion of the Balkans

The Avars, akin to the Huns, attempted to follow the route of their predecessors. Driving the remaining Germanic tribes from the Pannonian plain, they established there a new state that loosely controlled the Slavs settled in modern Czechoslovakia, the Balkans, and even north of the lower Danube in south Russia. Both the Avars and their Slavic tributaries undertook plundering expeditions throughout the Balkans. Eventually, the invaders took effective control of many areas as far south as the tip of the Peloponnese and as far east as the great wall of Thrace, the first line of defense of Constantinople. Usually, the ravaging of territory was followed by permanent Slavic settlements being made in the devastated villages. It was

only after 591 that the Byzantine army, finally withdrawn from other theatres of war, attacked the Avar–Slavic armies in the Balkan Peninsula; but by then the situation was hopeless. The Byzantine troops committed to driving the Avars back beyond the Danube revolted in 602, and soon the emperor was paying tribute to the Avars while the latter's Slavic vassals moved unimpeded through the Balkan Peninsula, populating it anew. Remnants of the local agricultural population probably remained among the Slavic settlers, but many of the Greek-speaking and Latin-speaking inhabitants took refuge in and around fortified Byzantine cities, particularly along the coasts. The safety of these cities, however, was in no way guaranteed. Thessalonica, the leading city of Macedonia and the second-richest city in the Byzantine Empire, several times came close to falling to the Slavs. In the interior, no imperial islands remained. Singidunum (Belgrade), Naissus (now Niš in Yugoslavia), Sardica (Sofia), Sirmium (now Sremska Mitrovica in Yugoslavia)—the mighty fortresses guarding the middle Danube frontier—fell, with the surrounding countryside, to the Avars and Slavs. Contemporary sources often speak of the Balkans simply as Sclavinia (Slavdom). Byzantine control of this region had been so completely disrupted that the emperor Heraclius omitted the Balkan provinces from the administrative reorganization of the rest of his realm. Rather, to preserve his capital, Constantinople, Heraclius paid massive tribute to the Avar khan (chief) before he dared move against the equally serious Persian threat on the empire's eastern border.

In 626, however, the strength of the fabled walls of Constantinople and Byzantine naval superiority gave the Byzantines a wholly unexpected but decisive victory against a joint attack by the Avars and Persians. From this point on, both the Persian and Avar states began their rapid decline. Already weakened by a revolt of their Slavic subjects in what is now Czechoslovakia, the Avars were further threatened when Emperor Heraclius invited two strong tribes from beyond the Carpathian Mountains, the Croats and the Serbs, to settle in land held by the Avars in the northwest Balkans. Aided by the Byzantine navy and the few fortified imperial cities left on the Dalmatian coast, these tribes took control first of Dalmatia and then of the territory that, after their occupation, bore their names, the Croatia and Serbia of modern Yugoslavia. Once settled, theoretically as vassals of Byzantium, they mingled with the earlier Slavic settlers.

Return of Byzantine power

For the great new empires that emerged in the early Middle Ages from settlement of the invading people, see separate Balkan country articles.

The 11th century saw further invasions of the Balkans by people from the Volga region, namely the Pechenegs and the Kumans. At the beginning of the 13th century, however, Greek Byzantine rule over the south of the peninsula was replaced, except in Epirus, by Latin, through the diversion of the Fourth Crusade to Constantinople.

Under Ottoman rule. *Ottoman conquest.* The Ottoman Empire had its origins in a small Turkish emirate established in the second half of the 13th century in northwest Anatolia, near the Sea of Marmara and close to the borders of the Byzantine Empire. The name Ottoman was derived from Osman I, a Turkish chieftain who ruled the emirate from 1281 to 1324. An active and aggressive military organization of dedicated frontier warriors, this state first established itself on the European mainland in 1354 on the Gallipoli Peninsula. In 1362 the Ottoman armies took Adrianople, which soon became the Ottoman capital. From this centre the Ottoman armies moved up the Maritsa and Vardar valleys against the Christian states of the Balkans.

The early Ottoman successes were made possible by their superior military power that was based on a mobile light cavalry and by excellent leadership and organization. The Christian powers were hampered by their failure either to cooperate among themselves or to secure effective European assistance. Pope Boniface IX's army, sent in an attempt to aid the Hungarians, was crushed by the Ottoman army at Nicopolis in 1396. A second effort in the next century by Pope Eugenius IV collapsed with the defeat of the Christian armies led by János Hunyadi and King Wladyslaw III Warneńezuk of Poland and Hungary at Varna in 1444.

The victorious march of the Ottoman forces into Europe was checked briefly, at the beginning of the 14th century, by internal problems within the empire and by Timur's threat in the east. Under the leadership of Mehmed I (1413–21), who won the struggle for succession, and his successor Murad II (1421–51), the march forward was resumed. The empire at this time also had to develop a navy to deal especially with Venice. Until the end of the 18th century this city-state was a major power in the eastern Mediterranean and waged with the Ottoman Empire a continual duel for control of the islands and the peninsulas of this sea—in particular, Crete, Cyprus, and the Morea.

The greatest single Ottoman military achievement occurred in the reign of Mehmed II the Conqueror (1451–81). Although Ottoman territories now reached far into Europe, the Byzantine Empire, reduced to the city of Constantinople and a small surrounding area, still stood. After long and careful preparation, Mehmed took the city in 1453, a date that compares in Greek history to that of Kosovo in the Serbian national memory. The fall of Constantinople marks also the completion of the process of the subjugation of the Balkan Peninsula to Ottoman rule. The empire now held most of Serbia, Bosnia, Bulgaria, and Greece. Although some isolated areas still resisted, no major centre of opposition remained. With the capture of the imperial city, the Ottoman conquerors made it their capital and developed an administrative system that was to dominate the Balkan region for the next five centuries.

Fall of Constantinople

The Ottoman Empire reached its height in power and prestige during the reign of Supleyman I, known in the West as "the Magnificent" and in Ottoman history as "the Lawgiver" (1520–66). A great military leader and an able administrator, Supleyman extended Ottoman territory far into the centre of Europe. Belgrade fell in 1521; in 1526 Supleyman defeated a Hungarian army at Mohaacs; and in 1529 the Ottoman army reached Vienna and lay siege to that imperial city. With the failure of this attempt, the limits were set to Ottoman westward expansion, but Supleyman's campaigns had left his country with control over most of Hungary and Transylvania. Transylvania, like the principalities of Walachia and Moldavia, taken in the 15th and 16th centuries, became an autonomous tributary province of the empire.

Although some territorial changes occurred in the Balkans during the next century and a half, the Ottoman Empire lost no significant territory to Christian Europe until after 1683. Moreover, it kept at least titular control of the majority of the Balkan lands until 1878. Thus, for almost five centuries the greater portion of the Balkan people lived under an alien Muslim rule. It is the common experience of the peninsula and one that has profoundly shaped its present condition.

Character of Ottoman rule. When the Balkan people first fell under Ottoman control, they became a part of one of the great empires of world history, a worthy successor to the Roman and Byzantine states, which, too, once had their centres at Constantinople. The negative opinion often held of Ottoman civilization is usually based on judgments made of conditions in the 18th and 19th centuries, when the state was in a period of obvious decline. In the 15th and 16th centuries, however, Ottoman institutions may have offered the Balkan Christian a better life than he had led previously under his native feudal governments.

To understand the political conditions in the Balkans in the Ottoman era, it is first necessary to emphasize that the conquering power was a Muslim and not a Turkish national state. The Ottoman leaders regarded their peoples as divided by religious faith rather than by nationality. Any individual could join the ruling group by converting to Islām. Basing the political structure on this concept, the non-Muslim peoples were divided into five religious communities called millets: Orthodox, Gregorian Armenian, Roman Catholic, Jewish, and Protestant. Each group was under the direction of its religious head. Thus for the Balkan people, the vast majority of whom were Orthodox, the titular leader was the patriarch of Constantinople. But

in practice, during the years of Ottoman rule, the Balkan Orthodox church organization became divided into its national components. Constantinople became a Greek centre; the Serbians had their own patriarchate at Peca, and the Bulgarians had a metroplitanate at Ohrid. The Romanians had similar national institutions. Thus national separateness and local tradition were preserved through the ecclesiastical organizations. The Ottoman government expected the church authorities to assume many civil functions, in particular judicial and tax-collecting duties. The Christian churches thus became, in a sense, part of the Ottoman state system.

The Ottoman government had a regularly organized administrative system with its centre in Constantinople and extending over the entire empire. The Balkan lands were part of the region (*beylerbeylik*) of Rumelia; the territories under direct administration were subdivided into provinces and then into lower administrative units. The principal interest of the Ottoman officials was the collection of taxes, which were to pay for the military power and the administration of the empire. Soldiers, policemen, and judges were also present in the administrative centres to maintain conditions beneficial to the proper collection of revenue.

Although most of the Balkan lands were administered by Ottoman officials, some areas enjoyed unusual rights of self-government. The Danubian principalities of Walachia and Moldavia, for instance, as well as Transylvania, being autonomous tributary regions, were governed by their own aristocracy. The merchant city of Dubrovnik was an autonomous republic. Certain other cities and regions either had won special rights or were too remote and primitive to arouse Ottoman concern. They enjoyed almost complete self-government. Even among the areas under direct administration, the Ottoman authorities remained chiefly concerned with taxation and the maintenance of public order. They did not attempt to regulate other details of Balkan Christian life. These fell under the jurisdiction of the local village authorities or the church.

The *devşirme* system

Although only Muslims could hold office in the empire or serve in its military forces, converted Balkan Christians came to occupy the highest positions at this time. In fact, these converts came to form the main basis of the military and administrative apparatus of the state. By the time of Süleyman, the highest state offices were held by slave administrators. These were either purchased, were prisoners of war, or were acquired through the *devşirme* system: from the 14th century to about the middle of the 17th, approximately every five years one out of four boys in the Balkans, aged between 10 and 20, was taken from his parents, brought to Constantinople, and converted to Islām. The most able of these were sent to the palace school, where they were trained as administrators, and then assigned posts throughout the empire. Other children recruited in this manner became members of the Janissary Corps—an elite infantry unit armed with muskets—which became the most effective fighting arm of the Ottoman army. Forbidden at first to marry, this body of new converts proved to be fanatic and dedicated soldiers.

In the first period of Ottoman rule, conditions for the Balkan peasant on the land were probably not overly onerous; in fact, he may have enjoyed a better position than his counterpart in western Europe. In theory, the sultan held all of the land taken in war; he was free to dispose of it at will. In practice, the Balkan lands were used to support cavalry units of the army. The members (spahis) were given grants of land (timars) in return for which they had to provide military service. At first, the amount that the peasant had to contribute in dues and services was carefully regulated. Later, however, the system tended to break down. With the introduction of the wide use of firearms the effectiveness of the cavalry was reduced. The government became more interested in increasing the tax load to pay for new weapons.

Despite the fact that the Ottoman government showed a high degree of toleration for non-Muslim faiths, there was no question of equality between religious groups. To rise in Ottoman society Balkan Christians had to abandon their faith. There were few attempts at forced conversion, but the subordinate position of the Christians was constantly emphasized. They could not wear conspicuous or rich clothes, for instance, or the colour green, sacred to Islām. If when on horseback a Christian passed a Muslim, he was forced to dismount. Old churches could be repaired, but new ones could not be built. They could not have bells or bell towers or be constructed in a place or manner likely to "offend the eyes of the faithful."

Even more important, Balkan Christians were forced to carry an unequal share of the tax burden of the empire. Although Christians were not subject to military conscription, they were assessed a special tax as a replacement. Christians were also liable to other services and payments, particularly in time of war, which were connected to their secondary status within the state. Despite these and other severe disabilities, it is interesting to note that there were relatively few examples of mass conversions among the Balkan people, with the exception of those that occurred in Bosnia, Albania, and Crete, where local conditions were unusual.

Despite the fact that the Ottoman system of government served to maintain the separation of the national groups, great similarities nevertheless existed in the conditions of life of all of the people who lived under direct Ottoman rule. The vast majority of the Balkan population lived as farmers or herdsmen, either on estates or in the less economically valuable hills and mountains on virtually independent family farms. They inhabited small primitive houses, clustered in villages, where their local and personal affairs were managed by the village elders and the church. Larger towns and cities served principally as trading and administrative centres. Ottoman officials usually resided in the largest cities, as did those Muslims who held large estates. Handicrafts and trades, with their centres in these cities, were organized in a guild system.

The average Balkan peasant in these circumstances lived a primitive existence, cut off from the rest of the world. Since education was controlled by the church and was extremely limited, illiteracy was nearly universal. The life of the individual was limited to his village. Nevertheless, in each region a unique peasant culture was retained. Each area and national group had its own style of houses, decorative patterns, and, most important, its traditional songs and stories that substituted for a written literature.

While the smaller Balkan communities reflected the local peasant culture, the Balkan cities, as the military and administrative centres for the Ottoman authorities, were built according to the architectural preferences of the ruling group. Typical of Ottoman architecture were the massive stone bridges, fortresses, mosques, baths, covered markets, and caravansaries.

Although most of the Balkan peoples tended to prefer to remain on the lands of their ancestors, they could, of course, as citizens of one state, move about within the empire. During the four centuries of Ottoman domination, there was not only much internal migration, due to war and similar causes, but certain nationalities tended to specialize in occupations that drew them out of their national area. For instance, Greek merchants could be found in most major Balkan cities; Vlach (Romanian) shepherds roamed the mountains of the entire peninsula. Some Turks moved from Anatolia to areas such as the Black Sea coast, Macedonia, Bosnia, and Thrace. Albanians served as guards or police in many localities. This intermixture of population caused complications later when it became necessary to draw the boundaries of the national states.

The decline of the Ottoman Empire

Despite the long existence of the empire, signs of weakness appeared in its structure as early as the reign of Süleyman. The Ottoman system could not function well without strong direction from the top. Up through Süleyman the state had been remarkably fortunate in its rulers, but among the 17 sultans who followed, few were men of ability. With the lack of firm direction from above, the administration based on slave officials who rose by merit inevitably changed. Born Muslims and sons of officials naturally sought to enjoy the rewards of public office. The *devşirme* system came to an end by the middle of the 17th century. Parallel with these developments was the increasing corruption of all aspects of Ottoman political life. High

offices were regularly sold; the officials who purchased their positions were primarily concerned with recovering their expenses and enriching themselves. The breakdown of the Ottoman administration, which commenced at the centre in Constantinople, spread out to encompass the entire governmental network.

Most dangerous for the future of the empire was the decline of the Janissary Corps. Though its members had gained the right to marry and to enlist their children in its ranks, they were soon forced to earn extra money and many became craftsmen on the side. Thus, this once-powerful military unit degenerated into a weak organization whose chief role in the state was to exert influence on the government. The corps, like the rest of the Ottoman military forces, failed to keep up with Western advances in military technology and organization. The military preeminence and the qualities of leadership that had won for the Ottoman Empire its vast domain were now lost.

The Christian, together with his Muslim neighbour, was directly affected by the decline of Ottoman administration. His chief grievances became the willful and arbitrary manner of the local officials and the grave abuses that were now associated with the collection of taxes. The central government relied for this service on tax farmers who competed yearly in bidding for the position. The tax farmer was chiefly interested in making a profit from his investment. Forceful methods and unfair standards of assessment were regularly used to extract high payments. The system was economically ruinous for the peasant and the central government alike.

The Balkan peasant found his standard of life also deeply affected by a change in the landholding system. As mentioned previously, under the timar system, strictly controlled nonhereditary land grants had been made in return for military service. With the relative decline in importance of the cavalry and the reduction in the number of military campaigns, large estates tended to pass into the hands of those who wished to work them at a profit and who now held them as hereditary rights. The peasant who worked the fields of the estates (now called *chifliks*) was reduced to the position of a sharecropper; his dues in kind and in labour were greatly increased.

Among the most severe consequences of the breakdown of the Ottoman government was the rise of lawlessness throughout the peninsula. As the central authority grew weaker, local Muslim leaders throughout the empire established what were in effect small principalities from which they were able to defy Constantinople and war among themselves. The most important of these for Balkan affairs were Ali Pasha of Janina and Pasvanoglu of Vidin. Christian bands of robbers (haiduks or klephts) also existed in large numbers. Their activities, together with the destruction caused by the Ottoman wars of the 18th century, made certain areas uninhabitable for long periods. These conditions were the direct cause of both the Serbian revolt of 1804 and the Greek revolution in the Morea in 1821.

Although all of the inhabitants of the empire suffered from this situation, certain Balkan Christians nevertheless did benefit from the growing weaknesses in the Ottoman system. Of the non-Muslim people, certainly the Greeks associated with either the Ottoman administration or the commercial world were in a privileged position. By the 18th century a number of Greek families residing in Constantinople, known as Phanariotes, had come to secure regularly certain high appointments in the Ottoman service. Most important were the offices of grand dragoman (secretary of state) and the governorships (*hospodar*) of Walachia and Moldavia. At the same time, through the patriarchate at Constantinople the Greek hierarchy was able to dominate the other national churches. In the middle of the 18th century, it secured control of the Bulgarian and Serbian ecclesiastical centres at Ohrid and Peca. In addition, Greek merchants were established in the cities of the peninsula, and they controlled a large part of the trade of the Black and the Mediterranean seas. These merchant communities were the most susceptible to the influence of Western political and social ideas.

The position of the Greeks in the empire

After the Greek, the Romanian landowner of the privileged provinces of Walachia and Moldavia held a superior position within the empire. These principalities were never under direct Ottoman administration, although they did pay tribute and they were subject to regulation from Constantinople. In both, a native aristocracy held control over an enserfed peasantry. After the Russian invasions of the early 18th century, during which Romanian leaders cooperated with the enemy, the Ottoman government appointed Greek rather than Romanian governors for the provinces. The Phanariote Greek regime, which lasted until 1821, was extremely corrupt and strongly disliked.

In comparison with these groups, the Serbs and Bulgarians—without a native aristocracy and largely peasants—were at a distinct disadvantage. Both nations suffered severely from the corruption and lawlessness of the time. Their lands were the scene of repeated struggles between local Muslim leaders and of battles with the invading great powers. In the 18th century, both found their national churches taken over by Greeks.

A special word must be said concerning the Albanians and the Montenegrins. As the only Balkan people among whom a majority (70 percent) converted to Islām, the Albanians were not subject to many of the problems of their Christian neighbours. Many rose high in the Ottoman state and in military service. Moreover, the central government made little effort to control closely this distant, backward, mountainous region. Relatively content with their status, the Albanians were the last to establish an independent state. Montenegro, too, although a Christian state under a prince-bishop, because of its poverty and the difficulty of access to its lands, was usually able to govern itself.

Increased role of European powers. Although the internal problems had become acute long before, in international affairs the empire was able to preserve its territories almost intact until the end of the 17th century. In the second half of that century, a last great offensive was begun and additional lands in the Ukraine were acquired. The symbolic turning point for the Ottoman offensive against Europe was the second siege of Vienna, in 1683. The failure to take the city marked a reversal in Ottoman fortunes. Henceforth, the European states moved against Ottoman possessions. Russia, Venice, Poland, and Austria now joined in a victorious coalition. In 1699 the empire was forced to sign the Treaty of Carlowitz, the first time that the Ottoman government appeared on the diplomatic stage as a clearly defeated power. By this treaty, Austria gained control of most of Hungary, Transylvania, Croatia, and Slavonia; Venice took Dalmatia and the Morea; Poland gained Podolia. Thus large numbers of Balkan people passed from Ottoman to Austrian and Venetian control. The agreement established what was to be until 1878 a relatively stable Austrian–Ottoman frontier along the Danube and Sava rivers. The main territorial gains at the expense of the Ottoman Empire in the next century were to be made by Russia.

Russian invasions

The Russian advances against Ottoman lands began under Peter I the Great at the end of the 17th century. The major acquisitions were made, however, by Catherine II the Great, who by the end of her reign had won the lands north of the Black Sea to the Dnestr River. For the future of the Balkan Peninsula the most important agreement between Russia and the Ottoman Empire was the Treaty of Küçük Kaynarca (1774). By its terms, Russia gained territory in the Black Sea region and certain trade rights, but, most important, because of the ambiguous wording of the treaty, Russia was later to claim that it had won certain rights to speak in behalf of the Orthodox Christians of the empire. In 1781 Catherine joined with Joseph II of Austria in what was to be the first of numerous partition agreements among the great powers. According to this agreement, both Austria and Russia were to receive territory, but the majority of the lands of the Ottoman Empire in Europe were to be organized into a Kingdom of Dacia, under Russian influence, and a Greek kingdom with its capital at Constantinople and with Catherine's grandson as its ruler. General European events prevented the success of this venture, but, by the end of the century, it was apparent that Russia was both the chief threat to the empire and the major hope of the Balkan people for outside aid.

French Revolution and Napoleonic era. The wars of this period were to have a deep effect on the Balkans, both from the international aspect and because of the response that French Revolutionary ideology evoked within the peninsula. This period also witnessed the entrance of France and Britain actively into the diplomatic and military struggle over the control of the Ottoman Empire. That state, in the midst of another grave internal crisis, now found its lands under attack from France. In 1797 France took the Ionian Islands from Venice. A year later, Napoleon launched an invasion of Egypt and Syria. As a consequence of this action, the empire was in alliance with Britain and Russia from 1798 to 1802.

After a period of peace from 1802 to 1806, the Ottoman government shifted sides and joined France against Russia. This alliance proved disastrous, because in the next year, by the Treaty of Tilsit, France came to terms with Russia. It is interesting to note that at this time these two governments discussed a partition of the Ottoman Empire, but they were unable to agree on the fate of Constantinople. In 1812 Russia, faced with an impending French invasion, signed the Treaty of Bucharest with the Ottoman Empire, which ceded the Romanian territory of Bessarabia to Russia.

The Balkan peoples were affected not only by the wars that were waged in the area, but also by the political arrangements made by the several powers. Most influential for the future was the introduction into the Balkans of French Revolutionary institutions and ideology through the Napoleonic conquests. For example, in the Ionian Islands, which France held from 1797 to 1799 and again from 1807 to 1814, a constitutional government on Western patterns was established. The Septinsular Republic was the first Greek national government in modern times. French political experiments also involved South Slav lands. In 1809 Austria was forced to cede to France its Balkan territories. From Dalmatia, Slovenia, Istria, Trieste, and parts of Croatia a new political entity, the Illyrian Provinces, was formed and became a part of the French Empire. As in the Ionian Islands, the government was based on revolutionary principles. The Illyrian Provinces has been described as the first Yugoslav state, as it included within its boundaries Serbs, Croats, and Slovenes.

This period also brought about important changes in Moldavia and Walachia and in Serbia. The Ottoman Empire was forced to agree to an increase of Russian direct influence in the Danubian principalities; henceforth, no governor of either province could be dismissed without Russian consent. Russia thus acquired what was, in effect, a protectorate over the country. This power also played a major role in the events of the first Serbian revolution (see below).

The defeat of Napoleon in 1815 resulted in the restoration of the political and territorial status quo in the Balkans, with only a few exceptions—Russia, of course, retained Bessarabia; the Ionian Islands were placed under British protection; the Illyrian Provinces disappeared as a political entity; and Austria received back its former land and, in addition, acquired Dalmatia. The Ottoman Empire, which did not attend the Congress of Vienna in the aftermath of the Napoleonic Wars, retained most of its territories intact. It was also to be aided by the conservative reaction that followed in a Europe tired of war and revolutionary upheaval.

In Balkan history the French Revolution and the wars of Napoleon mark the shift from the long period of Ottoman domination into the era of the national revolutions of the 19th century. To some extent, the stage had already been set in the 18th century, when both Russia and Austria appealed to the Balkan subject populations for assistance against the Ottoman Empire. These wars, as well as those of Napoleon, had shown how weak the Ottoman military forces really were. Moreover, Balkan nationals had fought in these wars. They had learned modern military methods and they were armed. Equally important, the period of war and internal upset had opened the area to outside influences. The national and liberal ideology of Revolutionary France provided a program that would allow Balkan leaders to combat not only Ottoman political control but also the stifling cultural influence of their Christian church hierarchies.

Although the Balkans were now ready for revolt, the international situation was to change to the detriment of such actions. After the period of war, which had wasted the resources and energy of Europe before 1815, the powers desired, above all, peace with political and social stability. There occurred in all the European states a conservative reaction directed against revolutionary methods and liberal–national programs. Subject Christian peoples could thus not expect aid or sympathy from abroad. At the same time, it was also apparent to the European powers that the Ottoman Empire was in a dangerous condition of internal decay and military weakness. The question of the fate of the Ottoman territories and of the control of that government became perhaps the most important single diplomatic problem for Europe in the century after the Congress of Vienna. This issue, the so-called Eastern Question, was the direct cause of the two great wars of that period—the Crimean War and World War I—and the occasion of repeated less serious conflicts among the powers.

The basic problem in the disposition of the Ottoman lands was their strategic position across three continents. Because of its past history and its close links with the Balkan people through the Orthodox Church, Russia was in the best position to gain predominant influence in the area. Russia's chief rivals were Austria and Great Britain. The Habsburg monarchy could not afford to allow Russia more political control in lands along its frontiers. The British feared for their communications with their empire, which ran through the eastern Mediterranean, and their control over India. Moreover, despite its favourable relations with the Balkan people, the Russian government, too, because of its abhorrence of revolutionary activity and liberal reform programs, stood after 1815 against national revolt. Thus, although the era of the French wars had prepared the Balkan people for revolution, the international situation was not favourable. Nevertheless, the next 65 years were to witness the establishment of independent, or autonomous, governments for almost all of the Balkan national groups.

BALKANS FROM 1815 TO 1914

The era of national liberation

The 19th century saw the political and social development of the Balkans and the formation of independent states. During this period the Balkans were drawn, economically and culturally, into the orbit of contemporary Europe. While it is somewhat arbitrary to divide this exceedingly complex and dynamic era into segments, it is possible to delineate four general periods.

The first period brought the beginning of the process of national liberation during the first Balkan revolutions, 1800–30. At the beginning of the 19th century the Habsburgs ruled the northwestern part of the Balkans, while the Ottomans dominated the main central and southern regions of the peninsula. The development of a national renaissance of the Balkan peoples varied according to the different political and social conditions prevailing in these areas: political development occurred under the Habsburgs, and revolutionary upheavals predominated under the Ottomans.

Three main factors influenced the national revolutions in the areas under the Ottomans: the decline of the Ottoman Empire, provoked by a general crisis of Ottoman feudalism; the gradual shaping of a new Balkan society, as a result of the economic traffic with Europe and the development of local autonomy; and the influence of the "outer" Balkan world on the "inner" (*e.g.*, the influence of the Greek, Bulgarian, and Albanian trading colonies in the Mediterranean and Black Sea areas, of the Serbs in southern Hungary, etc.). The Balkan revolutions had two general aspects: nationalistic and agrarian. The national aspect was expressed in the drive for national liberation, the creation of national economies and cultures, and the political organization of national states. The agrarian aspect was marked by the endeavours of the peasantry to get rid of the Ottoman landlords and take possession of the land.

Next came a period (1830–78) of political and social

development. A third period saw the inclusion of the Balkans in Europe during the age of imperialism, 1878–1903. Russian designs on the Balkans culminated in the Treaty of San Stefano in 1878, which created a greater Bulgaria extending from the Danube to the Aegean and from the Black Sea to Albania, but at the Congress of Berlin in the same year the other European powers, alarmed at this extension of Russian influence, redrew the map. Bulgaria was limited to the country between the Danube and the Balkan Mountains; Serbia received Nish, Pirot and Vranje; the Dobruja was ceded to Romania; and Austria-Hungary was accorded the right to occupy Bosnia-Hercegovina provisionally. Serbia and Romania at the same time became independent principalities, while the independent status of Montenegro was recognized. In 1881 Greece secured Thessaly and Aarta from Turkey. The Bulgarian occupation of Eastern Rumelia (recognized in 1878 as an autonomous Turkish province) in 1885 was followed by the Serbo-Bulgarian War. The Greco-Turkish War of 1897 led to a rectification of the Thessalian frontier in Turkey's favour. For further history of this era, see separate Balkan country articles.

Imperialism, nationalism, and Balkan unity

Before World War I (1903–14). The fourth period involved the Balkans in European crisis on the eve of World War I. At the beginning of the century the Balkans were increasingly the scene of international conflict. The European powers had clashing military, political, and economic interests in the peninsula. The Baghdad railway project at the turn of the century symbolized German ambitions to push eastward, challenging French financial dominance in the Ottoman Empire. The growing weakness of Russia upset the Austro-Russian balance in the Balkans, while growing Italian strength in the Adriatic stimulated efforts by Austria-Hungary to reach the port of Thessaloníki and the Aegean.

Among the Balkan countries themselves, developing national strength resulted in a general movement toward political and economic emancipation. In the Yugoslav lands the year 1903 marked a turning point. Croatian political forces began to organize against Hungarian repression, and a similar restlessness was apparent in Bosnia. The assassination of King Alexander Obrenovica in Serbia prepared the way for a dynamic, nationalistic foreign policy (under the rule of Peter I Karageorgević). In Macedonia in 1903, the Ilinden uprising displayed to Europe the weakness of Turkey, the "sick man on the Bosphorus." In Croatia and Dalmatia, a coalition of political parties put forward a program calling for Yugoslav unity and social and political reform. The movement received support from Serbia and Montenegro and began to gather impetus. The Balkan states started to arm, buying artillery and ammunition in Europe.

Two tendencies were at work in the Balkans: rivalries among the states for territorial aggrandizement, and a common hostility toward interference from outside powers. A growing inter-Balkan struggle next occurred over Macedonia and was expressed in the activities of armed bands from neighbouring states that brought much suffering to the local populations. The fear of European intervention and the needs of common defense imposed a political rapprochement on Serbia and Bulgaria in 1904, joined later by Montenegro. A peasant revolt in Romania in 1907, put down by the military at the cost of thousands of lives, demonstrated that Romania's agrarian problem was far from solved.

Crisis of 1908–09. A diplomatic struggle among the European powers began in 1908 over railway projects in the Balkans. These projects expressed the political tendencies of the states involved: Austria's push toward the Aegean (the Novi Pazar railway project), Russia's and Serbia's toward the Adriatic (the Danube–Adriatic project), Italy's penetration of southern Albania (the Vlóre–Munushtir railway), and the effort of Greece and Bulgaria to absorb central Macedonia. In July 1908 the Ottoman garrison in Thessaloníki rebelled, as a result of the revolutionary activity of the Young Turks. The officers (among them Mustafa Kemal, the future leader of the Turkish republic), forced the Sultan to proclaim a constitutional era. The Young Turks wanted a regime that would give liberty and equality to all the nationalities within the empire. On October 5 the Bulgarians took advantage of the confusion to proclaim their full independence. On October 6 Austria-Hungary announced its annexation of Bosnia-Hercegovina. Two days later the Cretans proclaimed their union with Greece. The system created by the Congress of Berlin in 1878 had collapsed. The major crisis was over Bosnia-Hercegovina: the Serbs protested vehemently against the annexation; England opposed Austria; Russia backed Serbia and Bulgaria. The Austro-Turkish conflict was settled by an indemnity paid by Austria to the Turks, but the Austro-Serbian conflict brought Europe to the edge of war. Pressure from Berlin in March 1909 forced Russia to yield. Serbia had to follow the Russian example and accept Bosnia's annexation by Austria-Hungary. The crisis of 1908–09 foreshadowed coming events.

Revolt of the Young Turks

Balkan Wars of 1912–13. The situation in the Balkans remained uneasy. The Albanians, after a short-lived collaboration with the Young Turks, revolted against them in 1909–12. The Cretan *énosis* with Greece failed because of European opposition. This caused general indignation in Greece, opening the way to a military coup and the premiership of the Cretan political leader Eleuthérios Venizélos. The Venizélos government introduced reforms, reorganized the army, and revised the constitution. Macedonia continued to be a magnet for the nationalist ambitions of the Serbs, Bulgars, and Greeks. The Young Turks banned political parties and national organizations in Macedonia in 1909, and, consequently, the armed bands of various nationalities reopened their guerrilla warfare. An opportunity for an effective cooperation against Turkey presented itself in September 1911, when war broke out between Turkey and Italy in North Africa.

The Balkan alliance against the Ottomans

War against the Ottomans required a Balkan alliance. Serbo-Bulgarian negotiations started in the fall of 1911 but were troubled by differences in regard to the future delimitation of Macedonia. Finally an agreement was reached under Russian auspices in March 1912, providing for the division of Macedonia. A Greek–Bulgarian agreement was reached in May 1912, without touching the delimitation problem. Montenegro then joined the alliance, which disposed of some 550,000 troops. The war began in October 1912. The Balkan allies were soon victorious: the Serbs defeated the Ottomans at Kumanovo, joined forces with the Montenegrins to enter Skopje, achieved another victory at Munushtir, and reached the Adriatic at Durrës. The Bulgarians defeated the main Ottoman forces at Kirklareli and Lüleburgaz, advancing to the Çatalca lines in front of Constantinople. The Greeks seized Thessaloníki and laid siege to Ioánnina. The Ottomans lost all their territories in Europe, except a small strip around Constantinople. On December 3 an armistice was concluded. Peace negotiations opened in London on December 16.

The victories of the Balkan allies affected many Austro-Hungarian interests. Vienna reacted vehemently, demanding the withdrawal of Serbian troops from the Albanian coast where an Albanian state had been proclaimed on November 28, 1912. The Austro-Serbian conflict automatically developed into an Austro-Russian one.

The crisis was settled at a conference of ambassadors in London in December 1912 that recognized the new Albanian state and obliged Serbia to withdraw its troops from the Adriatic. But at the end of January 1913, after a coup d'etat in Constantinople by the nationalistic Young Turks, the war with the Ottomans was resumed. The allies were again victorious: Ioánnina fell to the Greeks, and Adrianople to the Bulgarians. Another crisis arose when the Montenegrins refused to leave Shkodër, which the London ambassadors' conference had given to Albania. The Montenegrins were forced to yield under the threat of a European naval blockade of their coast. A peace treaty, signed in London on May 30, 1913, gave all the territory west of the Enez–Midye line to the Balkan allies. Crete was united with Greece.

Serbia and Greece against Bulgaria

The territorial settlement produced discords among the Balkan allies. Serbia refused to give up the parts of Macedonia assigned by the 1912 treaty to Bulgaria on the ground that it had already been forced to withdraw from the Adriatic. A bitter struggle then ensued between the

Greeks and the Bulgarians over Thessaloníki and Thrace. Romania, as a price for remaining neutral, demanded from the Bulgarians a part of the Dobruja. Both Serbia and Bulgaria were reluctant to accept Russian arbitration. Austro-Hungarian diplomacy sought to undermine the Balkan alliance. Serbia and Greece allied themselves against Bulgaria on June 1, 1913. King Ferdinand of Bulgaria, backed by army circles, ordered his troops to attack Serbia and Greece in Macedonia on June 30. But the Bulgarian armies were defeated by the Serbs and Greeks. At the same time the Romanians entered the Dobruja and the Ottomans recaptured Adrianople. An armistice was concluded on July 31 and a peace treaty signed in Bucharest on August 10, 1913.

Results of wars. As a result of the Balkan Wars, Greece obtained Thessaloníki, Kaválla, and a large coastal part of Macedonia; Serbia gained the northern and central part of Macedonia; Montenegro acquired a portion of the sanjak of Novi Pazar, establishing a common frontier with Serbia; Bulgaria retained a part of eastern Macedonia; and Romania procured its part of Dobruja. The long decline of Ottoman rule in the Balkans had ended.

The political consequences of the Balkan Wars were considerable. Apart from Turkey, the real loser was Austria-Hungary. The partitioning of the sanjak of Novi Pazar between Serbia and Montenegro made it impossible, in the subsequent crisis of June–July 1914, for Austria-Hungary to intervene in the Balkans by occupying the sanjak. The success of Serbia and Montenegro stimulated the Yugoslav movement for union in the Habsburg Empire. The wars similarly altered the structure of alliances in the Balkans. Dissatisfied, Bulgaria henceforth looked to Austria-Hungary for support, while Romania tended to move away from its allies and toward Russia.

The most alarming result was the growth of tension between Austria-Hungary and Serbia. Serbia had extensive claims upon Albanian territory. Having obtained an assurance of German support, Austria-Hungary delivered an ultimatum in October 1913 to compel Serbia to withdraw from the Albanian borderlands. This, however, did not solved the Southern Slav question for Austria-Hungary, and it emerged once again in an acute form with the assassination of the archduke Francis Ferdinand on June 28, 1914, in Sarajevo, Bosnia. This event was followed by Austria-Hungary's ultimatum and by its declaration of war on Serbia (July 28, 1914) and the outbreak of general war in Europe in August.

BALKANS AFTER 1914

World War I and peace settlements, 1914–23. *Role of Balkan states in war.* The World War of 1914–18 was triggered by Balkan revolutionary nationalism. Gavrilo Princip, who shot and killed Archduke Francis Ferdinand, heir apparent to the Habsburg throne, was one of many young Serbs who pinned so much faith upon the advantages of national unification as to risk their lives and flout existing political authorities, both in Serbia itself and in the adjacent lands of Bosnia, Hercegovina, Dalmatia, and southern Hungary, where Serbian populations lived under Habsburg rule.

Nationalist movements

The breakup of traditional peasant styles of life among the South Slav peoples—Slovenians, Croats, Serbs, and Macedonians—fuelled this revolutionary movement. Population growth made subdivision of peasant holdings necessary; but, when a father had to divide his land among several sons, the new families could not hope to live as their parents had. This pushed innumerable young men off the land and into revolutionary activities. The pattern was as follows: as life on the farm became impossible, ambitious persons tried to get enough formal education to qualify for a desk job in town. But even after several years of secondary schooling, desk jobs were hard to come by, and many young men who had gone that far by desperate effort were not willing to wait patiently until something turned up. Instead, they listened eagerly to those who preached extreme revolutionary action.

According to Serbian nationalist agitators, justice, freedom, and a decent regard for Serbian dignity required that all speakers of the South Slav tongue who were also of the Orthodox faith should belong to the same state. The fact that the end of Habsburg rule over Serbian populations would inevitably mean more government jobs for educated Serbs was an attractive additional advantage.

Some thought that Muslim and even Roman Catholic speakers of the South Slav language should join their Orthodox brethren in a new South Slav state. But this Yugoslav (*Yug,* "South") ideal had little appeal for most Serbs. It attracted support mainly in Dalmatia, where rural Serbs (Orthodox South Slavs) and Croats (Roman Catholic South Slavs) found it easy to cooperate against the Italians, who had long dominated town life along the Adriatic.

Despite Serbia's tiny size compared to the vastness of the Habsburg monarchy, many high Austrian officials concluded that intransigent Serbian nationalism constituted a serious threat to their state. The power of the nationalist ideal in the Balkans had been demonstrated in 1912 by the First Balkan War, which all but drove the Ottoman Empire from Europe and added substantial new territories to Serbia. This success fanned the Serbian nationalist ambition which was already at white heat, to complete the task of liberation by disrupting the Habsburg state.

What made such a program so frightening to the Austrians was that other nationalities within the Habsburg Empire shared in some degree the ambition to win greater control over their own affairs; a few even dreamed of achieving complete national independence. But in 1914 nationalist movements among the Czechs, Germans, Italians, Poles, Magyars, Slovenes, and Croats were less emotionally intense and, therefore, less immediately threatening to constituted authority than was the case among the Serbs.

Outbreak of war

Austrian officials, therefore, felt that a showdown was desirable. They decided to seize upon Archduke Francis Ferdinand's assassination to settle accounts with Serbia. Preliminary investigation failed to turn up definite evidence that the Serbian government had been connected with the assassination; but historians have since shown that the Serbian premier, Nikola Pašić knew what was going on and tried indirectly to warn Austrian authorities of the plot. Col. Dragutin Dimitrijević, chief of intelligence for the Serbian general staff and leader of the Black Hand—a secret society that planned the assassination and armed Princip (with several others)—was one of Pašić's political enemies and rivals. In hinting to the Austrians of what was afoot, Pašić did as much as he dared to thwart Dimitrijević's risky, revolutionary plans. But the Austrian authorities failed to take the hint. The Archduke was assassinated, official Europe was horrified, and on July 23, 1914, the Habsburg government delivered an ultimatum to the Serbs requiring suppression of patriotic societies and the establishment of a joint commission to investigate and punish those persons who were responsible for organizing the assassination.

The ultimatum was designed to be unacceptable. Despite a conciliatory reply, the Austrians declared war against Serbia on July 28, exactly a month after the assassination itself. Within a week Europe's alliance system swung creakily into action, pitting the Triple Entente—Russia, France, and Great Britain—on Serbia's side against the Central Powers—Germany and Austria-Hungary.

Russia's entry into the war upset Austrian plans for crushing Serbia. Austrian troops had to be diverted to the Russian front, and, when the Austrians were finally ready to attack the Serbs, on August 13, they were repulsed. A Serbian counteroffensive soon petered out, however, and by the end of 1914 the battle line stood very close to the prewar frontiers.

The outbreak of hostilities provoked intense diplomatic–military activity elsewhere in the Balkans. On August 11, 1914, two German cruisers that had been trapped in the Mediterranean Sea at the beginning of the war arrived in Constantinople. A fictitious sale transferred them to the Turks, and in October these vessels sailed into the Black Sea to bombard Russian coastal towns. The Allies (Entente) promptly declared war against Turkey (November 4–5, 1914).

Both the Entente and the Central Powers sought to rally support for their cause among the three remaining un-

committed Balkan states—Bulgaria, Romania, and Greece (Albania was in chaos and lacked a central government). Bulgaria wanted the parts of Thrace and Macedonia that had been annexed by the Serbs and Greeks after the Second Balkan War in 1913; Romania wanted the Habsburg territories of Transylvania and adjacent regions in Bukovina and the Banat; Greece had ambitions in Anatolia and, above all, desired to possess Constantinople (later Istanbul).

The Central Powers were able to promise the Bulgars most of what they desired, because Serbia would be the main loser. Similarly, the Entente was able to appease the Romanian appetite at the expense of the Habsburgs. Greece's ambitions were more difficult, however, because Russia, too, aspired to possess Constantinople, and so did the Bulgars.

Courtship of the Balkan states

The work of diplomats and intelligence agents in lining up allies in the Balkans for one side and the other came slowly to fruition in 1915–17. The Bulgars were the first to commit themselves, by allying with the Central Powers in September 1915. By then the strategic situation in the Balkans had altered greatly. The first important move was British. In February 1915 the Royal Navy bombarded Turkish forts in the Dardanelles, and a month later British warships tried unsuccessfully to force their way through the straits. Then, on April 25, 1915, British troops went ashore on the Aegean side of the Gallipoli Peninsula in order to take the Turkish defenses of the Dardanelles in the rear and open a way to Constantinople for the Royal Navy; but once again the Turks were able to stop the British advance before it could achieve strategic success.

The next major move came from Italy. Despite a defensive alliance with Germany and Austria-Hungary, the Italian government remained neutral in 1914, arguing that, because Austria had attacked Serbia, the terms of the alliance had not been fulfilled. Italian nationalists wished to add Dalmatia and other Austrian provinces to their country; despite Serbian objections, the Allies agreed to most of the Italian demands in the secret Treaty of London (signed in April 1915). Accordingly, in May the Italians declared war and opened a new front against Austria in the Alps. Italian troops also crossed the Adriatic to occupy Albania.

The intervention of Italy, plus Austrian commitments on the Russian front, forced the Austrians to postpone large-scale action against the Serbs until fall; but by the first week in October everything was ready for a major assault. The Bulgars saw in this their opportunity to take Macedonia from Serbia, and they prepared to join the attack.

In a last-minute effort to aid the Serbs, French and British troops landed at the Greek city of Thessaloníki on October 3. This touched off a major crisis inside Greece, pitting Prime Minister Eleuthérios Venizélos, champion of the Allies, against King Constantine, who favoured continued neutrality, at least until it was clearer which side would win the war. (In the end, by bombarding the royal palace in Athens the Allied powers were able to compel Constantine to abdicate in June 1917; and Venizélos, who took command of Greek affairs, promptly caused Greece to declare war against the Central Powers.)

A combined Austrian and German force attacked Serbia on October 6, 1915, with overwhelming strength. Allied troops in Salonika were too few to check the simultaneous Bulgar advance from the east. In November, the Serbian Army began a painful retreat through the mountain passes of Albania. A remnant 125,000 strong found refuge on the Greek island of Corfu, where the Serbian government, still being led by the aged and ailing King Peter I, with Nikola Pašić as prime minister, established a temporary headquarters (January 1916) in what had formerly been the German kaiser's holiday palace.

Meanwhile, the British withdrew their troops from Gallipoli, transferring most of them to the Thessaloníki front, which soon extended westward to link up with the Italians in Albania. Trench warfare then set in along a battle line extending all the way from the Adriatic coast to Kaválla on the Aegean. Neither side could break through the other's prepared defenses.

Stalemate of 1916

This stalemate was demonstrated in 1916, when the last Balkan neutral, Romania, entered the war on the Allied side on August 27. This move was planned to coincide both with a Russian offensive against Austria and with a major push against Bulgaria along the Thessaloníki front. Despite substantial reinforcement by rested and re-equipped Serbian troops transferred to Thessaloníki from Corfu, the offensive failed. The Romanian Army proved ineffective, and by January 1917 most of Romania was in the hands of the Central Powers.

The year 1917 saw no large-scale military action in the Balkans, but revolutions in Russia changed the realities of power profoundly. The Bolsheviks publicly renounced Russian claims on Constantinople; instead, Lenin appealed over the heads of all governments to the peoples of Europe to rise against their exploiters and inaugurate Socialism. Balkan response to Lenin's revolutionary summons was slight, partly because the Western Allies forestalled the Bolshevik appeal by endorsing a different revolutionary formula—national self-determination. The war thus became far more ideological than before, pitting nationalist revolutionary ideals against Socialist revolutionary ideals through the whole of eastern Europe.

The most complex nationality issue in the Balkans centred around Serbia's relationship to the other South Slav peoples of the Habsburg Empire. On July 20, 1917, the Serb prime minister Pašić signed a document declaring that Serbs, Croats, and Slovenes should form a single state after the war, under the Serbian Karageorgević dynasty but with appropriate local autonomies. Exiled Dalmatian politicians, claiming to represent the Croats and Slovenes, also signed this Pact of Corfu. In April 1918 a Congress of Oppressed Nationalities met in Rome and, with Italian blessing, reiterated the idea that Serbs, Croats, and Slovenes belonged together, despite the fact that the three nationalities distrusted one another profoundly and that most Croats remained loyal to the Habsburg cause.

End of the war

The long military stalemate in the Balkans ended on September 30, 1918, when the Bulgars sued for an armistice. From the Thessaloníki front, Allied troops under French command marched northward to the Danube. As the Allied forces approached, the Romanians, who had signed a peace with the Central Powers in May 1918, reentered the war on November 8, just in time to count as an Allied and victorious power at the peace conference.

The Ottoman government followed the example of the Bulgarians by suing for an armistice on October 30, 1918. The armistice terms allowed a British force to advance through the Dardanelles and to occupy Constantinople on November 13. Amid the crash of falling empires, the Habsburg monarch also signed an armistice on November 3; but events had stripped the Emperor of his power to influence affairs. The Romanians (already in possession of the former tsarist province of Bessarabia) hastened to take possession of as much of Transylvania, Banat, and Bukovina as possible; meanwhile, Serbian troops were moving into the former Habsburg territory from the south, while the victorious Italians also sought to get control of as much of the Adriatic coastlands as they could.

Under these circumstances the Slovenes and Croats had little room for manoeuvre. On October 29, 1918, a national council meeting in Zagreb proclaimed the independence of "Yugoslavia" (meaning the former Habsburg lands in which the South Slavs lived). Before a stable settlement with the Serbs could be achieved—a preliminary agreement reached at Geneva on November 9 was later repudiated—the Zagreb National Council broke apart.

Kingdom of Serbs, Croats, and Slovenes

A faction opted for immediate union with Serbia; and, accordingly, on December 1, 1918, the Serbian monarch formally proclaimed a new Kingdom of Serbs, Croats, and Slovenes from his capital, Belgrade. A political coup unseated the Prince of Montenegro, whose state was also merged with the new South Slav state. But whether the new state would be federal or unitary and the question of how Serbs and Croats would come to an understanding with one another remained completely unsettled.

Results of peace conferences. The Paris Peace Conference that began in January 1919 drew up separate treaties for each defeated enemy state. By the Treaty of St. Germain (signed September 10, 1919) Austria ceded Sloveni-

an and Dalmatian territory to the new Kingdom of Serbs, Croats, and Slovenes; by the Treaty of Neuilly (signed November 27, 1919) Bulgaria lost a strip of the Aegean coast (acquired in 1912) to Greece and surrendered small border territories to Serbia; by the Treaty of Trianon (signed June 4, 1920) Hungary transferred Transylvania and part of the Banat to Romania and surrendered Croatia, Slavonia, and the rest of the Banat to the new South Slav state; lastly, by the Treaty of Sèvres (signed August 10, 1920), Turkey assigned most of Thrace as well as the hinterland of Smyrna (Izmir) in Asia Minor, to Greece.

From L. Stavrianos, *The Balkans Since 1453*; Holt, Rinehart and Winston, Inc.

The Balkans after World War I.

These treaties settled many of the territorial questions that had long distracted Balkan politics and in a comparatively enduring way. But "national self-determination" proved a difficult formula to apply in lands where mixtures of nationalities were the rule rather than the exception. Some issues, such as the delineation of the Italian–Yugoslav boundary, were never resolved at the peace conference; instead, bilateral negotiation eventually (1924) defined a frontier that satisfied neither side.

Albania's borders were not fixed until 1926, when an international commission, established in 1912, finally concluded its labours. This settlement, which left large Albanian populations inside the new South Slav state, made little difference politically until after World War II, because Albanian national self-consciousness was weakened by traditional loyalties to rival kindred groupings, on the one hand, and by religious (Muslim, Orthodox, Roman Catholic) and linguistic (Gheg, Tosk) differences on the other.

Turkish and Tartar minorities along the Black Sea coast in Bulgaria and Romania shared the Albanians' prepolitical status; and Jews, important mainly in Romania and at Thessaloníki, offered no systematic resistance to the new masters of the Balkans. The case was far different, however, with other Balkan nationalities. Those who had formerly enjoyed a leading position in society and government found it all but impossible to accept willingly the loss of former privileges. The victors often retaliated by subjecting German, Magyar, Turkish, and (in Dalmatia) Italian minorities to flagrant administrative discrimination.

Hence, the territorial settlement of 1919–20 had the effect of transferring to Romania and the new Kingdom of Serbs, Croats, and Slovenes many of the nationality problems that had plagued the Habsburg and Ottoman empires before 1914. In Romania, Magyar, German, Jewish, Ukrainian, Bulgarian, and Turko-Tartar minorities amounted to at least 4,500,000 in a total population of about 18,000,000; but the Romanian majority did give a solid core to the new state. The situation in the Kingdom of Serbs, Croats, and Slovenes was less stable, for, although the Serbs constituted the largest single nationality, they were still a minority in the state considered as a whole.

Continuing nationality problems

German, Magyar, Albanian, and Romanian minorities totalled over 1,600,000; about 1,300,000 Muslims (mostly speaking Serbo-Croatian) constituted another distinct bloc; Slovenes (about 1,100,000) formed a self-conscious, distinct nationality, too; but the really critical matter was the relation between the Croats (about 3,500,000) and the Serbs (about 5,500,000). A new constitution went into effect on January 1, 1921, establishing a unitary state on democratic lines. Croats felt that the Corfu Declaration of 1917 had committed the Serbs to federalism, and they refused to accept the new arrangement. This was especially dangerous for the new state, because the Italian government was dissatisfied with the peace settlement in the Adriatic and set out actively to encourage disruptive forces. The death of King Peter in 1921 made small difference; his heir, King Alexander I (reigned 1921–34), had already exercised the royal powers for several years.

The southern Balkans were even more distracted in the first postwar years. A Turkish nationalist movement, headed by Kemal Atatürk, refused to accept the terms of the Treaty of Sèvres. The Greeks, seeking to make good their claim to even more extensive territories in Asia Minor, invaded the interior in hope of forcing the Turks to yield; in 1921 they met defeat, and the angry Turks forced all Greeks and other Christian inhabitants from the land. About 1,500,000 survivors fled across the Aegean in 1921–22.

A new peace, agreed to at Lausanne, Switzerland, in 1923, provided for the systematic exchange of populations between Greece and Turkey under League of Nations supervision. Greek and other Christian inhabitants of Constantinople were exempted from this exchange; in return, Greece promised to allow Turkish peasants in western Thrace to remain on their land. This treaty provided that the Greeks relinquish their claim to Asia Minor entirely and retroceded eastern Thrace to Turkey.

Exchange of population with Bulgaria was also arranged by a separate agreement. The result, by about 1927, when major transfers ceased, was the sorting out of the populations of Greece, Bulgaria, and the western parts of Turkey by nationality.

No politically important national minorities remained in the southern and eastern Balkans. This radical surgery allowed nationality frictions to subside slowly in later decades. In the northern Balkans, however, such frictions continued to constitute a major axis of politics during the interwar years.

Interwar developments, 1923–39. The upheavals of World War I did little to solve the underlying problems of Balkan society. It is not strange, therefore, that revolutionary discontent found new channels of expression after the war. Success for certain nationalities (Serbs, Romanians) automatically meant frustration for others (Bulgars, Magyars, Croats, Macedonians), so that in some regions of the peninsula old-fashioned nationalist conspiracy and agitation continued as before, but now directed against the new masters of the land.

But national self-determination lost much of its glamour as it became clear that old problems were not really relieved by the changes in political boundaries and shifts in dominating nationalities that had occurred in 1918–19. Two new revolutionary movements, therefore, surged to the fore: peasantism and Communism.

New revolutionary movements. The three main bearers

Peasantism

of the peasantist idea were the Peasant Party of Bulgaria, led by Aleksandŭr Stamboliyski; the Croatian Peasant Party, led by Stjepan Radić; and the Romanian Peasant Party, led by Iuliu Maniu. The latter, based mainly in Transylvania, opted for parliamentary and peaceable agitation and, even when Maniu briefly became prime minister (served 1928-30), accomplished little to reform rural conditions. The Croatian Peasant Party quickly became identified with Croat nationalism, thanks to Radić's unbending opposition to Serbian preponderance in the new Kingdom of Serbs, Croats, and Slovenes; by boycotting the parliament, Radić showed how shaky the new state really was, but his policy, too, accomplished no positive ends. Stamboliyski, on the other hand, came to power in 1919 on a wave of revolutionary feeling; but his efforts to end bureaucratic oppression and overthrow the parasitic classes that fattened on peasant labour only succeeded in putting crude former peasants and party men into administrative roles they were poorly equipped to handle; in 1923 a coup d'etat led to Stamboliyski's assassination and to the establishment of a shaky parliamentary regime, closely controlled from behind the scenes by royal, military, and semi-military manipulators.

The basic reason for the failure of peasant parties to achieve their ends lay in their programs. Men who wished to destroy the so-called social parasites (*i.e.,* everyone who did not work with his hands and raise his own food) could not take power without themselves becoming that which they wished to abolish: bureaucrats and paper shufflers.

Opposition to those who ruled was the only role such parties and movements could accept comfortably. Only in this way, in fact, could they hope to remain faithful to the perennial distrust their peasant constituencies felt toward government in any form and toward city people generally.

Rise of Communist parties

The other new revolutionary movement, Communism, was better equipped ideologically and organizationally. Communist parties had arisen in each Balkan state by 1921. In Bulgaria and Yugoslavia, the new parties met with rapid initial success but subsided into small, quarrelsome groups of hunted revolutionaries when the two governments officially outlawed Communist agitation, in 1921 (Yugoslavia) and 1923 (Bulgaria). In Romania, Greece, and Albania, Communist organizers met with only slight response in the 1920s, but they did succeed in creating revolutionary cadres ready and willing to operate outside the law.

The frustration of peasantist aspiration and the prevalance of nationalist revolutionary sentiment among such groups as the Macedonians seemed in 1924 to offer Communists a chance to unite all the disaffected elements of Balkan society into a grand alliance. The Comintern (Communist International, formed in 1919 to help in spreading Leninism around the world) approved the formula of Balkan federation as a solution for the peninsula's political and economic ills and instructed each national party to form "popular fronts" with any and all available groups. Radić flirted openly with Moscow; so did leaders of the Internal Macedonian Revolutionary Organization (IMRO).

But these incompatible bedfellows soon parted. IMRO leaders who cooperated with the Communists were killed by rivals within the organization. Radić became a cabinet minister for a brief period (1925–26), but in 1928, during a session of the parliament, he was shot and killed by a Montenegrin Serb. This assassination provoked a strong reaction among the Croats against any kind of cooperation with the Serbs. King Alexander, therefore, decided to scrap the constitution, which had done so little to heal the fissures within his kingdom; he proclaimed a dictatorship, dissolved the political parties, and officially renamed the state Yugoslavia.

Government reactions. By resorting to authoritarian government, Alexander openly admitted the breakdown of effective government by consent. The same thing had also happened in Bulgaria, Romania, and Albania; but in those countries the pretense of parliamentary elections and the rituals of party coalitions were preserved, and they sometimes registered real adjustments in public mood.

This was the case, for instance, when Maniu forged a coalition of peasant parties in Romania and emerged as premier in 1928. In Greece, too, the parliamentary elections that returned Venizélos to power in 1928 registered public feeling quite accurately; but Greek electoral politics were frequently punctuated by coups d'etat (1922, 1925, 1933, 1935), that were sometimes successful, sometimes not.

Official efforts to cope with the problems of Balkan society and to meet the new revolutionary thrusts directed against existing governments were not entirely fruitless. They took three forms: land reform, industrialization, and police repression.

Land reform

Widespread redistribution of land shifted ownership toward the peasant families who actually worked the soil. Such reforms went fastest and farthest when land could be taken from owners of a different nationality. Thus, Magyar estates in the north and Turkish estates in the south disappeared at once. Gradual reapportionment took place elsewhere, too, so that by 1939 large estates had almost disappeared from the Balkans, surviving only in those parts of Romania where the landlords were politically powerful Romanian nationals.

The new owners of small peasant plots were often unable to cultivate the soil as efficiently as had the large-scale operators. Redistribution of land was no solution to the ills of Balkan overpopulation and underdevelopment. Each of the Balkan states recognized this fact by attempting to forward industrialization. Romania, building upon an oil industry that had arisen in the 19th century, made by far the most substantial progress in this direction; Bulgaria made almost none. Tariff protection, subsidy, and state enterprise were the devices used to develop industry. But shortage of capital and a generally low level of skills made progress painfully slow.

Industrialization

The only important policy difference in the interwar period with respect to industrialization was about how to treat foreign capital. Romania tried to finance industrial development from internal sources and actively discouraged further foreign investment in the oil industry, fearing that control and the real benefits of such expansion would pass exclusively to the foreign capitalists. Albania, under Ahmed Zogu (president, 1925–28; king, 1928–39), on the other hand, depended wholly on Italian capital for whatever modernization was achieved. Greece and Yugoslavia gave an ambivalent welcome to foreign investment, but political instability in these countries kept the flow of foreign capital to modest proportions. Bulgaria was more xenophobic and attracted almost no foreign funds.

Repression

The third major official response to the problems of Balkan society was repression. High-handed police action was common; elections were often "made" to suit the interests of those in power by overt use of army and police personnel; and sharp legal restriction, if not outright prohibition, was generally imposed on revolutionary political organizations and propaganda. Such measures were generally successful, even when the revolutionaries received systematic encouragement from abroad.

The Soviet Union provided at least moral support for Communists throughout the period. Fascist Italy spun a web of intrigue aimed against Yugoslavia; by the late 1920s Mussolini's agents had entangled the Bulgaria-based IMRO, extreme Croat nationalists (Ustaše) based in Rome, and some elements of the Hungarian government in plots to dismember King Alexander's state.

In Romania, too, a Fascist movement, founded in 1927 (renamed the Iron Guard in 1928) by Corneliu Codreanu, rose quickly to prominence; it owed little to foreign patronage, however, because Codreanu based his appeal mainly upon harsh anti-Semitism, for which the ground was already well prepared.

Alliances

The Balkan governments took steps to counter the foreign threat. A diplomatic alliance of Czechoslovakia, Romania, and Yugoslavia—the so-called Little Entente—dated from 1921. In 1934 a new Balkan Pact allied Greece, Yugoslavia, Turkey, and Romania. Such treaties were aimed mainly against Hungarian (Little Entente) and Bulgarian (Balkan Pact) aspirations for frontier revision. In the background, France played the role of great-power patron vis-à-vis Romania and Yugoslavia, rivalling Italy, the patron of Bulgaria and Hungary; Britain, the patron

of Greece; and the Soviet Union, the pan-Balkan patron of Communists.

Balkans in 1930s. The economic depression that settled upon world trade in the early 1930s put the Balkan countries at a great disadvantage. Farm prices plummeted, making it all but impossible for high-cost Balkan peasant producers to compete on world markets. Hardship and a pervasive sense of failure and confusion in official quarters encouraged revolutionary sentiment, especially among the young. But the enhanced power of revolution, especially in Communist guise, provoked more ruthless repression. Democratic and parliamentary government seemed to have failed everywhere, especially after 1933, when the rising influence of Nazi Germany began to make itself felt in the Balkans.

The effects of the world depression

German trade policy, as a matter of fact, brought an effective solution to the economic crisis that had paralyzed the Balkan markets since 1930. The Nazis offered to buy agricultural products from the Balkans, at prices fixed through bilateral negotiation, and offered manufactured goods in exchange, again at prices fixed by interstate bargaining. Germany often drove a hard bargain in these trade negotiations, but only in Germany could high-cost Balkan farm products find any kind of sale. Hence, these exchanges—administered by state trading agencies and kept track of through blocked accounts managed by state banks—benefitted all parties. Romania and Bulgaria, in particular, began to market substantial surpluses in Germany.

By the late 1930s the resulting economic improvement allowed precarious political stabilization in both countries. Thus, for example, King Boris III of Bulgaria suppressed IMRO by condoning, if not instigating, a coup d'état in 1936, which brought to power a military group that dispersed IMRO's gunmen and drove its leaders from the country. Similarly, in 1938 King Carol II of Romania was able to have Codreanu killed and to repress the Iron Guard through unscrupulous and highhanded police methods; the king owed his success to the fact that, despite the strong appeal of its anti-Semitism, trade revival had taken the cutting edge from the Iron Guard's revolutionary agitation.

Greece followed a similar route. The early 1930s saw a rising political crisis, with abortive coups d'état in 1933 and 1935. A "managed" plebescite in 1936 led to the recall of King George II from exile; but new elections resulted in a parliamentary deadlock between royalists and republicans. A handful of Communist deputies held the balance of power, being in a position to make or break any parliamentary majority. King George reacted to this situation by entrusting the government to Gen. John Metaxas, who ruled without a parliament until his death in 1941.

In Albania, however, King Zog's regime collapsed in 1939, and Italians took over direct administration of the country. This allowed Mussolini to threaten both Yugoslavia and Greece across a new frontier. Such an advance of Italian power was profoundly disturbing to both Balkan governments. Ever since 1934, when King Alexander of Yugoslavia had been assassinated in Marseilles by an IMRO gunman (supported by Hungary and Italy), Yugoslavia's internal problems had offered the Italians a promising field of action. Prince Paul, brother of the assassinated Alexander, took over the reins of government as regent for the heir, King Peter II, who was still too young to rule. Paul, like other Balkan monarchs, experimented with authoritarian rule and worked diligently for diplomatic rapprochement with both Bulgaria and Italy.

Italian control of Albania

In the late 1930s Paul decided that the kingdom could survive only by coming to a basic understanding with the Croats. Prolonged negotiations led to an agreement, in August 1939, by which a generously defined Croatia would enjoy extensive autonomy; the Croatian Peasant Party accepted this arrangement, and only certain extremists (the Ustaša) remained irreconcilable. This drastically weakened the Italian leverage inside Yugoslavia; but it did not solve the state's problems, because most Serbs balked at the prospect of giving the Croats control over territory in which large numbers of Serbs lived. Yet nothing less would satisfy Croatian ambitions. As a result, the federal structure for Yugoslavia as promised in the 1939 agreement had not been fully implemented when the flood tide of World War II overwhelmed the Balkan Peninsula.

World War II, 1939–45. *Axis victories and occupation.* Stalin's cooperation with Hitler against Poland (September 1939) turned previously proferred Anglo-French guarantees of Polish and Romanian territorial integrity into worthless scraps of paper. The Romanian government had to yield Bessarabia and Bukovina to the Soviet Union (June 1940), part of Transylvania to Hungary (August 1940), and the southern Dobruja to Bulgaria (September 1940). King Carol, discredited by such losses, fled the country, and Gen. Ion Antonescu became dictator. Antonescu invited German troops to enter Romania in October 1940. The Romanian government remained a loyal and relatively enthusiastic ally of the Nazis in their struggle against the Soviet Union until 1944.

The Italian government, jealous of Germany's successes in Poland, Scandinavia, and France, decided to make Greece into a dependency. When the Greek dictator Metaxas refused to yield to Mussolini's ultimatum (October 1940), Italian troops crossed the Albanian frontier, expecting to meet only token opposition. To their surprise and discomfiture, the politically divided Greeks joined ranks against the Italians and drove the attacking forces back across the Albanian border. Fearing the provocation of German military intervention, the Greek government reacted coolly to the British offers of air and naval support; these fears assumed new plausibility when Nazi troops moved secretly into Bulgaria in March 1941.

German advance into Bulgaria

The German move attempted to forestall Soviet influence in Bulgaria and to compel Greece and Yugoslavia to repudiate ties with Great Britain and align themselves with the Nazi cause. As a diplomatic prelude to their planned attack on the Soviet Union, the Germans demanded that each Balkan state adhere to the Tripartite Pact (concluded initially by Germany, Italy, and Japan in September 1940). Hungary and Romania did so in November 1940; Bulgaria followed suit on March 1, 1941; and the Yugoslavs reluctantly did the same a few weeks later (March 25). News of this act led Serbian radicals, already bitterly opposed to the deal the government had concluded with the Croats, to seize power in Belgrade. This act of defiance outraged Hitler, then at the summit of his diplomatic success; to safeguard his southern flank for the thrust against the Soviet Union, he determined to crush the Yugoslavs and Greeks in a lightning campaign.

Accordingly, on April 6, 1941, German forces attacked and speedily overran Yugoslavia. A hastily assembled British expeditionary force scarcely reached Greece from North Africa when headlong retreat began. In May 1941 German parachutists invaded Crete, and by the end of the month they had won a costly victory there; German arms had thus chalked up yet another brilliant success, but it forced them to postpone by a few weeks the start of the campaign against the Soviet Union. Whether this delay actually helped the Soviets survive Hitler's assault can probably never be definitively decided; but Greeks and Yugoslavs easily convinced themselves that the defeat of their countries was an essential precondition for the later Soviet victory.

Greece and Yugoslavia remained under Axis occupation until 1944–45. Bulgarian, Italian, and German troops carved out separate zones of occupation—most of Macedonia was incorporated into Bulgaria; an independent Croatia, ruled by the Ustachi and under Italian patronage, was proclaimed. Romania, too, staked out its territorial claims in Bessarabia and adjacent regions of the Ukraine.

Resistance movements. The first people who actively opposed these arrangements were Serbs, who had everything to lose and whose national tradition of heroic outlawry against the Turks inspired guerrilla action. Bands of demobilized Serbian soldiers—"Chetniks"—formed under the leadership of Col. Draža Mihailović and engaged in acts of sabotage. When the Germans sent fresh troops into Serbia and retaliated brutally against communities suspected of harbouring guerrillas, the Chetniks abandoned active operations, husbanding their strength against a future day of liberation when Germany's defeat would per-

mit them to restore the Serbian nation to its accustomed and proper position in the Balkans.

After the Nazi attack on the Soviet Union (June 22, 1941), the Communist parties of the Balkans, which had previously cooperated with the Germans, made an abrupt about-face. Communist activity in Bulgaria remained marginal until 1944; in Romania the party was of trifling importance, being firmly identified with Jewish and other minority groups.

But in Yugoslavia, Albania, and Greece the Communists rapidly built up powerful resistance organizations. The Communist Party line was to cooperate with all anti-Fascist elements in the population; hence, popular fronts, uniting Socialists, peasants, nationalists, and anyone else willing to cooperate with Communists, sprang into existence. Communist policy systematically played down Marxism and disguised the extent of party control over the network of political and military resistance organizations they created.

The popular fronts

An essential strength of the Communists in this situation was the availability in each Balkan state of underground party cadres already accustomed to survival in face of police harassment. In addition, the Communists emphasized political organization in towns and countryside, thus providing the armed guerrilla bands responsive to their leadership with far firmer support than any rival organizations could offer. Finally, Communist policy aimed at helping the common cause of Socialist revolution by fighting Germans wherever they could be found; this simple policy had the effect of attracting the support of restless and active men throughout the western Balkans. Hence, by degrees the Communists were able to live down their traitorously antinational past and to far outdistance all rivals.

In Yugoslavia, when the Communist-led Partisans first took the field in the summer of 1941, they tried, in accord with popular front tactics, to cooperate with Mihailović's Chetniks. But violent quarrels soon broke out. Two points were at issue: whether to persist in active operations against the Germans and Italians, despite the cost to civilian populations, as Tito (Josip Broz), the Communist leader, desired (Mihailović opposed this); and whether to welcome all Yugoslavs into the resistance, as Tito assumed, or accept only Serbs, as Mihailović felt was necessary.

Behind these disagreements lay rival visions of the future: Mihailović desired above all to protect Serbdom against Croat, Bulgar, and other threats, whereas Tito envisioned a revolutionary brotherhood of all Balkan nationalities, as projected by the Comintern ever since the 1920s. In the long-drawn-out struggle that ensued, advantage lay overwhelmingly with Tito; his policy of conducting active operations against the occupiers of his country won the backing of the British, American, and Soviet allies (Tehrán Conference, November 1943). In addition, as the war continued to bite into their daily lives the peoples of Yugoslavia more and more rallied to the only active transnational Yugoslav organizations in sight: Tito's Partisans and the Anti-Fascist Peoples' National Council (AVNOJ), the political voice and arm of the movement.

In Greece, the Communist resistance also took the form of an armed guerrilla organization, known by the acronym ELAS, and of a political organization, called National Liberation Front (EAM). British policy, however, never abandoned support for the Greek government in exile, headed by King George II. In Greece itself, anti-Communist guerrilla groups that were supported by British agents and supplies continued to divide the ground with ELAS.

The situation in Albania was similarly confused; rival Albanian resistance groups achieved effective organization only after 1943, when the surrender of the Italian government to the Allies in September put large stocks of arms into Albanian hands. The Italian surrender also allowed the resistance movements of Yugoslavia and Greece to acquire large quantities of Italian weapons and briefly to take possession of territories formerly policed by Italian troops.

Acquisition of Italian weapons

The Germans were able to reoccupy major cities and lines of communication in the former Italian zones of occupation; but by 1944 the continuing retreat of German armies in the Soviet Union made clear the ultimate outcome of the war. On August 23, 1944, when the advancing Soviet forces were close to the Romanian border, King Michael kidnapped General Antonescu and adroitly changed from the Axis to the Allied side. As a result, Soviet troops were able to advance rapidly toward the Danube, and Romanian units, formerly Hitler's allies, switched allegiances, turning their arms against the Germans.

The Bulgarians followed suit on September 9, as the Soviet advance guard neared their border. The Soviets then turned westward, passing through a corner of Yugoslavia (Belgrade was liberated, October 20, 1944) on their way to Budapest.

This abrupt military reversal compelled the Germans to withdraw what forces they could from the western Balkans. Isolated German garrisons, such as that in Crete, were left behind; but the main forces moved northward in good order, and at the end of the war (May 1945) much of Croatia still remained under German occupation.

The last Germans left Greece in October 1944, whereupon the royal Greek government returned to Athens (October 18) with the protection of a handful of British troops; quarrels soon broke out, however, and, during six bitter weeks (December 1944 to January 1945), fighting flared in Athens between British forces and the resistance guerrilla army, ELAS.

World War II to present. *Postwar settlement, 1945–49.* The armed collision between British- and Communist-led forces in the streets of Athens showed how hard it was, as the prospect of final victory over Germany came closer, for the great Allied powers to keep on cooperating. In May 1944 the United States had reluctantly approved a division of the Balkans into British and Soviet spheres of influence; this was confirmed and adjusted in favour of the Soviet Union when the British prime minister, Winston Churchill, visited Moscow that September to settle details of armistice arrangements with Romania and Bulgaria and to clear the way for the British landing in Greece.

British and Soviet spheres of influence

The Soviets were not unhappy to see the British resort to high-handed means to enforce their will in Greece. They hoped for a similar free hand in Romania, where the absence of any strong native Communist party made it particularly difficult for them to establish a government they could depend on. The Bulgarian Communist Party, though, was relatively strong and, operating through a "Fatherland Front," gave the Russians no cause for concern.

In Yugoslavia, however, where Tito easily took power as the Germans withdrew, the Soviets found it hard to restrain the Partisans' revolutionary enthusiasm. In the first flush of victory, Tito's followers were eager to put the Communist recipe for Balkan federation into effect and saw no reason to compromise with the "effete capitalist imperialists" of Britain and the United States over such an issue as control of Trieste. When the Soviets, in conformity to their deal with Churchill to divide the Balkans into spheres of influence, advised Tito to make conciliatory gestures towards the Royal Yugoslav government (in exile since 1940), the Yugoslav Communists reluctantly complied by allowing King Peter's representative, Ivan Šubašić, briefly to join the Cabinet. But the Partisans had come to power by taking risks and despising compromise; accordingly, many of Tito's followers were appalled at the Soviet Union's unwillingness to act upon revolutionary principles in the immediate postwar period.

The British, American, and Soviet foreign ministers slowly negotiated peace treaties with the former enemy states of the Balkans (1945–47). As long as the Anglo-Americans had not recognized the postwar Communist-dominated regimes of Bulgaria, Romania, and Hungary, the Soviets had a powerful argument for moderation; Stalin accordingly supported collaboration with all anti-Fascists and opposed further revolutionary adventures in the Balkans.

In Romania and Hungary there were strong practical reasons why out-and-out Communist Party dictatorships could not come to power—national feeling was distinctly anti-Soviet Union and, therefore, anti-Communist, and pre-existing Communist Party structures were extremely weak. But even where there were strong Communist parties, as in Bulgaria and Yugoslavia, until 1947 Soviet policy continued to support anti-Fascist popular fronts in

which Communist preponderance was at least partially camouflaged.

Changes in Soviet policy toward the Balkans

In 1947 two events altered Soviet policy toward the Balkans. First, the United States and Great Britain signed peace treaties with Bulgaria, Romania, and Hungary in February. These treaties did not much alter interwar boundaries, although Romania did lose Bessarabia and Bukovina to the Soviet Union once again and ceded a small strip of Dobruja to Bulgaria; Greece acquired the Dodecanese Islands in the Aegean from Italy; and Yugoslavia annexed a strip of territory in the Istrian peninsula. The treaties did, however, disband the Allied armistice commissions, which had exercised some control over local affairs in both Bulgaria and Romania following 1944.

The cancellation of British and American legal claims to authority in these countries freed local Communists from what had been a real, if modest, hindrance. At the same time, because the treaties authorized the Soviets to maintain their forces in Romania as "lines of communication" troops to forward garrisons in Austria, the Communists retained what was still an essential guarantee of their hold over Romania.

The second event that altered Soviet policy occurred in March 1947, just a month after the peace treaties came into effect, when Pres. Harry S. Truman asked the United States Congress to authorize military and economic aid to Greece and Turkey, both of which were under Communist pressure. After prolonged debate, Congress voted to approve Truman's request and thereby endorsed the "Truman Doctrine," according to which the United States undertook to defend the rule of law internationally by coming to the aid of any government threatened with Communist subversion.

The U.S. decision to back the Greek government against a renewed Communist guerrilla movement signalized a sharp shift of American policy vis-à-vis the Soviet Union. Quarrels over the peace treaties, over Poland, over the occupation regimes in Germany and Japan, as well as over the rising tide of Communist power in China and several other parts of Asia, all contributed to the change of American public attitudes; but events in the Balkans were also important.

Attempts to form a Balkan confederation

In 1945 and 1946, through a series of diplomatic notes, the Soviets tried to convince the Turks to give them military bases on the straits between the Black Sea and the Aegean. In addition, Tito and other Balkan Communists, most notably Georgi Dimitrov of Bulgaria, set about actively implementing their ideal of Balkan confederation. A first step was to federalize Yugoslavia itself. A new constitution, closely modelled on that of the U.S.S.R., was accordingly promulgated in January 1946; it established six federal republics (Serbia, Slovenia, Croatia, Montenegro, Macedonia, Bosnia) and several autonomous regions. King Peter lost his throne, and Communist Party control came fully into the open.

A second step was to bring a reliable Communist regime to power in Albania. The Yugoslavs sent troops and technical advisers to help Albanian revolutionaries seize firm control (January 1946). Soon the entire country began to behave much like another constitutive republic of Tito's emerging Balkan superstate.

The next item on the Communists' agenda was never realized: the creation of a united Macedonia that would combine the portions of that land belonging to Greece, Bulgaria, and Yugoslavia into a single whole. Bulgaria did in fact briefly cede Pirin Province to the new Macedonia, but the Greeks refused to cooperate. Accordingly, Tito sent Greek veterans of the wartime guerrilla force (who had retreated into Yugoslavia at the end of the war) back to Greece, where they formed the core around which fresh bands of guerrillas quickly formed in 1946 and 1947.

The Greek government's efforts to repress this renewal of guerrilla activity were ineffective; British resources were too straitened at home to permit fresh involvement in Greece. Hence, for a few months—until the United States committed itself fully to stopping the Communist advance—Tito's revolutionary policy seemed on the verge of paying off.

Such heady prospects encouraged the Yugoslavs to be aggressive along their frontier with Italy, demanding further territory to unite all Slovenes in the new Slovene republic. They also entered into negotiations with the Bulgarians (and perhaps also with Romanian Communists) for merging their countries into the proposed Balkan federal state.

Tito's activity antagonized the United States and was a potent factor in persuading the U.S. Congress to support the embattled Greek government in 1947. Tito's effect on Soviet policy remains a matter for speculation; Stalin probably backed the idea of abandoning the popular-front tactic and the placing of out-and-out Communist regimes in power. Soviet diplomatic agents played the key role in driving King Michael from the Romanian throne (December 1947), thus instituting Communist Party dictatorship in that country. Bulgarian Communists needed no outside help to achieve the same result by the end of 1947.

Tito's break with the Soviet Union

Eventually, Tito's efforts to overthrow the royal government in Greece and to press ahead with Balkan federation alienated the Soviets. Perhaps Stalin feared Tito's independence or attributed the United States' involvement in Greece and Turkey to what he saw as Tito's recklessly revolutionary policies.

At any rate, in 1948 the Soviet dictator decided to call Tito to heel. With characteristic guile, Stalin set out to overthrow Tito by stirring up an intrigue within the ranks of the Yugoslav Communist Party. It did not work; Tito's prestige inside his own country was too great for an outsider—even Stalin—to succeed in unseating him.

When the quarrel between the Soviet Union and Tito came out into the open (June 1948), the entire strategic situation in the Balkans altered abruptly. Albanian Communists, warmly backed by the Soviet Union, broke away from Tito and unceremoniously evicted all the Yugoslav experts and advisers who had until then been running the country. Bulgaria and Romania disclaimed all sympathy for Tito and hurried to participate in Stalin's economic blockade and propaganda war against the stubborn Yugoslavs (all the more energetically because of their previous associations with the new heretic).

In Greece, the Tito–Stalin split meant, first of all, the cessation of Yugoslav aid to the Greek guerrillas. This crippled the guerrilla cause. Then, in an effort to get the Macedonians to imitate the Albanians and secede from Tito's dominion, the Cominform (Communist Information Bureau, established in 1947) announced its program of a united Macedonia. Radio broadcasts failed to stir the Yugoslav Macedonians to action; but the news did create consternation in the ranks of the Greek guerrillas, who were unwilling to fight for a cause that, it now appeared, would lead, if successful, to the surrender of Greek territory to a Slav people. Greek Communist morale therefore collapsed, so that Greek government troops—equipped, advised, and assisted by a large American military mission—found it easy to win a decisive victory in the summer of 1949.

Changes caused by the war. By August 1949, therefore, active military operations in the Balkans ceased; nearly a decade of war thus came to a conclusion. By that date Communist Party dictatorships were firmly established in all the Balkan countries except Greece and Turkey.

Yet, in spite of this fact, the changes World War II brought to the Balkans were distinctly less drastic than those that came during the 20th century's earlier decade of Balkan fighting, 1912–23. Boundaries shifted only slightly after 1945; and war or postwar population movements wiped out or greatly reduced the numbers of some national minorities—Germans and Jews in particular. The effect was to confirm and enhance the sorting out of Balkan populations into territorially defined national states, according to the patterns of 1918–19. From a political and social point of view the one-party regimes of the northern Balkans, set up by Communists, proved a little more ruthless and persistent in pursuit of the same goals that interwar Balkan governments had pursued: economic and political mobilization of the peasant mass—by police hectoring and controlled elections if need be—to hasten the development of industries and cities.

Confirmation of national states

All in all, the major crisis of transition from a traditional peasant and premodern style of life apparently took place

in the Balkans before and after World War I, whereas after 1949 somewhat more stable patterns of modernization and mobilization established themselves in Greece as well as in Communist lands. The waning force of revolutionary movements of every kind, a marked feature of the post-World War II Balkan scene, is an index of this basic transformation.

1950s and after. Though Balkan politics after 1949 were less tumultuous and also less bloody than was the case earlier, confusion and upheavals were not lacking. Among the Communist states, fluctuating relations with the Soviet Union defined major shifts of government policies.

Political changes

From 1948 to 1953 all of Stalin's satellites in eastern Europe purged real or imagined "Titoists"—*i.e.,* Communist leaders suspected of putting national interests ahead of subservience to the Soviet Union. The countries formerly closest to Tito were, not surprisingly, the most energetic in carrying out such purges.

Albania broke away from Yugoslavia and became for a while the Soviet Union's most enthusiastic satellite. The Bulgarian government staged a show trial against Titoists and used the occasion to implicate personnel of the United States embassy, with the result that official U.S.–Bulgarian relations were broken off in 1950, not to be resumed until 1966. Yet, in spite of a trade blockade and various kinds of subversive activity aimed against the Yugoslav regime, Tito's power remained unshaken. By cautiously accepting proffered American aid, the Yugoslavs managed to survive even Stalin's wrath.

After Stalin's death in 1953, the Soviet government tried to mend its fences with Tito. As a result, the Yugoslav government was able to balance itself between the Soviet Union and the United States—playing one off against the other and even getting economic and military aid from both great powers at once. The advantages of such an independent policy were obvious to other Balkan Communist states. But not until after 1961, when a quarrel between Chinese and Soviet Communists came into the open, did first the Albanians and then the Romanians venture to defy the Soviet Union.

Albanian and Romanian policies

Albania's policy was governed mainly by fear of a lasting Yugoslav–Soviet rapprochement; when that seemed likely in 1961, the Albanians threw out their Soviet advisers, as they had earlier ejected Yugoslavs from similar roles, and allowed the distant Chinese to become their new patrons. The Chinese sent substantial aid to Albania in the ensuing years, but after Mao Tse-tung's death in 1976, the Albanians quarrelled with Mao's successors. In 1978 the Chinese accordingly cut off all aid, leaving Albania without a foreign protector for the first time since 1912.

A far more serious break in the Communist ranks occurred in 1964, when Romania declined to accept the role assigned to it by the Soviet Union in the Comecon (Council for Mutual Economic Assistance) plan for economic development of eastern Europe. Instead of concentrating on agriculture for export, the Romanian government wished to continue to emphasize industry, even if this meant duplicating Czech or East German factories. Romanian national feeling, always strongly anti-Soviet, surged to the surface in support of the government's stand. In the following 15 years, until Polish trade unionists began to challenge Communist practices, Romania remained the Soviet Union's most vocal eastern European critic and gadfly. In late 1979, however, President Nicolae Ceaușescu found it necessary to import Soviet, in lieu of Iranian, oil. It was thought that this new link might diminish friction between the two countries, or at least muffle its public expression.

Post-war policy in Greece

In Greece, U.S. policy played a dominating role after 1947. Until 1956 U.S. funds continued to provide the Greek government with substantial aid for economic rehabilitation; the United States thus was able to control a number of important aspects of government policy.

After 1956, aid tapered off, and U.S. diplomats deliberately tried to pull out of Greek politics. Such aloofness was all the more attractive to the United States because Greece seemed to have attained a relatively stable parliamentary regime under Marshal Alexandros Papagos—hero of the Albanian war of 1940 and of the guerrilla war of 1947–49—and, after his death (1955), Konstantinos Karamanlis, who survived three elections to remain prime minister for an unprecedentedly long period (1955–63).

Yet the fact that Greece had become a member of the North Atlantic Treaty Organization (NATO) in 1952 meant that U.S. influence upon the Greek armed forces remained strong even after other forms of aid had stopped. And, because the political attitudes of the army mattered in Greek politics, U.S. efforts to leave the Greek politicians alone were never very successful. Thus, in April 1967, when a clique of army officers seized power by coup d'état, U.S. agents were generally suspected of being responsible for what was probably planned in secret by a narrow and purely Greek circle. The policy of the new Greek government lent some colour to the charges, however, because the officers who emerged as rulers of the country combined an enthusiastic anti-Communism with an earnest courtship of private foreign (principally U.S.) investment, which they needed to balance their international accounts.

Military government in Greece

The authoritarian military dictatorship, headed by Col. Georgios Papadopoulos, lasted until 1974, when intrigues aimed at bringing Cyprus into union with Greece miscarried and the military rulers of the country were discredited. Karamanlis was recalled from his self-imposed exile in Paris and organized new elections, thus restoring the legal forms of parliamentary democracy. In a national referendum in December 1974 the Greeks rejected the monarchy and became a republic. Relations with Turkey, which had nearly come to war over Cyprus, were not easily healed; and indignation at the failure of NATO commanders to support the Greek cause against Turkey led Karamanlis to cancel most of the agreements that had bound the Greek armed forces to NATO. Subsequently, long drawn-out negotiations restored some links with NATO, but the Greek government retained fuller control over NATO and U.S. bases on Greek soil than before. Efforts to normalize relations with Turkey met with more limited success. Far more important was the fact that on January 1, 1981, Greece became a member of the European Economic Community (EEC). Important changes in the Greek economy were likely to flow from this move, which also symbolized a psychological and cultural commitment on the part of the Greeks to the Western style of life.

Establishment of the Greek republic

Economic changes

The political upheavals and often noisy public debate that characterized Greece stood in sharp contrast to the strict controls on public discussion that Communist regimes enforced in other Balkan countries. This difference, however, should not disguise some important resemblances in the pattern of economic development pursued by all Balkan governments in the post-World War II era. Everywhere officials preferred industrial to agricultural investment and subjected most economic activities to elaborate, often cumbersome, bureaucratic control. Everywhere, too, cities grew rapidly, offering scope for peasant sons and daughters to pursue new careers more attractive than those their parents had known. This permitted a general relaxation of political and economic dissatisfactions, despite the numerous unsolved problems that persisted within all the Balkan nations.

The only important difference in economic policies was that Greece continued to import private foreign capital and linked internal prices to world markets, whereas Communist countries maintained a command price system at home and regarded all private investment as detestable capitalist exploitation. When importing foreign capital, the Communist countries were willing only to borrow from other governments—a policy that seemed unlikely to make borrowers any less dependent on lenders.

Tangible increases of industrial output occurred in every Balkan state. This did not mean, however, that problems of distribution and assortment of products to suit the consumers had been solved. Quite to the contrary, in the Communist lands shoddy goods and insufficient supplies of some commodities were still the rule; and many of the new factories, built as a result of political decisions, were distressingly inefficient when it came to costs. In Greece, unsolved problems were primarily those of income distribution among the different social classes, although inefficient and high-cost factories that sheltered behind tariff and quota protection were not absent.

Decentralization in Yugoslavia

Recognition of the difficulty of adjusting a command economy to consumer needs led the Yugoslavs to experiment with freer market prices on the one hand and with worker–management partnership in factory administration on the other. A new constitution, effective in 1963, gave enhanced power in economic matters to the federal republics in the hope that this, too, would bring planning and management closer to the people. Yugoslavia met some success in this effort at decentralization and bureaucratic simplification. But divisive tendencies among the various nationalities of the state were perhaps reinforced by allowing greater popular participation in management decisions. Nevertheless, on June 30, 1971, the federal government instituted amendments that further enhanced the powers of the republics and municipalities, granting the right, for example, to initiate investment projects. The federal government reserved for itself sole power in the areas of defense, foreign policy, and general economic policies. A new constitution promulgated in 1974 went even further toward institutionalizing "workers' " management, but frictions among the separate nationalities of Yugoslavia were not completely damped down by the policy of decentralization; and Marshal Tito's death in May 1980 brought the regime a new test of its stability. Tito was succeeded by a collective presidency representing each of the six republics and two autonomous provinces into which Yugoslavia was juridically divided. Internal diversity was thus proclaimed symbolically, but a far more centralized Communist party structure continued to counterbalance fissiparous tendencies.

In the 1970s, all of the Balkan countries faced new difficulties in economic management. Sharp increases in the international price of oil on the one hand and rising expectations at home on the other made planned investment in heavy industry difficult to achieve and, in fact, planned expansion was not realized. Deficit accounts in international trade and domestic inflation at rates considerably higher than those experienced by the more highly industrialized nations of Europe registered the gap between current productivity and the demand for goods and services.

Despite these unsolved difficulties and the repressive police regimes that supported the Communist monopoly of political power, Balkan history since 1949 must be considered an overall success, at least when compared to the earlier decades of the 20th century. Peace prevailed as seldom before, coups d'etat were far fewer, and urbanization increased by leaps and bounds. Most significant of all, age-old routines of peasant life altered profoundly. Family patterns changed so that lower birth rates prevailed in every country except Albania. This reduced population pressure and made rising standards of living easier to attain. Even in poor villages, new, powerful farm machinery lightened the labour of cultivation and increased production. The Balkan peasantries also accepted new ideas about the world and its possibilities, largely through daily exposure to radio and television broadcasts.

The Balkan nations, in short, struggled to transform themselves into provincial variants of the world-girdling Western style of civilization. Their success provoked new dissatisfaction to be sure, but nothing to compare with the fierceness of older revolutionary aspirations that had kept the Balkans in turmoil from 1850 to 1950. (For additional information on contemporary affairs in the Balkans, see ALBANIA; GREECE; ROMANIA; BULGARIA; and YUGOSLAVIA.)

BIBLIOGRAPHY

Physical and human geography. The physical geography of the Balkans is summarized in the AUSTRIAN EAST AND SOUTHEAST EUROPE INSTITUTE'S *Atlas of the Danubian Countries*, issued in fascicles (1971–). Long-term relations between the landscape and its human occupants are examined in FRANCIS W. CARTER (ed.), *An Historical Geography of the Balkans* (1977). The economic history of the region is provided by JOHN R. LAMPE and MARVIN R. JACKSON, *Balkan Economic History, 1550–1950: From Imperial Borderlands to Developing Nations* (1982); NICHOLAS V. GIANARIS, *The Economies of the Balkan Countries: Albania, Bulgaria, Greece, Romania, Turkey, and Yugoslavia* (1982). Detailed studies of the region prior to World War II include M.I. NEWBIGIN, *Southern Europe*, 3rd ed. (1949); C.S. COON, *The Races of Europe* (1939); M.E. DURHAM, *Some Tribal Origins, Laws and Customs of the Balkans* (1929); M.S. FILIPOVIĆ "The Bektashi in the District of Strumica (Macedonia)," *Man.*, vol. 54, no. 7 (1954); M.M. HASLUCK, *The Unwritten Law in Albania* (1954); E. PITTARD, *Les Peuples des Balkans* (1920) and *Race and History* (1926).

History. *The Balkans to 1815:* JOHN ALEXANDER, *Yugoslavia Before the Roman Conquest* (1972), a general summary of archaeology and protohistory of this country with emphasis on the early Iron Age; DUMITRU BERCIU, *Romania* (1967), archaeological survey from the Paleolithic period to the Geto-Dacian civilization; STANLEY G. EVANS, *A Short History of Bulgaria* (1960), includes an outline of the early historical and cultural setting for the state; JOSEPH WIESNER, *Die Thraker* (1963), a detailed account of the Thracians to AD 600; MARIJA GIMBUTAS, "The Neolithic Cultures of the Balkan Peninsula," in *Aspects of the Balkans* (1972), a concise account of the Neolithic and Chalcolithic civilizations, 6500–3500 BC; *The Bronze Age Cultures in Central and Eastern Europe* (1965), a monograph that includes the Danubian region and Urnfield migrations; *The Slavs* (1971), an account of the early history of the Slavic peoples and their migrations to the Balkan peninsula; and the *Symposium* of the Centre d'Études Balkaniques, Sarajevo (1964), a collective work dedicated to the distribution and chronology of the prehistoric Illyrians. (*The Balkans in the Middle Ages*): Two books are quite reliable and up to date: GEORG STADTMULLER, *Geschichte Südosteuropas* (1950); and DIMITRY OBOLENSKY, *The Byzantine Commonwealth: Eastern Europe, 500–1453* (1971). The Byzantine context of the Balkan peninsula in the medieval period is well treated in the relevant chapters of *The Cambridge Medieval History*, 2nd ed., vol. 4 (1966); and on a more modest scale in GEORGE OSTROGORSKY, *Geschichte des byzantinischen Staates* (1965; Eng. trans., *History of the Byzantine State*, 2nd ed., 1968). On the Slavic states in the Balkans, see the two studies of FRANCIS DVORNIK, *The Slavs: Their Early History and Civilization* (1956) and *The Slavs in European History and Civilization* (1962). All these books have substantial bibliographies. (*The Ottoman era*): The best general discussion is to be found in L.S. STAVRIANOS, *The Balkans Since 1453*, ch. 3–12 (1958). The Ottoman conquest of the Balkans and its effects are ably described in PAUL COLES, *The Ottoman Impact on Europe* (1968); STEVEN RUNCIMAN, *The Fall of Constantinople, 1453* (1965); and THOMAS M. BARKER, *Double Eagle and Crescent: Vienna's Second Turkish Siege and Its Historical Setting* (1967). STEVEN RUNCIMAN, *The Great Church in Captivity* (1968), covers the patriarchate in Constantinople from the Turkish conquest until the 1820s. DIMITRIJE DJORDJEVIC and STEPHEN FISCHER-GALATI, *The Balkan Revolutionary Tradition* (1981), examines the nationalist revolutions in several Balkan states from the 16th to the 20th century.

The Balkans from 1815 to 1914: The period from the early 19th century until the 1950s is treated in RENE RISTELHUEBER, *Histoire des peuples balkaniques* (1950; Eng. trans., *A History of the Balkan Peoples*, 1971). A short review of the Balkans in the 19th century may be found in L.S. STAVRIANOS, *The Balkans, 1815–1914* (1963); and in CHARLES and BARBARA JELAVICH, *The Balkans* (1965). See also the relevant sections in Stavrianos' excellent general history, *The Balkans Since 1453*. The comparative method is applied in L.S. STAVRIANOS, *Balkan Federation* (1944); TRAIAN STOIANOVICH, *A Study in Balkan Civilization* (1967); CHARLES and BARBARA JELAVICH (eds.), *The Balkans in Transition* (1963), *The Establishment of the Balkan National States, 1804–1920* (1977); DIMITRIJE DJORDJEVIC, *Révolutions nationales des peuples balkaniques, 1804–1914* (French trans. 1965); DOREEN WARRINER (ed.), *Contrasts in Emerging Societies: Readings in the Social and Economic History of Southeastern Europe in the Nineteenth Century* (1965).

The Balkans after 1914: (*General works*): HUGH SETON-WATSON, *Eastern Europe Between the Wars, 1918–1941*, 3rd rev. ed. (1967); and *The East European Revolution*, 3rd ed. (1956); ROBERT L. WOLFF, *The Balkans in Our Time* (1956); R.V. BURKS, *The Dynamics of Communism in Eastern Europe* (1961); GHITA IONESCU, *The Politics of the European Communist States* (1967); YORICK BLUMENFELD, *Seesaw: Cultural Life in Eastern Europe* (1968); C.A. MACARTNEY and A.W. PALMER, *Independent Eastern Europe* (1962); A.W. PALMER, *The Lands Between: A History of East-Central Europe Since the Congress of Vienna* (1970); CHARLES JELAVICH (ed.), *The Balkans in Transition* (1963); STEPHEN A. FISCHER-GALATI (ed.), *Eastern Europe in the Sixties* (1963); NORMAN J.G. POUNDS, *Eastern Europe* (1969); J.F. BROWN, *The New Eastern Europe: The Khrushchev Era and After* (1966). (*On more specialized pan-Balkan topics*): W.E. MOORE, *Economic Demography of Eastern and Southern Europe* (1945); IRWIN T. SANDERS (ed.), *Collectivization of Agriculture in Eastern Europe* (1958); ALFRED BOHMANN, *Menschen und Grenzen*, vol. 2, *Bevölkerung und Nationalitäten in Südosteuropa* (1969).

(M.I.N./S.H.Br./W.B.T./A.Lu./W.C.B./M.G./G.P.M./B.Je./C.J./D.V.D./W.H.McN.)

Bangkok

Bangkok is the capital and chief port of Thailand. It is the only cosmopolitan city in a country of small towns and villages and is Thailand's cultural and commercial centre.

Bangkok is located on the delta of the Chao Phraya River, about 25 miles (40 kilometres) from the Gulf of Thailand. It was formerly divided into two municipalities—Krung Thep on the east bank and Thon Buri on the west—connected by several bridges. In 1971 the two were united as a city-province with a single municipal government. In 1972 the city and the two surrounding provinces were merged into one province, called Krung Thep Maha Nakhon (Bangkok Metropolis). Bangkok Metropolis has a total area of 604 square miles (1,565 square kilometres). It is a bustling, crowded city, with temples, factories, shops, and homes juxtaposed along its roads and canals.

The name Bangkok, used commonly by foreigners, is, according to one interpretation, derived from a name that dates to the time before the city was built—the village or district (*bang*) of wild plums (*makok*). The Thai call their capital Krung Thep, which is the first part of its mellifluous and lengthy official name meaning "the City of Gods, the Great City, the Residence of the Emerald Buddha, the Impregnable City (of Ayutthaya) of God Indra, the Grand Capital of the World Endowed with Nine Precious Gems, the Happy City Abounding in Enormous Royal Palaces Which Resemble the Heavenly Abode Wherein Dwell the Reincarnated Gods, a City Given by Indra and Built by Vishnukarm." The abbreviated name Krung Thep is often translated "City of Angels."

This article is divided into the following sections:

Physical and human geography

THE LANDSCAPE

Climate. The climate of Bangkok is hot throughout the year, ranging from 77° F (25° C) in the "cold" season in December to 86° F (30° C) at the height of the hot season in April. The mean annual rainfall totals 60 inches (1,500 millimetres), four-fifths of which falls in brief torrential downpours during the late afternoons of the rainy season, which lasts from mid-May through September; the dry season lasts from December to February. Mean monthly relative humidity varies from a low of 60 percent in the cold season to more than 80 percent during the rainy season.

The city layout. Modern Bangkok has undergone explosive growth, which the authorities have only recently attempted to direct by means of a series of master plans. The city centre, formerly enclosed by a wall, has long been densely developed, while later expansion has sprawled outward well beyond the administrative boundaries into the surrounding agricultural areas. Some districts have evolved into functional units as the inner city has become more institutional and commercial and the outer city more residential and industrial. Throughout the city, walled Buddhist temples and monasteries called wats, often sumptuously ornamented, serve as focal points for religious, cultural, and even commercial life.

Traditional areas. The governmental and commercial districts of the city occupy traditional sites. Government offices were originally housed in the walled compound of the 18th-century Grand Palace, but by the late 19th century they occupied surrounding palaces and mansions. The bureaucracy then spread outward into nearby colonial-style or Thai-style office buildings and homes along Ratchadamnoen Road. Multistoried buildings have been erected to meet the ever-increasing demand for space, and the traditional government compounds have become overbuilt. A number of large camps around and north of the National Assembly Hall constitute the military area.

Principal districts of the city

When Bangkok became the national capital in the 18th century and its citadel was moved to the east bank of the Chao Phraya River, Chinese merchants and tradesmen occupying the site moved a short distance southward to the area now known as Sam Peng. Business was at first carried on in one-story wood and thatch houses. By the early 1900s a number of streets had been lined with two-story masonry shop-houses. This ever-expanding district now contains rows of shop-houses that are sometimes five or more stories high. Warehouses line both banks of the river just south of Sam Peng, while industry is concentrated at Sam Rong, south of the port. Nightlife flourishes on Pat Pong Road. The financial district straddles Silom Road.

In the Floating Market a variety of foods and merchandise are sold daily from boats on the canals near Wat Sai. Formerly several such markets and innumerable door-to-door floating vendors served the daily needs of the city's residents.

Housing. Homes generally consist of small, detached one- or two-story wooden houses or row houses. Most of these are overcrowded because there are far too few of them to house the expanding population. Government programs alone are insufficient to meet the housing shortage, and funds from the World Bank have been used to build low-income housing, such as the Din Daeng and Hua Mak developments.

Housing shortage

The government allows squatters to occupy unused public land. The number of squatters is small, and most of them are concentrated in the Khlong Toei area near the port. Private real-estate developers provide homes for middle-income groups, and many government agencies provide homes for their employees. Homes may be crowded onto small lots with rudimentary sanitation facilities. These developments have spread out haphazardly on the periphery of the city.

Luxury housing, mostly for the wealthy foreign community, usually takes the form of large, modern, two-story masonry structures set in private compounds and equipped with separate servants' quarters and kitchens. Bang Kapi is perhaps the most affluent neighbourhood. High-rise offices, hotels, and condominiums are increasingly common.

THE PEOPLE

The population's outstanding demographic characteristics—its youth and the low proportion of non-Thais—are explained by the high rate of natural increase and by the restrictive foreign immigration quotas adopted after World War II. Two-fifths of the residents are under 20 years of age; another one-fifth is under 30. The birth rate has declined since the introduction of a birth-control program. At the same time, the net in-migration of young

adults, particularly females, has increased greatly, so that more than a quarter of the resident population of the city is made up of migrant Thais from all parts of the country.

Less than 3 percent of the population is non-Thai. The Chinese are by far the largest minority, but there are sizable communities of other Asians, North Americans, and Europeans. Despite their small size, the foreign communities tend to live in certain areas. The Chinese concentrate in the commercial area of Sam Peng, Indians gather around mosques in the Wang Burapha section, and the Western and Japanese communities reside in the affluent, modern eastern section of the city.

Of the foreign groups, the Chinese enter the most intimately into city life. They appear to assimilate readily, and intermarriage is frequent. Their offspring are Thai citizens, and many Chinese families take Thai surnames and are naturalized.

Population density is highest in the Chinese district of Sam Peng, where it reaches about 390,000 persons per square mile (150,000 per square kilometre).

THE ECONOMY

Industry. There are many factories in the metropolitan area, but most operate on a small scale. Larger plants are located in the vicinity of the port, near the warehouses that store imported materials. Manufacturing is chiefly confined to food processing, textiles, the assembly of electronic equipment, and the production of building materials. After 1976, however, no new factories were permitted to be built in Bangkok, and the government has given high priority to locating industry in industrial parks on the outer fringes of Bangkok. Tourism has increased greatly and is now a major source of revenue in Bangkok.

Finance. Bangkok houses about one-third of the country's banking units, holding three-fourths of all deposits. The Industrial Finance Corporation of Thailand, the Board of Investment, and the Securities Exchange of Thailand are also located in the city.

Transportation. Bangkok's transportation system was originally based on water travel. The city's maze of canals connected with the river earned it the name Venice of the East. The advent of the automobile, however, brought drastic changes. The number of vehicles in the city (including three-wheeled taxis, private cars, and buses—colour coded according to the region of service) has increased, and a shortage of road space has developed. The problem was met first by filling in most of the smaller and a number of the larger canals. This proved to be more than an aesthetic loss, however, because the waterway system had served to drain the waterlogged delta; flooding of the lower-lying parts of the city thus became increasingly frequent. Furthermore, the measure did not solve the problem of lack of space; traffic became so congested that movement was increasingly difficult. To help ameliorate these problems, an authority was established in the 1970s to oversee bus transportation in the city.

Lines of communication radiate outward from the city. Roads run north to Laos and Chiengmai, east to Kampuchea, and south to Malaysia; railways run to the borders of Laos and Malaysia, to Chiengmai in the north, and to Ubon Ratchathani and the Kampuchean border in the east. Don Muang airport, one of the busiest in Southeast Asia, is served by international airlines.

The port of Bangkok, located on the Chao Phraya, at Khlong Toei, is connected to the sea by a channel dug through the sandbar at the river mouth some 17 twisting miles downstream. The port handles nearly all of the nation's imports and exports.

ADMINISTRATION AND SOCIAL CONDITIONS

Government. The government of Bangkok Metropolis is administered by a governor and deputies. Developmental responsibilities rest with a large number of governmental agencies. Bangkok houses the headquarters of the United Nations Economic and Social Commission for Asia and the Pacific (ESCAP). In addition, the city houses various other UN agencies, including branch offices of the World Health Organization (WHO), the International Labour Organisation (ILO), the United Nations Children's Fund (UNICEF), and the International Bank for Reconstruction and Development (World Bank).

Public utilities. Most of the city's water supply comes from purification plants; it is drawn from the Chao Phraya and from deep wells. The pumping of water from wells has caused subsidence in parts of the city, which has increased flooding.

Sanitation facilities consist partly of sewage storm drains and the canals; large buildings are often equipped with septic tanks. Bangkok consumes more than half of the country's electric power.

Health. Bangkok has most of the country's hospitals and clinics. Special services are offered for patients with tuberculosis and venereal disease, and there are government homes for the indigent, handicapped, and aged. The Pasteur Institute and WHO supply vaccines. Family-planning clinics have proliferated in recent years.

Education. Because of its high proportion of school-age citizens, Bangkok's educational facilities are overburdened. There are too few schools, and the standard of instruction varies. Literacy is extremely high, however.

Many of the government-built preprimary and primary schools are located on monastery grounds. Private primary and secondary schools run by foreign religious missions train the children of the elite. There are many private Chinese primary schools and night schools. The city has several universities.

CULTURAL LIFE

The most important cultural feature of Bangkok is the wat. There are more than 300 such temples, representing classic examples of Thai architecture. Most are enclosed

Buddhist monk at the Wat Arun in Bangkok.

by walls. Many wats have leased a portion of their grounds for residential or commercial use.

The National Museum houses prehistoric and Bronze Age art relics, as well as royal objects dating to the 6th century AD. The city also houses the National Library and the Thai National Documentation Department. Jim Thompson's Thai House, named for a U.S. entrepreneur and devotee of Thai culture, is composed of several traditional Thai mansions; it contains the country's largest collection of 17th-century Thai religious paintings. There are also collections of Dvaravati and Khmer sculpture, in addition to examples of Thai and Chinese pottery and porcelain.

All of the country's daily newspapers and most of its weeklies and monthlies are published in Bangkok. Newspapers are printed in Thai, English, and Chinese.

Radio and television are controlled by government agencies and by the military. Most of the nation's radio stations and all of its television stations are located in or near Bangkok. Most programs are in Thai, but some special programs are in English and Chinese.

Motion pictures are extremely popular. There is a thriving Thai cinema industry, but films are also imported. Fairs, festivals, and "kite-fighting" contests are held in the parks; the Ratchadamnoen and Lumphini stadiums present exhibitions of Thai boxing. Silapakorn National Theatre presents dancing, drama, and music.

History

Bangkok became the capital of Siam (as Thailand was previously known) in 1782, when General Chao Phraya Chakkri, the founder of the ruling Chakkri dynasty, assumed the throne as Rama I and moved the court from the west to the east bank of the Chao Phraya River. The move appears to have been dictated by strategic considerations: the wide westward bend in the river constituted a wide moat guarding the northern, western, and southern perimeters of the new site. To the east stretched a vast, swampy delta called the Sea of Mud, which could be traversed only with extreme difficulty. Rama I modeled the new city on the former capital, Ayutthaya, 40 miles to the north. By the end of his reign the city was established. The walled Grand Palace complex and Bangkok's oldest temple, Wat Po, were completed. A new city wall, perhaps the most imposing structure, skirted the river and Khlong Ong Ang to the east; it was four and a half miles long, 10 feet thick, and 13 feet high, and it had 63 gates and 15 forts. The area enclosed amounted to one and a half square miles.

Strategic location of the city

More wats were built during the reigns of Rama II (1809–24) and Rama III (1824–51). They served as schools, libraries, hospitals, and recreation areas, as well as religious centres. During these years Wat Arun, noted for its tall spire, Wat Yan Nawa, and Wat Bowon Niwet were completed, Wat Po was further enlarged, and Wat Sutat was begun. There were, however, few other substantial buildings and fewer paved streets; the river and the network of interconnected canals served as roadways.

Rama IV (1851–68) developed the city while continuing, at a reduced rate, the traditional building of wats. The Grand Palace was improved, a number of substantial dwellings were constructed for members of the royal family, several new streets were laid down, and a reduction was made in the large number of floating houses anchored along the river front. A new route, Charoen Krung (New Road), leading southward was constructed, and a new city moat, Khlong Phadung Krung Kasem, parallel to the city's first canal, was dug and fortified; a long canal led from it to the present port area (Khlong Toei), thus allowing small boats to bypass the large bend in the river immediately south of the city. A pony path, now Phra Ram Thi 4 Road, was laid atop the mud heaped up beside this waterway.

The transformation of the city under Rama V

During the long reign of Rama V, King Chulalongkorn (1868–1910), the city was transformed through a program of public works. The great triple-spired Chakkri Building in the Grand Palace was completed by 1880; the Dusit Palace and an ancillary garden city were later built beyond the wall, being connected to the Grand Palace by the European-inspired Ratchadamnoen Nok Road. A road- and bridge-building program was embarked on in earnest, because King Chulalongkorn, an early automobile enthusiast, foresaw the effect that the motor vehicle would have on city development. Most of the now obsolete city wall was pulled down to build the roads, but two forts, a large gate, and a section of the wall were preserved. The centenary of the city, in 1882, was marked by the inauguration of many social reforms, manifested in the public buildings used for their administration, as well as by the completion of the great royal temple, Wat Phra Kaeo, which housed the Emerald Buddha. A post and telegraph service was organized in the 1880s, an electric tram service was instituted on Charoen Krung in 1892, and the first line of the State Railway, running from Bangkok to Phra Nakhon Si Ayutthaya, opened in 1900. Nor were aesthetic considerations forgotten, for other new buildings included the marble temple of Wat Benchamabopit (1900), elegant bridges in the French style, and the Italian-inspired National Assembly Hall (Throne Hall).

Rama VI (1910–25) continued the program of public works. He established Chulalongkorn University in 1916, built a system of locks to control the level of waterways throughout the city, and gave the public its first and largest recreational area—Lumphini Park. During Rama VII's reign (1925–35) municipal areas were delimited as part of a general administrative reorganization aimed at decentralization. In 1937 Bangkok was formally divided into the municipalities of Krung Thep and Thon Buri. At the time of their establishment, the two municipalities, approximately equal in area, together covered about 37 square miles; about four-fifths of the city's population lived in Krung Thep.

Since World War II Bangkok has grown with unprecedented rapidity. As a result of this growth, problems with transportation, communication, housing, water supply, drainage, and pollution have become acute. That those responsible for modernizing the metropolis are coping with these problems suggests the appropriateness of its official emblem: the God Indra seated atop a sacred white elephant, the four tusks of which denote its celestial status and its ability to accomplish the impossible. The city's uniquely Thai character, while perhaps diminishing, provides a vibrant backdrop for Bangkok's increasingly cosmopolitan image.

BIBLIOGRAPHY. WILLIAM WARREN and MARC RIBOUD, *Bangkok* (1972), gives a panorama of life in the city through imaginative text and evocative photographs. ERIK SEIDENFADEN, *Guide to Bangkok, with Notes on Siam,* 2nd ed. (1928, reprinted 1985), is a beautifully illustrated guide with a wealth of historical data. Three useful works are LARRY STERNSTEIN, *Planning the Developing Primate City: Bangkok 2000* (1971), a translation and critique of three Thai plans, *Thailand: The Environment of Modernisation* (1976), a profusely illustrated study that includes a long section on Bangkok, and *Portrait of Bangkok* (1982), a series of 11 essays, with rare pictures and maps, on historical and contemporary affairs, published on the occasion of the bicentenary of the city.

(La.S.)

Bangladesh

Bangladesh (in full People's Republic of Bangladesh; Bengali: Gana Prajātantrī Bangladesh) is an independent Asian state located in the delta of the Ganges and Brahmaputra rivers in the northeastern part of the Indian subcontinent. As the eastern portion of the historic region of Bengal, it formed, with the Indian state of West Bengal, the province of Bengal in British India. From the partition of 1947 until 1971 it was, as East Pakistan, one of five provinces of Pakistan, separated from the other four by 1,100 miles (1,800 kilometres) of Indian territory.

Bangladesh has an area of 55,598 square miles (143,998 square kilometres) and is one of the most densely populated areas in the world. It is bounded by the Indian states of West Bengal to the west and north, Assam to the north, Meghālaya to the north and northeast, and Tripura and Mizorām to the east, by Burma to the southeast, and by the Bay of Bengal to the south. The capital is Dhākā (formerly spelled Dacca).

This article is divided into the following sections:

Physical and human geography

THE LAND

Relief. Bangladesh constitutes the eastern two-thirds of the Ganges-Brahmaputra deltaic plain, which stretches northward from the Bay of Bengal. Except for small higher areas of jungle-covered old alluvium (rising to about 100 feet [30 metres]) in the northwest and north-centre—called, respectively, the Bārind and Madhupur tracts—the plain is a flat surface of recent alluvium, having a gentle slope and generally with an elevation of less than 30 feet above sea level. In the northeast and southeast the alluvial plains—called, respectively, the Sylhet and Chittagong hills—give place to ridges running mainly north–south, which form part of the mountain divide with Burma and India. Bangladesh is fringed on the south by the Sundarbans, a huge expanse of marshy deltaic forest.

In northwestern Bangladesh the Bārind Tract comprises a triangular wedge of land between the floodplains of the Ganges and Brahmaputra rivers. The soil of this region is hard, reddish clay, and the region is comparatively elevated. A depression called the Bhar Basin extends southeast of the Bārind Tract for about 100 miles between the floodplains of the Ganges and Brahmaputra rivers to their confluence. This area is inundated during the summer monsoon season. The drainage of the western part of the basin is centred in the vast marshy area called the Chalan wetlands, also known as Chalan Lake. The floodplains of the Brahmaputra (Jamuna), which lie north of the Bhar Basin and east of the Bārind Tract, stretch from the border with India (Assam) in the north to the confluence of the Ganges and Brahmaputra in the south. The Brahmaputra frequently overflows its banks in devastating floods. South of the Bhar Basin is the floodplain of the Ganges.

In north-central Bangladesh, east of the Brahmaputra floodplains, is the Madhupur Tract. It consists of an elevated plateau, with hillocks varying in height from 30 to 60 feet, and cultivated valleys. The Madhupur Tract contains *sal* trees, whose hardwood is comparable in value and utility to teak. East of the Madhupur Tract, in northeastern Bangladesh, is a region called the Northeastern Lowland. It encompasses the southern and southwestern parts of the Sylhet area (including the valley plain of the Surma River) and the northern part of the Mymensingh area and has a large number of lakes. The Sylhet Hills in the far northeast of the region consist of a number of hillocks and hills ranging from 100 feet to more than 1,100 feet in height.

In east-central Bangladesh the Brahmaputra River in its old course built up the Meghna Flood Basin, which includes the low and fertile Meghna-Lakhya Doāb (the land area between those rivers). This area is enriched by the Titās distributary, as well as the land areas formed and changed by the deposition of silt and sand in the riverbeds of the Meghna River, especially between Bhairab Bazar and Daudkandi. Dhākā is located in this region.

In southern Bangladesh the Central Delta Basins include the extensive lakes in the central part of the Bengal Delta, to the south of the Ganges (Padma). The basin's total area is about 1,200 square miles. The belt of land in southwestern Bangladesh bordering the Bay of Bengal constitutes the Immature Delta. The belt—a lowland of some 3,000 square miles—contains, in addition to the vast mangrove forest known as the Sundarbans, the reclaimed and cultivated lands to the north of it. The area nearest the Bay of Bengal is crisscrossed by a network of streams that flow around roughly oblong islands. The Active Delta, located north of the Central Delta Basins and east of the Immature Delta, includes the Dhaleswari-Padma Doāb and the estuarine islands of varying sizes that are found from the Pusur River in the southwest to the island of Sandwīp near Chittagong in the southeast.

Lying to the south of the Feni River in southeastern Bangladesh, the Chittagong region has many hills, hillocks, valleys, and forests and is quite different in aspect from other parts of the country. The coastal plain is partly sandy and partly composed of saline clay; it extends southward from the Feni River to the town of Cox's Bāzār and varies in width from one to 10 miles. The region has a number of offshore islands and one coral reef, St. Martin's, off the coast of Burma. The hilly area known as the Chittagong Hill Tracts in the far southeast consists of low hills of soft rocks, mainly clay and shale. The north–south ranges are generally below 2,000 feet in height.

Drainage. The rivers of Bangladesh have molded not only its physiography but also the way of life of the people. They may be divided into five systems—(1) The Ganges, or Padma, as the united streams of the Ganges and Brahmaputra are known, and their deltaic streams; (2) the Meghna and the Surma river system; (3) the Brahmaputra and its adjoining channels; (4) the North Bengal rivers;

and (5) the rivers of the Chittagong Hill Tracts and the adjoining plains.

The Ganges is the pivot of the deltaic river system of Bengal. The river and its tributaries enclose a large area of southwestern Bangladesh. The Ganges Delta itself covers about 20,000 square miles. The Ganges enters Bangladesh from the west and forms, for about 90 miles, the boundary between Bangladesh and West Bengal (India). It flows southeast to join the Brahmaputra and forms numerous distributaries and spill channels. Except at places where it is confined by high banks, the main channel of the river changes its course every two or three years; consequently no description of Bangladesh's topography retains its absolute accuracy for long. The river carries an immense volume of water mixed with silt, which gives it a muddy look. Silt deposits build temporary islands that reduce navigability but are so highly fertile that they have been for decades a source of feuds among peasants.

The Meghna is formed by the union of the Sylhet-Surma and Kusiyāra rivers. These two rivers are branches of the Barāk River, which rises in the Nagar-Manipur watershed in India. The main branch of the Barāk, the Surma, is joined near Azmiriganj in northeastern Bangladesh by the Kālni and farther down by the Kusiyāra branch. The Dhaleswari, a distributary of the Jamuna River, joins the Meghna a few miles above the junction of the Ganges and the Meghna. As it meanders south, the Meghna grows larger after receiving the waters of a number of rivers, including the Buriganga and the Sitallakhya.

The Brahmaputra and its adjoining channels cover a large area from north-central Bangladesh to the Meghna River in the southeast. The Brahmaputra receives waters from a number of rivers, especially on its right bank. The river, with its notoriously shifting channels, not only prevents permanent settlement along its banks but also inhibits communication between the northern area of Bangladesh and the eastern part, where Dhākā is situated.

The Tista is the most important water carrier of northwestern Bangladesh. Rising in the Himalayas near Darjeeling (India), it flows southward. After the floods of 1787, however, the Tista changed its course, moving southeastward to join the Brahmaputra. A number of small and medium-sized rivers in the southwest are silting up, adversely affecting the economic life of that region.

Four main rivers constitute the river system of the Chittagong Hills and the adjoining plains—the Feni, the Karnaphuli, the Sangu, and the Mātāmuhari. Flowing generally west and southwest across the coastal plain, they empty into the Bay of Bengal. Of these rivers the longest is the Karnaphuli. It is dammed at Kaptai, about 30 miles upstream from its mouth near the city of Chittagong.

None of the major rivers of Bangladesh originates within the country's territory. The headwaters of the Surma are in India; the Ganges rises in Nepal and the Brahmaputra in China (Tibet), but they, too, reach Bangladesh across Indian territory. Thus, Bangladesh lacks full control over the flow of any of the streams that irrigate it. The construction of a barrage upstream at Farakka in West Bengal has led to the diversion of a considerable volume of water from the Ganges, and the flow to western Bangladesh is insufficient in the dry season from November to April. The equitable distribution of the river's waters has been since the 1970s a source of friction between India and Bangladesh.

Each year between June and October the rivers overflow their banks, rising most heavily in September or October and receding quickly in November. The inundations are both a blessing and a curse. Without them, the fertile silt deposits would not be replenished. But severe floods regularly damage crops and ruin hamlets and sometimes take a heavy toll on human and animal populations.

Three main types of soils

Soils. There are three main categories of soils: the old alluvial soils, the recent alluvial soils, and the hill soils, which have a base of sandstone and shale. The fertile recent alluvial soils, found mainly in flooded areas, are usually pale brown, sandy, micaceous, and chalky clays and loams. They are deficient in phosphoric acid, nitrogen, and humus but not in potash and lime. The old alluvial soils in the Bārind and Madhupur jungles are dark-brown clays and loams. They are sticky during the rainy season and hard during the dry. The hill soils are permeable and can support dense forest growth.

Climate. Bangladesh has a typical monsoon climate characterized by rain-bearing winds, moderately warm temperatures, and high humidity. In general, maximum temperatures in the summer months, from April to September, range between 91° and 96° F (33° and 36° C). April is the warmest month in most parts. January is the coolest month in the winter season, which lasts from about November to March.

The conditions of lowest atmospheric pressure occur in Bangladesh in June and July, the storm season. Winds are mostly from the north and northeast in winter, blowing at a rate of one to two miles per hour in northern and central areas and two to four miles per hour near the coast. During the period of the northwesters (March to May), wind speeds may rise to 30 or 40 miles per hour.

Rainfall

Bangladesh receives heavy rainfall; except for some parts in the west, it generally exceeds 60 inches (1,500 millimetres) annually. Large areas of the south, southeast, north, and northeast receive from 80 to 100 inches, and the northern and northwestern parts of the Sylhet area receive from 150 to 200 inches. The maximum rainfall occurs during the monsoon period, from June to September or early October.

In the early summer (April and May) and late in the monsoon season (September to November), storms of very high intensity often occur; they may create winds with speeds of more than 100 miles per hour, piling up the waters of the Bay of Bengal to crests as high as 20 feet that crash with tremendous force onto the coastal areas and the offshore islands, inundating them and causing heavy losses of life and property. Since the early 18th century, when records were first kept, more than 1,000,000 people have been killed in such storms, 815,000 of them in three storms occurring in 1737, 1876, and 1970. Another severe storm occurred in May 1985.

Plant and animal life. Bangladesh in general possesses a luxuriant vegetation, with villages appearing to be virtually buried in groves of mango, jackfruit, bamboo, betel nut, coconut, and date palm. About 15 percent of the country's land surface is covered with forests.

Bangladesh has four different areas of vegetation. The eastern zone, consisting of parts of the Sylhet and Chittagong areas, has many low hills covered with jungles of bamboo and rattan. The central zone, covering parts of the country extending north of Dhākā, contains a large number of lakes and swampy vegetation, as well as the Madhupur jungles. The area lying to the northwest of the Brahmaputra and to the southwest of the Padma forms a flat plain, the vegetation of which consists mostly of cultivated plants and orchards. Babul (*Acacia arabica*) is the most conspicuous plant. The southern zone along the Bay of Bengal contains the Sundarbans, with their distinctive mangrove vegetation. Many commercially valuable trees, such as the sundri, for which the Sundarbans are named (*Heritiera fomes* or *minor*); *gewa,* or *gengwa* (*Excoecaria agallocha*), a softwood tree used for making newsprint; and goran (*Ceriops roxburghiana*), a type of mangrove, grow in this vast forest.

Elephants

Elephants are found in the Chittagong Hill Tracts and northeastern Sylhet. The domesticated, or water, buffalo are used for plowing and pulling carts. The barking deer, the *barosinga* (or 12-horned deer), and the sambar deer, with its maned neck, are well-known. Of the carnivores, the royal Bengal tiger is the best known. The clouded leopard, dark gray and with spots that are oval or oblong in form, is smaller than the leopard. The ferocious leopard cat is about the size of the domestic cat but with longer legs.

There are three types of bear: the sloth bear, the Himalayan black bear, and the Malayan sun bear. The sloth bear is the most numerous. The jackal is a common animal, as is the mongoose. The Bengal, or rhesus, monkey is the most common primate.

The common house crow is seen everywhere. The bulbul, the magpie robin, and a wide variety of warblers are also found; some are migrants that appear only in winter.

Other species include the common game birds, parakeets, cuckoos, hawks, owls, kingfishers, hornbills, woodpeckers, and vultures. Among the eagles, the crested serpent eagle and the ring-tailed fishing eagle are the most common. There are also hoopoes, herons, storks, ducks, and wild geese.

Settlement patterns. The extremely high population density of Bangladesh, averaging 1,900 persons per square mile, varies widely according to the distribution of flat land. The highest density, over 2,800 persons per square mile, occurs in and around the capital, which is also the centre of the most fertile zone; the lowest, at just over 100 persons per square mile, occurs in the hills of Chittagong.

Population density

Rural settlement. The rural area throughout Bangladesh is very densely settled. The inundation of most of the fields during the rainy season makes it necessary to build houses on higher ground. Continuous strings of settlements along roads are common in areas south of the Ganges and in the floodplains of the Mahānanda, Tista, Jamuna, Ganges, and Meghna rivers. Similar settlements are also found in the hilly regions of southern Sylhet and in the Chittagong region. In central and western Sylhet and in the Chittagong Hill Tracts, settlements occur in a nucleated, or clustered, pattern. The traditional character of rural villages has changed with the addition of prefabricated one- or two-storied structures scattered among the thatched bamboo huts.

Urban settlement. Although industrial development has prompted migration to the cities, Bangladesh is one of the least urbanized areas in South Asia. Eighty percent of the population lives in villages. There are only three major cities: Dhākā, Chittagong, and Khulna. Dhākā, the capital, is the largest. Chittagong, the country's major port, is second in importance. A number of industrial areas, such as Kalurghat, Sholāshahar, and Faujdār Hāt, have developed around Chittagong. Khulna, in the southwest, has become a commercial and industrial centre; the opening of the port of Chālna nearby and the growth of the Daulatpur industrial area has increased its population.

THE PEOPLE

Ethnic composition and distribution. The proto-Australoids, sometimes called Veddas, were one of the earliest groups to enter the area. According to some ethnologists, they were followed by Mediterranean Caucasoids (whites), also known as Aryans. Armenoids (of Indo-European stock) are believed to have entered as well. With the coming of the Muslims in the 8th century AD, new elements were introduced. Persons of Arab, Persian, and Turkish origin moved in large numbers to the subcontinent. By the end of the 12th century they had entered what is now Bangladesh. A substantial proportion of the population is descended from these Muslim immigrants.

Most of the tribal peoples of Bangladesh inhabit the Chittagong Hill Tracts in the southeast. They are predominantly Buddhist, and some of the tribes are related to the peoples of Burma. Of the approximately 12 ethnolinguistic groups of the Chittagong Hill Tracts, the four largest are the Chakmā, the Marma, the Tripura (Tipra), and the Mro. Tribal peoples in other parts of Bangladesh include the Santāls, of the proto-Australoid group, the Khāsis, the Gāro, and the Hajang. The Santāls live in the northwestern part of Bangladesh, the Khāsis in Sylhet in the Khāsi Hills near the border with Assam, and the Gāro and Hajang in the northeastern part of the country.

The Bengalis

Apart from these tribes, the rest of the people are Bengalis—an ethnic as well as a linguistic group. The Bengalis, however, are not homogeneous in origin. In general, the people of the coastal areas with whom the Muslim merchants of the Middle East were in close touch show physical features that seem to be the result of the admixture of local people with those of Turkish and Semitic origin.

Linguistic composition. Bengali, the language spoken in Bangladesh, belongs to the Indo-Aryan group of languages and is related to Sanskrit, not directly but by way of Gaudiya Prākrit, the language from which it is derived. Bengali is the mother tongue of about 98 percent of the people. Tribal peoples have their own distinct dialects, some of which are related to the Tibeto-Burman group of languages. English is spoken in urban centres and among educated groups.

Bengali has two distinct styles: *sādhu bhāṣā,* the literary language, contains many words derived from Sanskrit, and *calit bhāṣā,* the colloquial language. Until the 1930s *sādhu bhāṣā* was the standard medium of formal writing, but *calit bhāṣā* is now the basic form. There are a number of dialects. Bengali contains a large number of loanwords from Portuguese, English, Arabic, Persian, and Hindi.

Religions. More than 85 percent of the population follows the religion of Islām. Most Muslims belong to the Sunnī sect, but there are a small number of Shī'ite Muslims, mostly descendants of immigrants from Iran. Hindus, who constitute about 10 percent of the population, are divided into scheduled (low) and nonscheduled castes. Buddhists form less than 1% percent of the population. Of the tribes in the Chittagong Hill Tracts, the Chakmā,

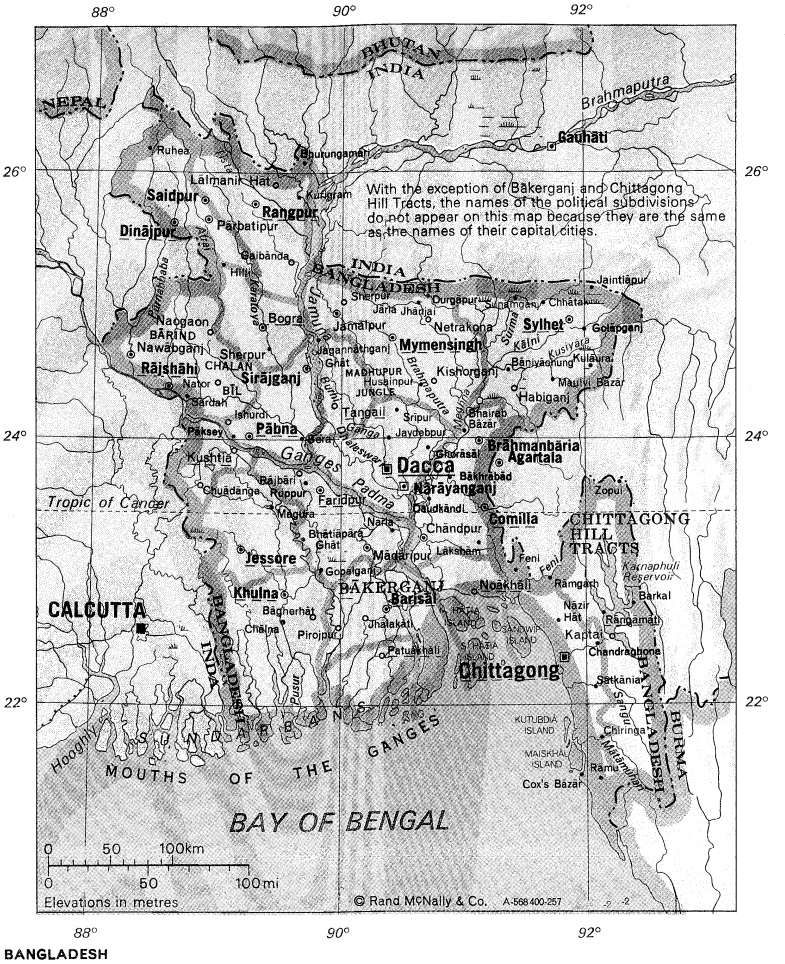
MAP INDEX

Political subdivisions

The name of a political subdivision if not shown on the map is the same as that of its capital city.

Cities and towns

Physical features and points of interest



BANGLADESH

Marma, and Mro are mostly Buddhists. The Kuki, Khomoi (Kumi), and some of the Mro are animists. While most of the Lushai are Christians, the Tripura are Hindus.

Demographic trends. Almost half of Bangladesh's population is under 15 years of age; the birth rate is high, and average life expectancy is about 50 years. The rate of infant mortality remains high. There has been almost no immigration since the 1970s. A relatively small number of Bangladeshis work in Britain and in Middle Eastern countries, and there has been a steady emigration of farm labourers into neighbouring Assam.

THE ECONOMY

Agriculture. Bangladesh is overwhelmingly agricultural, with some three-fifths of the population engaged in farming. Jute and tea, which are principal sources of foreign exchange, follow rice as the most important agricultural products. The country produces about one-fifth of the world's supply of raw jute. Other important agricultural products are wheat, pulses (leguminous plants, such as peas, beans, and lentils), sweet potatoes, oilseeds of various kinds, sugarcane, tobacco, and fruits such as bananas, mangoes, and pineapples.

Jute and rice production

Agriculture has in the past been wholly dependent upon the vagaries of the monsoon. A poor monsoon has always meant poor harvests or none and the threat of famine. Among the remedial measures adopted has been the construction of a number of irrigation projects designed to control floods and to conserve rainwater for use in the dry months. The most important are the Karnaphuli Multipurpose Project in the southeast, the Tista Barrage Project in the north, and the Ganges-Kabadak Project, to serve the southwestern part of the country. Economic planning has encouraged double and triple cropping, intercropping, and the increased use of fertilizers.

Irrigation and flood control

Fisheries. The rivers of Bangladesh are suitable for the breeding and raising of fish. Its rivers and seacoast offer opportunities for the operation of the usual types of fisheries, mostly in the estuaries of the Bay of Bengal. Among the varieties of fish caught are the marine *rupchanda,* or pomfret, and the freshwater hilsa, a relative of the shad.

Industry. The excessive—until recently almost exclusive—dependence upon agriculture leads to seasonal unemployment among peasants, as well as to a low standard of living. To counteract this imbalance, a policy of industrialization was adopted after 1947 and was pursued through five-year plans. The main obstacle to its fulfillment has been the comparative lack of mineral resources.

Power resources. Oil in marketable quantities has not been struck anywhere in Bangladesh. The country's first oil well, near Sylhet, was discovered in 1986. Natural gas is used mainly in the manufacture of fertilizer and for thermal power. More than half the proven gas reserves are in the Comilla area, and nearly all the rest in Sylhet.

Some deposits of coal have been found in northwestern Bangladesh in the Rājshāhi area. The thickest seams are located at relatively inaccessible depths of 3,000 to 3,500 feet. Smaller deposits of coal exist in northwestern Sylhet. The Chittagong Hill Tracts contain some brown coal and lignite. Peat deposits exist in several places, but some of the beds remain under water for half the year, making extraction difficult. Limestone is found in the Sylhet and Chittagong areas. Radioactive minerals have been detected in sand deposits along the beaches south of Cox's Bāzār.

Bangladesh's electricity is produced by thermal and hydroelectric processes. The main source of hydroelectricity is the Kaptai Dam in the Chittagong Hill Tracts.

Industrial development. Industrial policy between 1947 and 1971 was to give priority to industries based on indigenous raw materials such as jute, cotton, hides, and skins. The principle of free enterprise in the private sector was accepted, subject to certain conditions, which included the national ownership of public utilities. The policy also aimed to develop as quickly as possible consumer-goods

industries with a view to avoiding dependence on imports.

Autonomous corporations

Under Pakistani administration, new types of autonomous corporations were established to deal with industrial development, electricity, water and sewerage management, the development of forest industries, and road transportation. The Bangladesh government in 1972 nationalized these corporations and then established several new corporations to manage the nationalized enterprises. But hasty change, coupled with the inexperience of those placed in charge of the corporations, produced widespread disruptions, and industrial production came almost to a halt. The policy of nationalization was gradually revised and was replaced by a 19-point program announced in 1979. This program emphasized greater productivity and efficiency. In an attempt to encourage private investment, the government also returned many state-owned enterprises to the private sector.

Manufacturing and other industries. Because the export of raw jute is not highly remunerative, efforts were made between 1947 and 1971 to establish mills to produce and export jute products and thus earn foreign exchange. About 45 percent of the jute produced during that period was processed in the territory; the balance was exported raw. Next to jute, Bangladesh's main exports are tea and hides and skins. Among the minor exports are newsprint, ready-to-wear garments, shrimp, and frogs' legs.

The bamboo in the Chittagong Hill Tracts and the various softwood trees growing in the Sundarbans provide excellent raw material for papermaking. There are paper mills at Chandraghona, Chhātak, and at Pāksey, as well as a paper and board mill at Khulna.

Bangladesh has fertilizer factories, textile mills, sugar factories, glass works, and aluminum works. Its two cement factories, located at Chhātak, in the Sylhet area, are unable to meet the growing demand for cement. A shipyard has been opened at Khulna for repairing and reconstructing ships, and a steel mill is located at Chittagong.

Cottage industries

By far the most important cottage industry centres on the production of yarn and textile fabrics—mostly coarse and medium-quality fabrics. Carpets, ceramics, and cane furniture are also products of cottage industries.

Transportation. Central to the country's transportation system are its networks of waterways, roads, and railways, the latter mostly built during British rule. Inland waterways are of major importance, providing low-cost transport and access to areas where land transport cannot be economically developed. They carry most of the domestic and foreign cargo. The chief seaports are Chittagong and Chalna, and there are international airports at Dhākā and Chittagong.

The forms of transport used on Bangladesh's roads range from automobiles and buses to the bullock cart. Two-wheeled horse-drawn jigs and buffalo carts are still used in the north in Rājshāhi. Town- and city-dwellers both rely largely on the cycle rickshaw and two types of motorized transport, known locally as auto and tempo, both of which are three-wheeled.

ADMINISTRATION AND SOCIAL CONDITIONS

Government. The constitution of 1972 specified a parliamentary form of government under a prime minister and a president elected by a national assembly. In 1975, however, a military coup led to a regime of martial law. The form of government that obtained thereafter was a mixture of presidential and parliamentary systems, effective power remaining with the army. Following another coup in 1982, the constitution was suspended and the country placed under martial law. In 1986 martial law was lifted and parliamentary elections were held, but in 1987, following a series of strikes and riots, the government dissolved the parliament. A new parliament was elected in 1988.

A large-scale administrative reorganization was carried out in the 1980s. While the revenue divisions remained the same—namely, Dhākā, Chittagong, Rājshāhi, and Khulna—the older districts were subdivided and each subdivision raised to the status of a district. A new administrative unit, called *upazilla,* or subdivision, was created to facilitate decentralization of power. The *upazillas* are headed by an executive officer who has administrative and judicial functions.

Bangladesh has continued with substantially the same judicial system as had been in operation when the territory was a province of Pakistan, and which owed its origins to the system in operation under the British raj.

The 1972 constitution divided the Supreme Court of Bangladesh into Appellate and High Court divisions and mandated a complete separation of the judiciary and executive branches. During the subsequent authoritarian regime, however, the power of the Supreme Court was greatly reduced. In 1977 a Supreme Judicial Council was established to draw up a code of conduct for Supreme Court and High Court judges, who may be removed from office by the president upon the council's recommendation. The fragmentation of the High Court into five divisions located in different parts of the country—which had been decreed by the military in the 1970s—was rescinded in 1986. Provision was made, however, for the judges to go on circuit for part of the year to hear cases in other parts of the country.

Education. The foundation of the educational system in Bangladesh was laid down during the period of British rule; the system has three levels—primary, secondary, and higher education. Primary education, which is free but not compulsory, is for children up to about 10 years old. Only about half of all children attend primary school. Secondary education is divided into three levels—junior secondary, high school, and higher secondary (intermediate college)—with public examinations being held at the conclusion of each level of schooling. Schools in cities and towns are generally better staffed and financed than those in rural areas.

Higher education

There are more than 600 colleges, most of them affiliated with the University of Dhākā, the University of Rājshāhi, or the University of Chittagong. Other institutions include Jahangirnagar University on the outskirts of the capital, the Bangladesh Agricultural University at Mymensingh, the Bangladesh University of Engineering and Technology at Dhākā, and the Islāmic University at Tongi. Medical education is provided by several medical colleges and an institute of postgraduate medicine at Dhākā.

For vocational training Bangladesh relies on several engineering colleges and a network of polytechnics and law colleges. In addition, there exist a college of arts and crafts, an agricultural college, a college of home economics for women, and an institute of social welfare and research.

The demand for higher education has continued to rise. One of the problems that has continued to impede educational progress is student unrest.

Health and welfare. Malaria, cholera, and tuberculosis are the most serious threats to health. An effective approach to the treatment of cholera and tuberculosis has been developed by research laboratories and hospitals in Dhākā and Comilla. The incidence of malaria has been reduced by a malaria eradication program in which swamps and marshes are regularly sprayed with insecticides. A family planning program has also been introduced.

Social services are provided by private agencies and government departments. These services include community development projects, schools for handicapped children, youth centres, orphanages, and training institutes for social workers.

CULTURAL LIFE

The Bengali language, Islāmic religion, and rural character of Bangladesh all serve to unify the country's culture, although variations occur among ethnic, religious, and social minorities and in the urban centres.

Daily life. The typical household in Bangladesh, particularly in the villages, includes several generations of extended family. Most marriages are arranged by parents or other relatives, but educated men and women have tended increasingly to choose their own partners. Divorce is permissible among Muslims. Hindu marriage is sacramental, but a Hindu can obtain a separation by application to a court of law. Muslim law permits limited polygamy.

Religious festivals

The main festivals in Bangladesh are religious. The two most important are ʿId al-Fiṭr, which comes at the end

of Ramaḍān, the Muslim month of fasting, and ʿId al-Aḍḥā, or the festival of sacrifice, which follows two and a half months later. On both occasions families and friends exchange visits.

While rice, pulses, and fish continue to constitute the staple diet of Bangladeshis, shortages of rice since World War II have forced the acceptance of wheat and wheat products as alternatives. Meat, including goat and beef, is also eaten, especially in the towns.

The arts. There are four main types of music in Bangladesh—classical, light-classical, devotional, and popular. Classical music has many forms, of which *dhrupad*—Hindustani devotional songs—and *khayal*—a blending of the Perso-Arab and Indian musical systems—are the best known. Devotional music also is represented by *qawwāli* and *kīrtan,* forms that are part of the common musical heritage of the subcontinent. It is, however, in the field of popular music that Bangladesh can best claim originality. The forms known as *bhatiali, bhawaiya, jari, sari, marfati,* and *baul* have no exact equivalents outside the country. While they may lack the sophistication and artistry of the classical forms, they are characterized by a spontaneity and vigour missing in classical music.

Apart from such classical dances as kathakali and bharata natya—forms that are popular throughout the subcontinent—Bangladesh has evolved highly original indigenous dances. The best known are the *dhāli, baul,* Manipuri, and snake dances. Each form expresses a particular aspect of tribal or communal life and is danced on specific occasions. The best known among the academies devoted to music and dancing are the Bulbul Academy and the Nazrul Academy.

Painting in Bangladesh is a recently introduced art form. The main figure behind the art movement was Zainul Abedin, whose sketches of the Bengal famine of 1943 first attracted attention. He was able, after 1947, to gather around him a school of artists who experimented with various forms, both orthodox and original.

Traditional architecture in Bangladesh is represented by the many mosques, mausoleums, forts, and gateways that have survived from the Mughal period. These, like Muslim architecture elsewhere in the subcontinent, are characterized by the pointed arch, the dome, and the minaret. Although in style and conception traditional architecture belongs to the same school as medieval buildings in northern India, Bangladesh's distinctive contribution lay in the translation into brick and mortar of the sloping, four-sided, thatched roof found in the countryside.

Recreation. Association football (soccer) has in the course of the 20th century supplanted practically all traditional sports. Field hockey, cricket, tennis, badminton, and wrestling are also practiced. The best known of the indigenous games is *ha-do-do.* The rules require each team in turn to send out a player to raid the other's territory. The raider must, while chanting, touch as many opposing players as he can without taking a breath. Kite-flying is another traditional pastime enjoyed by young and old alike. The making of elaborate kites from cloth or paper is a distinctive visual folk art as well.

The game of *ha-do-do*

All towns and most villages have cinema houses. Plays are occasionally staged by amateur groups and drama societies in educational institutions and are broadcast regularly on radio and television. Musical concerts, though not as popular as the cinema, are well-attended. Especially popular in the countryside is *jatra,* a rudimentary form of opera that draws on local legends.

Press and broadcasting. Programs are broadcast on radio and television in English and Bengali; news on radio is also broadcast in Urdu, Hindi, Burmese, and Arabic. The Bengali newspapers have relatively small circulations, a fact that reflects the low level of literacy in the country. But for each reader there are commonly five or six listeners, sometimes more, so that the influence of the press on opinion is greater than the sales suggest. Although the circulation of the English dailies is even less, because their patrons are the educated classes they exercise a disproportionate influence. Both radio and television are controlled by the government. The majority of newspapers are privately owned, and the press is relatively free. (S.S.H.)

For statistical data on the land and people of Bangladesh, see the *Britannica World Data* section in the BRITANNICA WORLD DATA ANNUAL.

History

Bangladesh ("Land of Bengal," or "of the Bengalis") has existed as an independent state only since 1971, yet its national character dates to the ancient past (see also the history sections of the articles INDIA and PAKISTAN). This identity consists in three distinctive attributes—a land, a language, and a religion.

The land is shaped by the two great rivers Ganges and Brahmaputra, which join in central Bangladesh to become the Padma. They are the greatest of a series of rivers winding down to the Bay of Bengal. This region has always been isolated from the north Indian plain. In early times eastern Bengal was called Vaṅga, while western Bengal was known as Gauḍa.

The Bengali language began to assume a distinct form in the 7th century AD and by the 11th century had acquired its own literature. The "Bengali Renaissance" of the 19th century was centred in Calcutta, and its greatest figure was the poet Rabindranath Tagore. Almost all of the movement's literary and artistic celebrities were Hindus.

The Buddhism that under the Mauryan emperor Aśoka's patronage spread across the whole subcontinent in the 3rd century BC was driven out after the decline of Maurya power, as Brahmanical Hinduism reestablished its hold. However, in remote eastern Bengal Buddhism lingered on under the Pāla kings (8th–12th century) until their overthrow by the Vaishnavite Hindu Senas. The Senas encouraged the settlement of high-caste Hindus as lords of the land, but this did not greatly affect the general populace. Then, in about AD 1200, Muslim invaders from the northwest overthrew the Senas. Islām found a mass following among the Vaṅga people. In the eastern part of the country—Noakhali, Chittagong, and Sylhet—Arab traders also spread Islāmic teaching. Whereas in northern India the strength of caste Hinduism was enough to withstand centuries of Muslim dominance, culminating in the Mughal dynasty (16th–18th century), in eastern Bengal, Islām became the religion of the majority.

Introduction of Islām

As Mughal authority declined, the Suba, or Dominion, of Bengal—including Bihār and Orissa—became semi-independent. The threat to the Muslim rulers of the Suba came first from the east—from Arakanese pirates and Portuguese raiders. The capital was moved to Dhākā in 1608. When further invasion threatened from central India, from the rising power of the Marāṭhā kingdom, the capital was shifted to Murshidābād in 1704. It was during this period that the English East India Company established its base at Calcutta. From 1757 the British were the dominant political power in Bengal.

Reluctant to become involved in Indian administration, the British confirmed the landed magnates, or zamindars, in their charge of vast estates. Some were Muslims (such as the Nawab of Dhākā), but most were Hindu rajas, even in eastern Bengal. They were required to collect revenue from the land, and they appointed agents to ensure regular collection. These agents formed the new middle class of Bengal, the *bhadralok* ("respectable people"). Mainly upper-caste Hindus, they collected the revenue from peasants, who were mainly Muslims. The *bhadralok* resided in Calcutta and the larger towns; in time they became the most active advocates of Indian self-government.

The province of Bengal was almost impossible to administer, even though Assam was made a separate province in 1874. In 1905, largely at the initiative of the viceroy, Lord Curzon, two new provinces were created: Western Bengal, with Bihār and Orissa, and Eastern Bengal and Assam. The division followed one of the branch rivers of the Ganges from Rājmahāl in the north to the sea. The division was on a geopolitical rather than an avowedly communal basis. It gave Eastern Bengal, with its capital at Dhākā, a population of 31,000,000, all but 6,000,000 being Bengalis. Behind Curzon's move, besides greater efficiency, was the intention of encouraging the Bengali Muslims as a counterweight to the "seditious" Bengali Hindus.

The partition of Bengal

The partition elicited vociferous protest in Western Bengal, especially in Calcutta. A prominent part was played by Tagore, whose family had vast holdings along the Padma. The campaign included a boycott of British manufactures under the slogan "swadeshī" ("of our own country," or "India-made goods"). The Muslim notables, still loyal to the British, decided that they also needed to organize. Their principal leaders were in northern India, but in December 1906 they gathered at Dhākā under the patronage of Nawab Salimullah and set up the All-India Muslim League. Their efforts secured separate electorates and separate constituencies for the Muslims under the 1909 Reforms, but the campaign against the partition of Bengal went on, and in 1912 the province was reunited (Bihār and Orissa being separated and Assam reverting to separate status).

All-India Muslim League

Despite the separate electorates, the Muslim League had no majority in any province. In reunited Bengal, where Muslims formed a majority of the population (33,000,000 in a total of 60,000,000), they received 117 seats in the Bengal Legislative Council numbering 250. It was necessary to adopt coalition tactics. The politician most adept at this was Fazl ul-Haq, chief minister of Bengal from 1937 to 1943. He set up his own Peasants and Tenants Party, but he was also active in the Muslim League from its inception. When in 1940 the Muslim League held its annual gathering at Lahore, Fazl ul-Haq proposed a resolution calling for "independent states" for the Muslims. The press labeled this the "Pakistan Resolution," but for Fazl ul-Haq and many others it implied a plurality of states. Distrusted by the influential Indian Muslim politician Mohammed Ali Jinnah (the first governor general of Pakistan [1947–48]), Fazl ul-Haq was expelled from the league. In his place Khwaja Nazimuddin became chief minister. Nazimuddin, a relative of the Nawab of Dhākā, was loyal to Jinnah but lacked political finesse. He was displaced in 1945 by the more sophisticated Hussein Shaheed Suhrawardy. Suhrawardy was the main architect of the Muslim League's success in Bengal in the 1946 election. He became chief minister of Bengal in 1946.

After protracted negotiations it became clear that the Congress Party (Indian National Congress) could not expect to preserve a united India. A major factor was the intense intercommunal conflict in August 1946 known as the "Great Calcutta Killing." On his arrival as the new viceroy the following year, Admiral Lord Mountbatten of Burma drafted a plan to partition the subcontinent. Suhrawardy met with Sarat Chandra Bose, the acknowledged Hindu political leader in Bengal, and the two agreed that they should claim a separate, independent united Bengal. Jinnah was prepared to agree, as was Mountbatten, but Mahatma Gandhi and the Congress Party refused. Hence, when partition came it was decided by religion rather than language.

The boundaries of East Pakistan, as the region became, were determined by Sir Cyril Radcliffe, chairman of the Boundary Commission, there being total disagreement among his Hindu and Muslim colleagues. The boundary he defined did not follow any clear natural feature, as in the 1905 partition, nor was it wholly based on communal proportions. Excluded wholly or partly from East Pakistan were Murshidābād, Nadia, Jessore, and Dināj pur, each approximately 60 percent Muslim. Included were Khulna (49 percent Muslim) and the Chittagong Hill Tracts, where Muslims formed only 3 percent of the population. In addition, following a plebiscite, the Sylhet area (61 percent Muslim), formerly a part of Assam province, and a small area of Cachar (38 percent) were included.

he new oundary

On both sides of the new boundary, those who believed themselves a threatened minority moved into what they perceived as a place of refuge. Along with Muslim Bengalis arriving from Hindu majority districts, there were many Muslims who came from Bihār. One district, Purnea, had an actual Muslim majority and had been claimed by Jinnah. About 1,000,000 Bihāris settled in the new state.

At independence, Suhrawardy lingered in Calcutta, and Nazimuddin became chief minister of East Pakistan. From the beginning, the link between the two parts of Pakistan was tenuous; indeed, their only common interest was fear of Indian domination. Jinnah and his advisers believed that unification might be achieved through a common language, Urdu, which was used in the army and administration. The Bengalis perceived this as a threat. Their other major grievance was that their export products, jute and tea, provided most of Pakistan's foreign exchange; yet the central government mainly stimulated development in the West.

he two akistans

The Bengalis began to feel that they had no real power in Pakistan. When Jinnah died, Nazimuddin became governor general; but when Liaquat Ali Khan, the prime minister, was shot in October 1951, Nazimuddin took over, installing a Punjabi, Ghulam Mohammad, as governor general. Although Nazimuddin had a majority in the legislature, Ghulam Mohammad dismissed him in April 1953. The East Bengal electorate demonstrated its dissatisfaction when an election was held in March 1954. A "United Front" was formed, including the extreme right (religious fundamentalist) and left (quasi-Marxist). Its main leaders were the aged Fazl ul-Haq and his revamped Workers and Peasants Party and Suhrawardy, who made his comeback with a new party, the Awami League. The Front won 300 seats, while the Muslim League retained only 10. The Front ministers were dismissed after two months. Ghulam Mohammad appointed Major General Iskander Mirza governor of East Bengal. He announced a tough regime, and his task was simplified by the quarrels among the different elements of the United Front. The deputy speaker was killed in an assembly brawl.

In 1956 Pakistan at last obtained a proper constitution in which both wings were equally represented. Thus far, prime ministers had come and gone; Suhrawardy, who took office in September 1956 with a motley group of supporters, lasted only one year. In 1958, government by politicians was superseded by a military regime.

Military rule

Under the military the elite civil servants assumed great importance, which adversely affected the East wing. In 1947 there had been only one Bengali Muslim in the Indian Civil Service (ICS), whereas the West wing had produced about 40. Although recruitment policy was designed to diminish the difference, by 1960 only about one-third of the personnel in the Civil Service of Pakistan (successor to the ICS) were Bengalis, with none in senior positions.

Bengali discontent festered, finding a spokesman in Mujibur Rahman (known as Sheikh Mujib). Like previous leaders, Mujib belonged to a landed family. Mujib was one of the founders of the Awami League in 1949 and, after Suhrawardy's death, became its leading figure. Jailed repeatedly by the military, he acquired an aura of martyrdom, but he was an orator, not a statesman. He announced a six-point demand for autonomy. When in December 1970 President Yahya Khan ordered elections, the Awami League won 167 of the 169 seats allotted to East Pakistan, or Bangladesh as it was now popularly called, in the National Assembly. This gave the League an overall majority in a chamber of 313 members. In West Pakistan, however, the Pakistan People's Party, led by Zulfikar Ali Bhutto, won 81 of 144 seats; Bhutto saw himself as Mujib's rival.

Throughout March 1971 President Yahya Khan negotiated at length in Dhākā with Mujib while government troops poured in from West Pakistan. Then, on March 25, the army launched a massive attack in which there were heavy casualties, including many students. Mujib was arrested and flown to West Pakistan. Most of the Awami League leaders fled and set up a government-in-exile in Calcutta, declaring Bangladesh an independent state. Internal resistance was mobilized by some Bengali units of the regular army, notably by Major Zia ur-Rahman, who held out for some days in Chittagong before the town's recapture by the Pakistan army. He then retreated to the border and began to organize bands of guerrillas. A different resistance was started by student militants, among whom Abdul Kader Siddiqi with his followers, known as Kader Bahini, acquired a reputation for ferocity. Some 10,000,000 Bengalis, mainly Hindus, fled over the frontier into India. The Indian government watched the struggle with alarm. The Awami League, which they supported, was a moderate middle-class body like the Congress Party; but many guerrillas were leftist. The United States and

China, for different reasons, were committed to a united Pakistan; India and the Soviet Union wanted a Bangladesh dependent on India. Eventually, on Dec. 3, 1971, the Indian army invaded the territory of its neighbour. The Pakistani defenses surrendered on December 16. Mujib was released from jail and returned to a hero's welcome, assuming leadership of the new Bangladesh government in January 1972.

Invasion by India

Revenge was brought against those who had collaborated. Local paramilitary forces, known as Razakars, had been raised. The Bengali force was called Al-Badr, while another, Al-Shams, was recruited from Urdu speakers—still called Bihāris, though most had been born locally. A terrible retribution ensued, with Kader Siddiqi as public executioner. The Bihāris had to flee into enclaves where their numbers gave some security; many were killed. Their pleas to be allowed to go to (West) Pakistan were ignored by Bhutto. Ten years later, most remained in refugee camps.

Mujib preached a secular state: the new national anthem was a poem by Tagore. In 1973 an election gave Mujib a landslide majority, but the euphoria soon turned sour. Prices escalated, and in 1974 a great famine claimed 50,000 lives. Faced with crisis, Mujib became a virtual dictator; corruption and nepotism reached new depths. On Aug. 15, 1975, Mujib was assassinated along with most of his family.

Assassination of Mujib

Right-wing, pro-Pakistan army officers were behind the killing; there have been allegations of U.S. support. The reconstructed army split into rival factions. Some of those who had fought in the resistance were politicized, especially the soldiers. The 1,000 officers and 28,000 soldiers who had been serving in the West since 1970 were not repatriated until 1973–74; they were allegedly pro-Pakistan and jealous of the fighters whom Mujib had favoured. A third military group comprised those who had worked with the Pakistanis in their brutal repression. A second coup in November 1975 brought Major General Zia ur-Rahman into power. Despite his own resistance record he turned against India and favoured those considered pro-Pakistan. A referendum held in May 1977 gave him an enormous vote of confidence. This did not prevent several military coup attempts, and on May 30, 1981, he was assassinated by radical officers. The prompt action of the chief of staff, Lieutenant General Hossain Mohammad Ershad, foiled their plans, and the conspirators were hanged.

The civilian vice-president, Abdus Sattar, was confirmed as president by a nationwide election in 1981; but he was ill, and real power was exercised by Ershad and a National Security Council. On March 24, 1982, Ershad ejected Sattar and took over as chief martial-law administrator. In December 1983 he assumed the office of president. To legitimize his authority he called elections for a National Assembly. He formed his own National Party. The election of May 1986 was contested by many parties. The National Party won 210 of the 330 seats in the legislature, just short of the two-thirds majority required to pass a fundamental law to legalize the martial-law regulations and revert to constitutional practice.

Ershad retired from the military command the following August, demonstrating his confidence that the army was now under control. He called a presidential election for October. The main opposition remained the Awami League, now led by Mujib's daughter, Sheikh Hasina Wajad. One of Mujib's murderers, Lieutenant Colonel Farooq Rahman, who had been in exile, was also a candidate. Ershad received 84 percent of the total.

The political turmoil had little relevance to the country's basic problems. At the 1951 census the East Pakistan population numbered 42,000,000 (about 12,000,000 being Hindus); by the late 1980s there were more than 100,000,000, despite massive emigration to neighbouring Assam and Tripura in India and a smaller exodus over the Arakan border with Burma. Agriculture was still the occupation of more than half the labour force, and what economic development there had been was confined to the environs of Dhākā and Chittagong. (H.R.T.)

For later developments in the history of Bangladesh, see the *Britannica Book of the Year* section in the BRITANNICA WORLD DATA ANNUAL.

For coverage of related topics in the *Macropædia* and *Micropædia,* see the *Propædia,* sections 968 and 976.

BIBLIOGRAPHY

Physical and human geography: For information on the geography of Bangladesh, it is necessary to consult books and documents published both during the Pakistani period and since independence. SYED S. HUSAIN, *East Pakistan: A Profile* (1962), is a collection of essays on the country's geography. B.L.C. JOHNSON, *Bangladesh,* 2nd ed. (1982), is a brief, well-illustrated study. HAROUN ER RASHID, *Geography of Bangladesh* (1977), is a comprehensive work. For demographic, agricultural, and industrial statistics, see *Statistical Yearbook of Bangladesh* (annual); and *Statistical Pocket Book of Bangladesh* (irregular), both published by the government. Other useful works include DON YEO, *Bangladesh, a Traveller's Guide* (1982); A.B.M. SHAMSUDDOULAH, *Introducing Bangladesh Through Books: A Select Bibliography with Introductions and Annotations, 1855–1976* (1976); and A.B.M. SHAMSUL ISLAM, *Bibliography on Population, Health, and Development in Bangladesh* (1986). Works on the economy include NAFIS AHMAD, *A New Economic Geography of Bangladesh* (1976); and HAROUN ER RASHID, *An Economic Geography of Bangladesh* (1981). Postwar economic development is discussed in JUST FAALAND and J.R. PARKINSON, *Bangladesh: The Test Case of Development* (1976); and E.A.G. ROBINSON and KEITH GRIFFIN (eds.), *The Economic Development of Bangladesh Within a Socialist Framework* (1974, reprinted 1986). Rural conditions at the time of independence are discussed in ROBERT D. STEVENS, HAMZA ALAVI, and PETER J. BERTOCCI (eds.), *Rural Development in Bangladesh and Pakistan* (1976), a collection of articles. Traditional life-styles and customs of rural communities are examined in MOHAMMAD AFSARUDDIN, *Rural Life in Bangladesh: A Study of Five Selected Villages,* 2nd ed. (1979); M. HABIBULLAH and A. FAROUK (eds.), *Some Aspects of Rural Capital Formation in East Pakistan* (1963); JOSEPH F. STEPANEK, *Bangladesh, Equitable Growth?* (1979); BETSY HARTMANN and JAMES K. BOYCE, *A Quiet Violence: View from a Bangladesh Village* (1983); GUDRUN MARTIUS VON HARDER, *Women in Rural Bangladesh: An Empirical Study in Four Villages of Comilla District* (1981); and TAHRUNNESSA A. ABDULLAH and SONDRA A. ZEIDENSTEIN, *Village Women of Bangladesh: Prospects for Change* (1982). Broader social studies include M. ANISUZZAMAN, *Bangladesh Public Administration and Society* (1979); BEN WHITAKER, IAIN GUEST, and DAVID ENNALS, *The Biharis of Bangladesh,* 4th rev. ed. (1982); and CLARENCE MALONEY, K.M. ASHRAFUL AZIZ, and PROFULLA C. SARKER, *Beliefs and Fertility in Bangladesh* (1981).

(S.S.H.)

History: SUBRATA ROY CHOWDHURY, *The Genesis of Bangladesh: A Study in International Legal Norms and Permissive Conscience* (1972), examines the political history of the country. For the background of the civil war of 1971, see G.W. CHOUDHURY, *The Last Days of United Pakistan* (1974); HERBERT FELDMAN, *The End and the Beginning: Pakistan, 1969–1971* (1975); JYOTI SEN GUPTA, *History of Freedom Movement in Bangladesh, 1943–1973* (1974); and PRAN CHOPRA, *India's Second Liberation* (1973). The events of the civil war are chronicled in LAWRENCE LIFSCHULTZ, *Bangladesh, the Unfinished Revolution* (1979); MARCUS FRANDA, *Bangladesh, the First Decade* (1981); and TALUKDER MANIRUZZAMAN, *The Bangladesh Revolution and Its Aftermath* (1980). Both historical background and surveys of later developments are provided in CRAIG BAXTER, *Bangladesh: A New Nation in an Old Setting* (1984); and CHARLES PETER O'DONNELL, *Bangladesh: Biography of a Muslim Nation* (1984). The political forces that brought about the emergence of independent Bangladesh are discussed in G.P. BHATTACHARJEE, *Renaissance and Freedom Movement in Bangladesh* (1973); and MD. ABDUL WADUD BHUIYAN, *Emergence of Bangladesh and Role of Awami League* (1982). Also see MATIUR RAHMAN, *Bangladesh Today: An Indictment and a Lament* (1978); MATIUR RAHMAN and NAEEM HASAN, *Iron Bars of Freedom* (1980); ANTHONY MASCARENHAS, *Bangladesh: A Legacy of Blood* (1986); TALUKDER MANIRUZZAMAN, *Group Interests and Political Changes: Studies of Pakistan and Bangladesh* (1982); and ASOKA RAINA, *Inside RAW: The Story of India's Secret Service* (1981). The following works place the emergence of independent Bangladesh into regional and world perspective: KULDIP NAYAR, *Distant Neighbours: A Tale of the Subcontinent* (1972); and G.W. CHOUDHURY, *India, Pakistan, Bangladesh, and the Major Powers: Politics of a Divided Subcontinent* (1975).

(H.R.T.)

Banks and Banking

The principal types of banking in the modern industrial world are commercial banking and central banking. A commercial banker is a dealer in money and in substitutes for money, such as checks or bills of exchange. The banker also provides a variety of financial services. The basis of the banking business is borrowing from individuals, firms, and occasionally governments—*i.e.*, receiving "deposits" from them. With these resources and also with the bank's own capital, the banker makes loans or extends credit and also invests in securities. The banker makes profit by borrowing at one rate of interest and lending at a higher rate and by charging commissions for services rendered.

A bank must always have cash balances on hand in order to pay its depositors upon demand or when the amounts credited to them become due. It must also keep a proportion of its assets in forms that can readily be converted into cash. Only in this way can confidence in the banking system be maintained. Provided it honours its promises (*e.g.*, to provide cash in exchange for deposit balances), a bank can create credit for use by its customers by issuing additional notes or by making new loans, which in their turn become new deposits. The amount of credit it extends may considerably exceed the sums available to it in cash. But a bank is able to do this only as long as the public believes the bank can and will honour its obligations, which are then accepted at face value and circulate as money. So long as they remain outstanding, these promises or obligations constitute claims against that bank and can be transferred by means of checks or other negotiable instruments from one party to another. These are the essentials of deposit banking as practiced throughout the world today, with the partial exception of Soviet-type institutions.

Another type of banking is carried on by central banks, which are bankers to governments and "lenders of last resort" to commercial banks and/or other financial institutions. They are often responsible for formulating and implementing their country's monetary and credit policies, usually in cooperation with the government. In some cases—*e.g.*, the U.S. Federal Reserve System—they have been established specifically to lead or regulate the banking system; in other cases—*e.g.*, the Bank of England—they have come to perform these functions through a process of evolution.

Some institutions often called banks, such as finance companies, savings banks, investment banks, trust companies, and home-loan banks, do not perform the banking functions described above and are best classified as financial intermediaries. Their economic function is that of channelling savings from private individuals into the hands of those who will use them, in the form of loans for building purposes or for the purchase of capital assets. These financial intermediaries cannot, however, create money (*i.e.*, credit) as the commercial banks do; they can lend no more than savers place with them.

This article describes the historical development of various banking functions and institutions, the basic principles of modern banking practice, and the structure of a number of important national banking systems. Certain concepts not addressed here that are nonetheless fundamental to banking are treated in the articles ACCOUNTING and MONEY.

The article is divided into the following sections:

The development of banking systems

Banking is of ancient origin, though little is known about it prior to the 13th century. Many of the early "banks" dealt primarily in coin and bullion, much of their business being money changing and the supplying of foreign and domestic coin of the correct weight and fineness. Another important early group of banking institutions was the merchant bankers, who dealt both in goods and in bills of exchange, providing for the remittance of money and payment of accounts at a distance but without shipping actual coin. Their business arose from the fact that many of these merchants traded internationally and held assets at different points along trade routes. For a certain consideration, a merchant stood prepared to accept instructions to pay money to a named party through one of his agents elsewhere; the amount of the bill of exchange would be debited by his agent to the account of the merchant banker, who would also hope to make an additional profit from exchanging one currency against another. Because there was a possibility of loss, any profit or gain was not subject to the medieval ban on usury. There were, moreover, techniques for concealing a loan by making foreign exchange available at a distance but deferring payment for it so that the interest charge could be camouflaged as a fluctuation in the exchange rate.

loney
iangers
id
ierchant
ankers

Another form of early banking activity was the acceptance of deposits. These might derive from the deposit of money or valuables for safekeeping or for purposes of transfer to another party; or, more straightforwardly, they might represent the deposit of money in a current account. A balance in a current account could also represent the proceeds of a loan that had been granted by the banker, perhaps based on an oral agreement between the parties (recorded in the banker's journal) whereby the customer would be allowed to overdraw his account.

English bankers in particular had by the 17th century begun to develop a deposit banking business, and the techniques they evolved were to prove influential elsewhere. The London goldsmiths kept money and valuables in safe custody for their customers. In addition, they dealt in bullion and foreign exchange, acquiring and sorting coin for profit. As a means of attracting coin for sorting, they were prepared to pay a rate of interest, and it was largely

in this way that they began to supplant as deposit bankers their great rivals, the "money scriveners." The latter were notaries who had come to specialize in bringing together borrowers and lenders; they also accepted deposits.

It was found that when money was deposited by a number of people with a goldsmith or a scrivener a fund of deposits came to be maintained at a fairly steady level; over a period of time, deposits and withdrawals tended to balance. In any event, customers preferred to leave their surplus money with the goldsmith, keeping only enough for their everyday needs. The result was a fund of idle cash that could be lent out at interest to other parties.

Origin of the check

About the same time, a practice grew up whereby a customer could arrange for the transfer of part of his credit balance to another party by addressing an order to the banker. This was the origin of the modern check. It was only a short step from making a loan in specie or coin to allowing customers to borrow by check: the amount borrowed would be debited to a loan account and credited to a current account against which checks could be drawn; or the customer would be allowed to overdraw his account up to a specified limit. In the first case, interest was charged on the full amount of the debit, and in the second the customer paid interest only on the amount actually borrowed. A check was a claim against the bank, which had a corresponding claim against its customer.

Another way in which a bank could create claims against itself was by issuing bank notes. The amount actually issued depended on the banker's judgment of the possible demand for specie, and this depended in large part on public confidence in the bank itself. In London, goldsmith bankers were probably developing the use of the bank note about the same time as that of the check. (The first bank notes issued in Europe were by the Bank of Stockholm in 1661.) Some commercial banks are still permitted to issue their own notes, but in most countries this has become a prerogative of the central bank.

In Britain the check soon proved to be such a convenient means of payment that the public began to use checks for the larger part of their monetary transactions, reserving coin (and, later, notes) for small payments. As a result, banks began to grant their borrowers the right to draw checks much in excess of the amounts of cash actually held, in this way "creating money"—*i.e.,* claims that were generally accepted as means of payment. Such money came to be known as "bank money" or "credit." Excluding bank notes, this money consisted of no more than figures in bank ledgers; it was acceptable because of the public's confidence in the ability of the bank to honour its liabilities when called upon to do so.

Bank money or credit

When a check is drawn and passes into the hands of another party in payment for goods or services, it is usually paid into another bank account. Assuming that the overdraft technique is employed, if the check has been drawn by a borrower, the mere act of drawing and passing the check will create a loan as soon as the check is paid by the borrower's banker. Since every loan so made tends to return to the banking system as a deposit, deposits will tend to increase for the system as a whole to about the same extent as loans. On the other hand, if the money lent has been debited to a loan account and the amount of the loan has been credited to the customer's current account, a deposit will have been created immediately.

One of the most important factors in the development of banking in England was the early legal recognition of the negotiability of credit instruments or bills of exchange. The check was expressly defined as a bill of exchange. In continental Europe, on the other hand, limitations on the negotiability of an order of payment prevented the extension of deposit banking based on the check. Continental countries developed their own system, known as giro payments, whereby transfers were effected on the basis of written instructions to debit the account of the payer and to credit that of the payee.

The business of banking

The business of banking consists of borrowing and lending. As in other businesses, operations must be based on capital, but banks employ comparatively little of their own capital in relation to the total volume of their transactions. The purpose of capital and reserve accounts is primarily to provide an ultimate cover against losses on loans and investments. In the United States capital accounts also have a legal significance, since the laws limit the proportion of its capital a bank may lend to a single borrower. Similar arrangements exist elsewhere.

FUNCTIONS OF COMMERCIAL BANKS

The balance sheet

The essential characteristics of the banking business may be described within the framework of a simplified balance sheet. A bank's main liabilities are its capital (including reserves and, often, subordinated debt) and deposits. The latter may be from domestic or foreign sources (corporations and firms, private individuals, other banks, and even governments). They may be repayable on demand (sight deposits or current accounts) or repayable only after the lapse of a period of time (time, term, or fixed deposits and, occasionally, savings deposits). A bank's assets include cash (which may be held in the form of credit balances with other banks, usually with a central bank but also, in varying degrees, with correspondent banks); liquid assets (money at call and short notice, day-to-day money, short-term government paper such as treasury bills and notes, and commercial bills of exchange, all of which can be converted readily into cash without risk of substantial loss); investments or securities (substantially medium-term and longer term government securities—sometimes including those of local authorities such as states, provinces, or municipalities—and, in certain countries, participations and shares in industrial concerns); loans and advances made to customers of all kinds, though primarily to trade and industry (in an increasing number of countries, these include term loans and also mortgage loans); and, finally, the bank's premises, furniture, and fittings (written down, as a rule, to quite nominal figures).

All bank balance sheets must include an item that relates to contingent liabilities (*e.g.,* bills of exchange "accepted" or endorsed by the bank), exactly balanced by an item on the other side of the balance sheet representing the customer's obligation to indemnify the bank (which may also be supported by a form of security taken by the bank over its customer's assets). Most banks of any size stand prepared to provide acceptance credits (also called bankers' acceptances); when a bank accepts a bill, it lends its name and reputation to the transaction in question and, in this way, ensures that the paper will be more readily discounted.

Deposits. The bulk of the resources employed by a modern bank consists of borrowed money (that is, deposits), which is lent out as profitably as is consistent with safety. Insofar as an increase in deposits provides a bank with additional cash (which is an asset), the increase in cash supplements its loanable resources and permits a more than proportionate increase in its loans.

The way in which deposits may be increased

An increase in deposits may arise in two ways. (1) When a bank makes a loan, it may transfer the sum to a current account, thus directly creating a new deposit; or it may arrange a line of credit for the borrower upon which he will be permitted to draw checks, which, when deposited by third parties, likewise create new deposits. (2) An enlargement of government expenditure financed by the central bank may occasion a growth in deposits, since claims on the government that are equivalent to cash will be paid into the commercial banks as deposits. In the first instance, with the increase in bank deposits goes a related increase in the potential liability to pay out cash; in the second case, the increase in deposits with the commercial banks is accompanied by a corresponding increase in commercial bank holdings of money claims that are equivalent to cash.

Taking one bank in isolation, an increase in its loans may result in a direct increase in deposits. This may occur either as a result of a transfer to a current account (as above) or a transfer to another customer of the same bank. Once again, there is an increase in the potential liability to pay out cash. On the other hand, if there has been an increase in loans by another bank (including an increase

in central bank loans to the government), this may give rise to increased deposits with the first bank, matched by a corresponding claim to cash (or its equivalent). For these reasons a bank can generally expect that, if there is an increase in deposits, there will also be some net acquisition of cash or of claims for receipt of cash. It is in this way that an increase in deposits usually provides the basis for further bank lending.

Except in countries where banks are small and insecure, banks as a whole can usually depend on their current account debits being largely offset by credits to current accounts, though from time to time an individual bank may experience marked fluctuations in its deposit totals, and all banks in a country may be subject to seasonal variations. Even when deposits are repayable on demand, there is usually a degree of inertia in the deposit structure that prevents sharp fluctuations; if money is accepted contractually for a fixed term or if notice must be given before its repayment, this inertia will be greater. On the other hand, if a significant proportion of total deposits derives from foreign sources, there is likely to be an element of volatility arising from international conditions.

In banking, confidence on the part of the depositors is the true basis of stability. Confidence is steadier if there exists a central bank to act as a "lender of last resort." Another means of maintaining confidence employed in some countries is deposit insurance, which protects the small depositor against loss in the event of a bank failure. Such protection was the declared purpose of the "nationalization" of bank deposits in Argentina between 1946 and 1957; banks receiving deposits acted merely as agents of the government-owned and government-controlled central bank, all deposits being guaranteed by the state.

Maintaining liquidity

Reserves. Since the banker undertakes to provide depositors with cash on demand or upon prior notice, it is necessary to hold a cash reserve and to maintain a "safe" ratio of cash to deposits. The safe ratio is determined largely through experience. It may be established by convention (as it was for many years in England) or by statute (as in the United States and elsewhere). If a minimum cash ratio is required by law, a portion of a bank's assets is in effect frozen and not available to meet sudden demands for cash from the bank's customers. In order to provide more flexibility, required ratios are frequently based on the average of cash holdings over a specified period, such as a week or a month. In addition to holding part of the bank's assets in cash, a banker will hold a proportion of the remainder in assets that can quickly be converted into cash without significant loss. No banker can safely ignore the necessity of maintaining adequate reserves of liquid assets; some prefer to limit the sum of loans and investments to a certain percentage of deposits, not allowing their loan–deposit ratio to run for any length of time at too high a level.

Unless a bank held cash covering 100 percent of its demand deposits, it could not meet the claims of depositors if they were all to exercise in full and at the same time their rights to demand cash. If that were a common phenomenon, deposit banking could not long survive. For the most part, the public is prepared to leave its surplus funds on deposit with the banks, confident that they will be repaid if needed. But there may be times when unexpected demands for cash exceed what might reasonably have been anticipated; therefore, a bank must not only hold part of its assets in cash but also must keep a proportion of the remainder in assets that can be quickly converted into cash without significant loss. Indeed, in theory, even its less liquid assets should be self-liquidating within a reasonable time.

A bank may mobilize its assets in several ways. It may demand repayment of loans, immediately or at short notice; it may sell securities; or it may borrow from the central bank, using paper representing investments or loans as security. Banks do not precipitately call in loans or sell marketable assets, because this would disrupt the delicate debtor–creditor relationships and increase any loss of confidence, probably resulting in a run on the banks. Ready cash may be obtainable in this way only at a very high price. Banks must either maintain their cash reserves and other liquid assets at a high level or have access to a "lender of last resort," such as a central bank, able and willing to provide cash against the security of eligible assets. In a number of countries the commercial banks have at times been required to maintain a minimum liquid assets ratio. But central banks impose such requirements less as a means of maintaining appropriate levels of commercial bank liquidity than as a technique for influencing directly the lending potential of the banks (see below).

Kinds of assets of commercial banks

Among the assets of commercial banks, investments are less liquid than money-market assets such as call money and treasury bills. By maintaining an appropriate spread of maturities, however, it is possible to ensure that a proportion of a bank's investments is regularly approaching redemption, thereby producing a steady flow of liquidity and in that way constituting a secondary liquid assets reserve. Some banks, particularly in the United States and Canada, have at times favoured the "dumbbell" distribution of maturities, a significant proportion of the total portfolio being held in long-dated maturities with a high yield, a small proportion in the middle ranges, and another significant proportion in short-dated maturities. Following redemption the banks usually reinvest all or most of the proceeds in longer term maturities that in due course become increasingly short-term. Interest-rate expectations frequently modify the shape of a maturity distribution, and, in times of great uncertainty with regard to interest rates, banks will tend to hold the bulk of their securities at short term, and something like a T-distribution may then be preferred (mainly shorts, supported by small amounts of medium to longer dated paper). Investments and money-market assets merge into each other. The dividing line is arbitrary, but there is an essential difference: the liquidity of investments depends primarily on marketability (though sometimes it also depends on the readiness of the government or its agent to exchange its own securities for cash); the liquidity of money-market assets, on the other hand, depends partly on marketability but mainly on the willingness of the central bank to purchase them or accept them as collateral for a loan. This is why money-market assets are more liquid than investments.

INDUSTRIAL FINANCE

Long-term and medium-term lending. Banks that do a great deal of long-term lending to industry must ensure their liquidity by maintaining relatively large capital funds and a relatively high proportion of long-term borrowings (*e.g.*, time deposits, or issues of bonds or debentures), as well as valuing their investments very conservatively. Such banks, notably the French *banques d'affaires* and the West German commercial banks, have developed special means of reducing their degree of risk. Every investment is preceded by a thorough technical and financial investigation. The initial advance may be an interim credit, later converted into a participation. Only when market conditions are favourable is the original investment converted into marketable securities, and an issue of shares to the public is arranged. One function of these banks is to nurse an investment along until the venture is well established. Even assuming its ultimate success, a bank may be obliged to hold such shares for long periods before being able to liquidate them. In addition, they often retain an interest in a firm as an ordinary investment as well as to ensure a degree of continuing control over it.

The long-term provision of industrial finance in Britain and the Commonwealth countries is usually handled by specialist institutions, with the commercial banks providing only part of the necessary capital. In Japan, the long-term financial needs of industry are met partly by special industrial banks (which also issue debentures as well as accepting deposits) and partly by the ordinary commercial banks. In West Germany the commercial banks customarily handle long-term finance.

Term loans

Since World War II the commercial banks in the United States have developed the so-called term loan, especially for financing industrial capital requirements. The attempt to popularize the term loan began in the economic depression of the 1930s, when the banks tried to expand their business by offering finance for a period of years. Most

term loans have an effective maturity of little more than five years, though some run for 10 years or more. They are usually arranged between the customer and a group of lending banks, sometimes in cooperation with other institutions such as insurance companies, and are normally subject to a formal term loan agreement. Banks in Britain, western Europe, the Commonwealth, and Japan began during the 1960s to give term loans both to industry and to agriculture.

Short-term lending. Short-term loans are the core of the banking business even in countries where commercial banks make long-term loans to industry. Much short-term lending consists in the provision of working capital, but the banks also provide temporary finance for fixed capital development, aiding a customer until long-term finance can be found elsewhere.

Overdrafts

Much of this short-term lending is done by overdraft, particularly in the United Kingdom and a number of the Commonwealth countries, or by way of "current account lending" in many western European countries. The overdraft permits a depositor to overdraw an account up to an agreed limit. In theory, overdrafts are repayable on demand or after reasonable notice has been given, but often they are allowed to run on indefinitely, subject to a periodic review. An advance is reduced or repaid whenever the account is credited with deposits and recreated when new checks are drawn upon it, interest being paid only on the amount outstanding.

An alternative method of short-term lending is to debit a loan account with the amount borrowed, crediting the proceeds to a current account; interest is usually payable on the whole amount of the loan, which normally is for a fixed period of time. (In Britain arrangements are sometimes more flexible, and the term of the loan may be set by oral agreement.)

Discountable paper

In a number of countries, including the United States, the United Kingdom, France, West Germany, and Japan, short-term finance is often made available on the basis of discountable paper—commercial bills or promissory notes. Some of this paper is usually rediscountable at the central bank, thus becoming virtually a liquid asset, unlike a bank advance or loan.

Credit may be offered with or without formal security, depending on the reputation and financial strength of the borrower. In many countries a customer may use a number of banks, and these institutions usually freely exchange information about joint credit risks. In Britain and The Netherlands, however, most concerns tend to use a single banking institution for most of their needs.

Liability management

Traditionally bankers took the view that the liabilities of a bank (in particular, its deposits) were more or less stable and concerned themselves primarily with the investment of these funds. Since the late 1950s and 1960s, especially in North America and latterly in the United Kingdom, there has been a change in emphasis. Banks began to find it more difficult to obtain deposits. Interest rates rose to high levels, and banks were obliged to compete with each other and with other institutions for funds. At the same time, there was little point in paying a high rate of interest for money unless it could be employed profitably. Bankers began to relate the cost of borrowed money directly to the return on loans and investments. Previously the main limitation on a bank's expansion had been its ability to find profitable new business, but now the determining factor became the availability of funds to lend out. The essence of assets and liabilities management, as it came to be called, was deciding what kinds of new money to buy and what to pay for it. In the United States the liabilities side of bank balance sheets now included, *inter alia,* in much larger proportion than during the 1960s, repurchase agreements (under which securities are sold subject to an agreement to repurchase at a stated date), federal funds purchases (on the assets side, federal funds sales), excess balances of commercial banks and other depository institutions (regularly traded throughout the United States), negotiable certificates of deposit (which can be traded on a secondary market), and, for the larger banks, Eurocurrency borrowings, mostly Eurodollars (dollar balances held abroad). In the United Kingdom, "bought" money consisted of wholesale (*i.e.,* large) deposits (on which money market rates were paid), negotiable certificates of deposit, interbank borrowings, and Eurocurrency purchases. This bought money could then be used to finance the loan demand, including term loans, long favoured in the United States but a more recent innovation in the United Kingdom and elsewhere, where they were developed considerably in the 1970s. Although much of the lending financed by bought money was by way of term loans, these could be "rolled over," with an interest rate adjustment, every three or six months, and there could therefore be a measure of interest-rate matching and also sometimes a matching of maturities. In less sophisticated environments than North America and the United Kingdom, there was again an increasing emphasis on bought money to meet any expansion in loan demands (much of which was now term lending), with an adjustment at the margin when more funds were needed—*e.g.,* wholesale deposits, certificates of deposit, interbank borrowings, and purchases of Eurocurrencies.

The principles of central banking

The need for a central bank

The principles of central banking grew up in response to the recurrent British financial crises of the 19th century and were later adopted in other countries. Modern market economies are subject to frequent fluctuations in output and employment. Although the causes of these fluctuations are various, there is general agreement that the ability of banks to create new money may exacerbate them. Although an individual bank may be cautious enough in maintaining its own liquidity position, the expansion or contraction of the money supply to which it contributes may be excessive. This raises the need for a disinterested outside authority able to view economic and financial developments objectively and to exert some measure of control over the activities of the banks. A central bank should also be capable of acting to offset forces originating outside the economy, although this is much more difficult.

RESPONSIBILITIES OF CENTRAL BANKS

The first concern of a central bank is the maintenance of a soundly based commercial banking structure. While this concern has grown to comprehend the operations of all financial institutions, including the several groups of nonbank financial intermediaries, the commercial banks remain the core of the banking system. A central bank must also cooperate closely with the national government. Indeed, most governments and central banks have become intimately associated in the formulation of policy.

Relationships with commercial banks. One source of economic instability is the supply of money. Even in relatively well-controlled banking systems, banks have sometimes expanded credit to such an extent that inflationary pressures developed. Such an overexpansion in bank lending would be followed almost inevitably by a period of undue caution in the making of loans. Frequently the turning point was associated with a financial crisis, and bank failures were not uncommon. Even today, failures occur from time to time. Such crises in the past often threatened the existence of financial institutions that were essentially sound, and the authorities sometimes intervened to prevent complete collapse.

The willingness of a central bank to offer support to the commercial banks and other financial institutions in time of crisis was greatly encouraged by the gradual disappearance of weaker institutions and a general improvement in bank management. The dangers of excessive lending came to be more fully appreciated, and the banks also became more experienced in the evaluation of risks. In some cases, the central bank itself has gone out of its way to educate commercial banks in the canons of sound finance. In the United States the Federal Reserve System examines the books of the commercial banks and carries on a range of frankly educational activities. In developing countries such as India and Pakistan, central banks have also set up departments to maintain a regular scrutiny of commercial bank operations.

The most obvious danger to the banks is a sudden and

overwhelming run on their cash resources in consequence of their liability to depositors to pay on demand. In the ordinary course of business, the demand for cash is fairly constant or subject to seasonal fluctuations that can be foreseen. It has become the responsibility of the central bank to protect banks that have been honestly and competently managed from the consequences of a sudden and unexpected demand for cash. In other words, the central bank came to act as the "lender of last resort." To do this effectively, it was necessary that the central bank be permitted either to buy the assets of commercial banks or to make advances against them. It was also necessary that the central bank have the power to issue money acceptable to bank depositors. But if a central bank was to play this role with respect to commercial banks, it was only reasonable that it or some related authority be allowed to exercise a degree of control over the way in which the banks conducted their business.

The central bank as lender of last resort"

Role of central banks in the banking system

Most central banks now take a continuing day-to-day part in the operations of the banking system. The Bank of England, for example, has been increasingly in the market to ensure that the banks have a steady supply of cash, even during periods of credit restriction. It also lends regularly to the discount houses, supplementing their resources whenever the commercial banks feel the need to call back money they have on loan to them. In the United States the Federal Reserve System has operated in a similar way by buying and selling securities on the open market and by lending to dealers in government securities on the basis of repurchase agreements. The Federal Reserve may also discount paper submitted by the commercial banks through the Federal Reserve banks. The various techniques of credit control in use are discussed in greater detail below.

The evolution of those working relations among banks implies a community of outlook that in some countries is relatively recent. The whole concept of a central bank as responsible for the stability of the banking system presupposes mutual confidence and cooperation. For this reason, contact between the central bank and the commercial banks must be close and continuous. The latter must be encouraged to feel that the central bank will give careful consideration to their views on matters of common concern. Once the central bank has formulated its policy after a full consideration of the facts and of the views expressed, however, the commercial banks must be prepared to accept its leadership. Otherwise, the whole basis of central banking would be undermined.

The central bank and the national economy. *Relationships with other countries.* Since no modern economy is self-contained, central banks must give considerable attention to trading and financial relationships with other countries. If goods are bought abroad, there is a demand for foreign currency to pay for them. Alternatively, if goods are sold abroad, foreign currency is acquired that the seller ordinarily wishes to convert into the home currency. These two sets of transactions usually pass through the banking system, but there is no necessary reason why, over the short period, they should balance. Sometimes there is a surplus of purchases and sometimes a surplus of sales. Short-period disequilibrium is not likely to matter very much, but it is rather important that there be a tendency to balance over a longer period, since it is difficult for a country to continue indefinitely as a permanent borrower or to continue building up a command over goods and services that it does not exercise.

Maintaining the balance of foreign payments

Short-period disequilibrium can be met very simply by diminishing or building up balances of foreign exchange. If a country has no balances to diminish, it may borrow, but normally it at least carries working balances. If the commercial banks find it unprofitable to hold such balances, the central bank is available to carry them; indeed, it may insist on concentrating the bulk of the country's foreign-exchange resources in its hands or in those of an associated agency.

Long-period equilibrium is more difficult to achieve. It may be approached in three different ways: price movements, exchange revaluation (appreciation or depreciation of the currency), or exchange controls.

Price levels may be influenced by expansion or contraction in the supply of bank credit. If the monetary authorities wish to stimulate imports, for example, they can induce a relative rise in home prices by encouraging an expansion of credit. If additional exports are necessary in order to achieve a more balanced position, the authorities can attempt to force down costs at home by operating to restrict credit.

The objective may be achieved more directly by revaluing a country's exchange rate. Depending on the circumstances, the rate may be appreciated or depreciated, or it may be allowed to "float." Appreciation means that the home currency becomes more valuable in terms of the currencies of other countries and that exports consequently become more expensive for foreigners to buy. Depreciation involves a cheapening of the home currency, thus lowering the prices of export goods in the world's markets. In both cases, however, the effects are likely to be only temporary, and for this reason the authorities often prefer relative stability in exchange rates even at the cost of some fluctuation in internal prices.

Quite often governments have resorted to exchange controls (sometimes combined with import licensing) to allocate foreign exchange more or less directly in payment for specific imports. At times, a considerable apparatus has been assembled for this purpose, and, despite "leakages" of various kinds, the system has proved reasonably efficient in achieving balance on external payments account. Its chief disadvantage is that it interferes with normal market processes, thereby encouraging rigidities in the economy, reinforcing vested interests, and restricting the growth of world trade.

Whatever method is chosen, the process of adjustment is generally supervised by some central authority—the central bank or some institution closely associated with it—that can assemble the information necessary to ensure that the proper responses are made to changing conditions.

Moderating booms and slumps

Economic fluctuations. As noted above, monetary influences may be an important contributory factor in economic fluctuations. An expansion in bank credit makes possible, if it does not cause, the relative overexpansion of investment activity characteristic of a boom. Insofar as monetary policy can assist in mitigating the worst excesses of the boom, it is the responsibility of the central bank to regulate the amount of lending by banks and perhaps by other financial institutions as well. The central bank may even wish to influence in some degree the direction of lending as well as the amount.

An even greater responsibility of the central bank is that of taking measures to prevent or overcome a slump. Recessions, when they occur, are often in the nature of adjustments to eliminate the effects of previous overexpansion. Such adjustments are necessary to restore economic health, but at times they have tended to go too far; depressive factors have been reinforced by a general lack of confidence, and, once this has happened, it has proved extremely difficult to stimulate recovery. In these circumstances, prevention is likely to be far easier than cure. It has therefore become a recognized function of the central bank to take steps to preclude, if possible, any such general deterioration in economic activity.

For the central bank to be effective in regulating the volume and distribution of credit so that economic fluctuations may be damped, if not eliminated, it must at least be able to regulate commercial bank liquidity (the supply of cash and "near cash"), because this is the basis of bank lending. Monetary authorities in a number of countries have begun to resort increasingly to the management of monetary aggregates as a basic policy. This does not mean an uncritical acceptance of monetarist philosophy but rather what the U.S. economist and banker Paul A. Volcker has called "practical monetarism." In addition to the Federal Reserve in the United States, a growing number of western European countries have adopted the practice of setting growth targets for the money supply and sometimes other monetary targets as well (like domestic credit expansion), usually setting some range of allowable variation. Japan has had reservations and has preferred to indicate monetary projections or forecasts, partly because

of the difficulty of changing a set target should it become necessary. Nor is there any great degree of consensus as to which target or aggregate to employ. In general terms, choice of a particular aggregate as a basis for reference would be linked to the theories—more or less explicit—on which the actions of a particular central bank are based and also on the state of the country's economy and its financial environment. Where there are publicly declared targets, these can have an important effect by the very fact of being announced.

There is now little dispute about the broad objectives, though the techniques of control are various and depend to some extent on environmental factors. It would be incorrect to suppose, however, that the actions of the central bank can, unaided, achieve a high degree of stability. It can by wise guidance contribute to that end, but monetary action is in no sense a panacea; at all times, the degree to which it is likely to be effective depends on the provision of an appropriate fiscal environment (see GOVERNMENT FINANCE).

Developing the banking system

Banking services. Another responsibility of the central bank is to ensure that banking services are adequately supplied to all members of the community that need them. Some areas of a country may be "under-banked" (*e.g.*, the rural areas of India and the northern and more remote parts of Norway), and central banks have attempted, directly or indirectly, to meet such needs. In France, this need underlay the early extension of branches of the Bank of France to the *départements*. In India the authorities encouraged the opening of "pioneer" branches by the former Imperial Bank of India and its successor, the State Bank of India, latterly by all the nationalized banks, and particularly their extension to rural and semirural areas. In Pakistan, officials of the State Bank of Pakistan played an active part in the foundation of the semipublic National Bank of Pakistan with a similar objective in view.

A different sort of problem arises when the business methods of existing banks are unsatisfactory. In such circumstances, a system of bank inspection and audit organized by the central banking authorities (as in India and Pakistan) or of bank "examinations" (as in the United States) may be the appropriate answer. Alternatively, the supervision of bank operations may be handed over to a separate authority, such as France's Banking Control Commission or South Africa's Registrar of Banks.

In developing countries, central banks may encourage the establishment and growth of specialist institutions such as savings institutions and agricultural credit or industrial finance corporations. These serve to improve the mechanism for tapping existing liquid resources and to supplement the flow of funds for investment in specific fields.

The central bank as a public institution

Responsibilities to the government. Central banks have over the years acquired a number of well-defined responsibilities to their respective national governments. Some, notably the Bank of England, developed into central banks after being, in origin, bankers to the government. More recently it has become a matter of course for a new central bank to accept responsibility for the financial affairs of its government. The reasons are self-evident. Government transactions have become of increasing importance in influencing the workings of the economy, and the institution that holds the government's account is in a strategic position to cushion the commercial banks against the impact of large movements of cash originating in this way. As banker to the government, furthermore, the central bank has an obvious responsibility to provide routine banking services, such as arranging loan flotations and supervising their service, renewal, and redemption. The central bank also usually issues the currency.

Equally important are its responsibilities as an adviser on the probable monetary consequences of any proposed action. In this role the central bank should scrutinize the government's proposals with a certain amount of objectivity and state its point of view with vigour. One may cite a now-famous dictum of Montagu Norman as governor of the Bank of England:

> I think it is of the utmost importance that the policy of the Bank and the policy of the Government should at all times be in harmony—in as complete harmony as possible. I look upon the Bank as having the unique right to offer advice and to press such advice even to the point of "nagging"; but always of course subject to the supreme authority of the Government.

Many central banks are now nationalized, reflecting the increasingly general recognition of the significance of the central bank's role as a servant, if not a creature, of the government. This development is also, in a way, a final recognition of the central bank as a responsible public institution whose major function is to serve the community as a whole, untrammelled by narrow dictates of profit and loss. Most central banks, nevertheless, make very handsome profits.

TECHNIQUES OF CREDIT CONTROL

Central banks have developed a variety of techniques for influencing, regulating, and controlling the activities of commercial banks. These may be divided into (1) the so-called classical, or indirect, techniques and (2) various direct controls. The classical techniques make use of open-market purchases or sales by the central bank of certain types of assets that are invariably associated with fluctuations in interest rates. Direct, or quantitative, credit controls are employed to influence the cash and liquidity bases of commercial bank lending by means of freezing or unfreezing their liquid resources; sometimes ceilings are imposed on bank loans.

Purchases and sales of securities by the central bank

Open-market operations. The way in which open-market operations influence the cash reserves and, through them, the general liquidity of the commercial banks is essentially simple. If the central bank buys securities in the open market, the cash it offers in exchange adds to the reserves of the banks; if the central bank sells securities in the open market, the cash necessary to pay for them is either withdrawn from the banks' reserves or obtained by diminishing holdings of other assets (with the possibility of capital losses in consequence of these sales). It does not matter whether this buying and selling takes place between the central bank and the commercial banks directly or between the central bank and other financial sectors, including the public at large, since these are the customers of the commercial banks.

Open-market operations are invariably associated with related changes in one or more "strategic" rates of interest, the most influential of these rates being the minimum rate at which the central bank does business (the bank rate, or the discount rate), since other rates tend to move in sympathy with it. The central bank seeks to achieve an appropriate and consistent structure of interest rates. If a particular rate structure is desired (*e.g.*, prior to a new issue of government securities or in order to change the emphasis of institutional investment between, say, long-term and short-term securities), it may be necessary to precondition the market by means of open-market operations. To achieve its purposes the central bank must possess (if it is selling) or be willing to absorb (if it is buying) the appropriate types of securities.

In London the specialist banks known as discount houses effectively put to work the revolving fund of cash that circulates through the British banking system. If temporarily there is an inadequate supply of cash, the Bank of England either lends on a short-term basis or buys some of the assets held by the discount market. (From 1980 there was a shift in emphasis from lending to open-market operations, especially by dealing in bankers' acceptances.) Alternatively, the Bank of England may buy assets from the clearing banks (the large joint-stock banks), which then make the relevant moneys available to the market. On the other hand, if the discount market is oversupplied with funds, the Bank of England sells treasury bills, in this way mopping up the excess of cash. These transactions are known as smoothing-out operations. In addition, the Bank of England is also responsible for managing the national debt, and, whether the object is to influence the flows of money or not, such transactions in fact have monetary effects.

In the United States the Federal Reserve System regulates the money supply. Within the Federal Reserve System, the Federal Open Market Committee is the most important single policy-making body. It is presided over by the chair-

The Open Market Committee of the Federal Reserve

man of the Board of Governors, with the president of the Federal Reserve Bank of New York as its permanent vice chairman. The main responsibility of the Open Market Committee is to decide upon the timing and amount of open-market purchases or sales of government securities. Since open-market operations must obviously be consistent with other aspects of monetary and credit policy, it is in the committee that broad agreement is reached on matters such as changes in discount rates or reserve requirements.

One of the big differences between London and New York is that the central banking authorities in New York maintain direct relationships more or less continuously with the nonbank government securities dealers as well as with the commercial banks. The Federal Reserve Bank of New York may make temporary accommodation available to some 35 primary dealers (including certain banks) under a repurchase agreement, whereby securities are sold to the bank under an agreement that they be repurchased after a stipulated time. These agreements are made only for the purpose of supplying reserves to the banking system, but from the dealer's standpoint they are helpful in financing portfolios. Such repos, as they are called, may also be done with foreign official accounts. Since early 1966 the bank has also been prepared to mop up money by undertaking reverse repurchase agreements, in which the dealers act as intermediaries for large commercial banks with temporarily surplus money that they are prepared to place against bills, subject to the bank's repurchasing them a few days later; the commercial bank concerned lends the dealer the money to finance the holding of the bill. Similar arrangements are also made by the Federal Reserve directly with bank dealers.

All member banks of the Federal Reserve System, and now also other depository institutions, have direct access to the discount service of their Federal Reserve Bank, of which there is one in each of 12 districts. This is a privilege, however, and not a right. In the early years of the system the banks would sell discountable paper to the Federal Reserve, but now they usually borrow against a pledge of government securities held in safe custody with the Federal Reserve Bank in question. The Federal Reserve lends for a number of purposes but always at a time of general stress. It is assumed that, as the pressure abates, borrowing banks will repay their indebtedness as quickly as possible. Under ordinary conditions, the continuous use of Federal Reserve credit by a member bank over a considerable period is not regarded as appropriate.

Direct controls. The so-called classical techniques of credit control—open-market operations and discount policy—can only be employed where there is a sufficiently developed complex of markets in which to buy and sell assets of the type that commercial banks ordinarily hold. Direct credit controls have a wider range of application. They may be used either as a substitute for the classical techniques or as a supplement to them. Direct controls are more likely to be resorted to when the money market is undeveloped, because then a central bank can only impose its authority by means of direct action. This is often the situation facing a newly established central bank. Rather than wait for the slow evolution of a money market, the authorities may provide the central bank from the start (as in Pakistan, the Philippines, Sri Lanka, and Malaysia) with very full powers to control the banking system.

The aim in imposing a direct, quantitative regulation of credit is to curb inflationary pressures that may result from an expansion of commercial bank lending. This can be done in four main ways: (1) the commercial banks may be required to maintain stated minimum reserve ratios of cash to deposits, a stated liquid assets ratio, or some combination of both; (2) part of the cash resources of the commercial banks may be immobilized at the discretion of the central bank; (3) ceilings may be imposed on the amount of accommodation to be made available to the commercial banks at the central bank (sometimes referred to as "discount quotas"); and (4) a ceiling may be prescribed for commercial bank lending itself.

Variations in reserves

Minimum reserve requirements. The variation of minimum cash reserve requirements as a direct means of quantitative credit control has become increasingly general in recent years. The practice has largely derived from experience in the United States. In its origin the U.S. insistence on stated minimum reserve requirements for commercial banks was simply a means of prescribing minimum standards of sound behaviour. Only later did such ratios come to be seen as a useful supplementary quantitative credit control.

The power granted by the Banking Act of 1935 to the Federal Reserve System to determine the cash reserves of the commercial banks in the United States was employed for the first time during the boom of 1936–37, and periodic variation of minimum reserve requirements subsequently came to be recognized as an appropriate technique for controlling the money supply. The Federal Reserve Board's decisions were sometimes subject to considerable criticism, but, as it became more experienced in the use of this technique, variation in reserve requirements combined with other measures came to be regarded as a useful means of cushioning the economy against a recession. The variation of reserve requirements did not prove as effective in preventing inflation, largely because of the government's insistence that the Federal Reserve simultaneously support the prices of government bonds through open-market operations. This insistence was abandoned by the Treasury in March 1951. Since then, much greater emphasis has been placed on the use of open-market operations, which had become more effective, and the importance of varying minimum reserve requirements as a means of controlling the credit base has diminished in the United States. The technique is still widely used, however, in many countries.

In some countries the authorities require the maintenance of minimum liquid assets ratios. This is often combined with minimum requirements for cash reserves, as in India, Pakistan, and West Germany, though not always (in France, for example, until 1967 there were no minimum cash reserve requirements). Where prescribed minima relate to liquid assets and not to cash as such, reserves are held in the form of earning assets—an important distinction from the point of view of the commercial banks.

An important step toward a uniform and explicit minimum liquidity ratio for the London clearing banks was taken in 1951, when the governor of the Bank of England indicated to the banks that a liquidity ratio of from 32 to 28 percent would be regarded as normal and that it would be undesirable for the ratio to be allowed to fall below 25 percent. By 1957 a fairly rigid 30 percent minimum was in place (it was reduced to 28 percent in 1963). After 1946 the London clearing banks (but not the Scottish banks) also observed a more or less fixed cash ratio of 8 percent. A new element was introduced in 1960, when the Bank of England launched its system of "special deposits" as a means of reinforcing other methods of credit control. Calls were made from time to time on the London clearing banks to deposit with the Bank of England by a specified date some specified percentage of their gross deposits; similar arrangements applied to the Scottish banks, but the calls were smaller. This system lasted until 1971, when a new 12.5 percent minimum reserve ratio (excluding till cash) was introduced. This ratio related to "eligible liabilities" (primarily sterling deposits of up to two years maturity, including sterling certificates of deposit). The banks could also be required to place special deposits with the Bank of England. These arrangements were replaced in August 1981 by a voluntary holding of operational funds with the Bank of England by the London clearing banks ("for clearing purposes") and a uniform requirement of 0.5 percent of an institution's eligible liabilities that would be applied to all banks and licensed deposit-takers with eligible liabilities averaging more than £10,000,000. All banks that were eligible acceptors were also normally required to hold an average equivalent to 6 percent of their eligible liabilities either as secured money with discount houses or as secured call money with money brokers and gilt-edged jobbers, but the amount held in the form of secured money with a discount house was not normally to fall below 4 percent of eligible liabilities. This money became known as "club money."

The use of variable minimum reserve requirements as a means of credit control can, if carried far enough, produce results, especially when the requirements include the holding of cash balances. It is more useful as an anti-inflationary weapon than as a means of countering recession, since it cannot overcome a possible unwillingness of the banks to lend or of their customers to borrow. It is a somewhat clumsy technique, however, and cannot make adequate allowance for the special needs of different institutions.

Immobilization of cash resources. A second group of direct quantitative credit controls involves keeping a portion of the cash resources of commercial banks immobilized at the discretion of the central bank. Two leading examples of this technique were the use of the Treasury Deposit Receipt (TDR) in the United Kingdom during and after World War II and the "special account procedure" adopted in Australia in 1941. Both were means of immobilizing the increased liquidity deriving from wartime government expenditure.

The direct issue of Treasury Deposit Receipts at a nominal rate of interest to banks in the United Kingdom began in July 1940. They were not negotiable in the market nor transferrable between banks, but they could be tendered in payment for government bonds (and tax certificates); hence, during the war years they had a limited degree of liquidity. The Bank of England communicated to the banks collectively the amount of the weekly call, which was divided among them in proportion to their deposits. After the war, TDR's were replaced by treasury bills; in order to reduce the consequent high liquidity of the banks, there was a "forced funding" of £1,000,000,000 of treasury bills in November 1951, which were required to be invested in Serial Funding Stocks.

Special accounts in Australia

The special account procedure introduced in Australia in 1941 had a similar objective. The surplus investable funds of the Australian trading banks, defined as the amount by which each bank's total assets in Australia at any time exceeded the average of its total assets in Australia in August 1939, were required to be placed in special deposit accounts with the Commonwealth Bank (then the central bank) at a nominal rate of interest. A bank was not to withdraw any sum from its special account except with the consent of the Commonwealth Bank; during the war years, the bank generally directed the trading banks to lodge in their special accounts each month an amount equal to the increase in their total assets in Australia during the preceding month, although as a rule a lodgment was not required if it was known that a rise in assets would be followed by an early fall. Legislation in 1945 adopted the special account procedures as a means of general credit control (*e.g.,* to curb inflation), but the provisions were made more flexible. In 1953 a more complicated formula was introduced, and in 1960 the system was abandoned in favour of minimum reserve ratios.

Accommodation ceilings. Some countries have tried placing a limit on the amount of accommodation that the central bank may make available to the commercial banks. The difficulty in this type of quantitative credit control is to make it effective while at the same time allowing for changes in the economy; its most obvious use is as a means of checking inflation, but if the upward pressures on prices are strong, there is a temptation to increase the ceilings so that the restraint then becomes little more than a temporary check.

Usually, it is only when a control begins to be felt and to affect bank profits that the banks become really sensitive to changes in credit policy and the implementation of the control becomes truly effective. The postwar experience

The French *plafonds*

of France is a case in point. *Plafonds,* or "ceilings," were first introduced in France in 1948. Rediscount ceilings (or discount quotas) were fixed for each bank, though some categories of paper were excluded. Ceilings could be increased or (after 1957) reduced.

From the authorities' point of view, the chief difficulty in operating this control was the persistent building up of pressure against the ceilings. This was met partly by upward revisions in the ceilings themselves and partly by instituting a number of safety valves. The degree of elasticity required constituted the chief weakness of the ceiling technique. The central bank was constantly under pressure to adjust the ceilings upward. Some upward revisions were unavoidable, but the problem was to decide which claims were legitimate and which not. Much bilateral bargaining took place between the Bank of France and individual commercial banks, but the banks continued to complain that the strictness of the control was excessive and that the technique was lacking in flexibility.

The inadequacies of the *plafonds* technique in its original form became apparent when prices began to rise rapidly during the Korean War boom, and even the built-in safety valves failed fully to accommodate the pressures on bank liquidity. The need to strengthen the mechanism was obvious, and this was attempted in 1951. Previously, rediscounts had frequently exceeded the ceilings during the month and were only brought within the *plafonds* by special action (*e.g.,* through open-market purchases). The situation was brought under control by introducing a secondary ceiling to which a penalty rate of interest was applied. This was extended in 1958 to permit rediscounts even beyond the secondary ceiling, provided a further penalty was paid; each application, however, was scrutinized by the Bank of France. The system lasted until about the spring of 1964, though it did not finally disappear until 1968, when it was largely replaced by Bank of France operations in the open market. After early 1967 banks also were subject to minimum reserve requirements.

Plafonds, or discount quotas, have also been employed in West Germany, being introduced in 1952 and strengthened in 1955. Quotas may be reduced periodically (after 1964 they were also used to discourage institutions from borrowing abroad). Again there were safety valves (although less generous than in France) and the possibility of extra accommodation (Lombard credits) at a higher rate. In some circumstances, supplementary quotas might be approved for up to six months. A bank might also raise funds through the money market, though likely at higher cost. Discount quotas are still an important tool of credit control in West Germany.

Other countries have employed this technique, including Sweden, where for a time the central bank imposed formal or informal ceilings on banks and sometimes on finance companies. If the banks failed to observe the ceiling, a penalty was applied based on the amount of the excess borrowing and its duration. In Finland, commercial banks have at times been able to borrow limited amounts from the Bank of Finland by way of traditional credit quotas. Beyond these quotas, funds could formerly be obtained as supra-quota credit at a higher rate, but banks now are forced into the official call money market. Denmark, too, has permitted borrowing from the central bank in tranches, with higher (penalty) rates applying after the first tranche of the loan quota has been resorted to, a practice that can be expensive.

General ceilings on credit. Attempts have been made to prescribe a general ceiling within which the quantity of commercial bank lending must be held. This is even more difficult to achieve. One example of such an attempt was the adoption of a "rising ceiling" by Chile in 1953. All banks were required not to expand the volume of their loans to businesses and individuals by more than 1.5 percent a month, using as their basis the average of a bank's advances on selected dates in 1953. Certain types of loans were forbidden, and bank resources were to be directed to productive and distributive activities that really contributed to the expansion of the national economy. Banks were also required to provide information on the destination of their loans. In succeeding years adjustments were made on several occasions in the maximum permitted credit increase, expressed either as a percentage of advances or sometimes as a total for the banking system as a whole. In 1959 all quantitative credit restrictions were removed, and banks were permitted to advance funds up to their financial capacity, provided that they operated within the general banking law. There was no evidence the controls had been effective, but the major problem in Chile was budgetary rather than monetary. A temporary ceiling on loans was imposed by agreement in Canada (in 1951–52), The Netherlands (1957–58), and France (1958–59).

Ceilings on bank advances in the United Kingdom

The United Kingdom had considerable experience with this type of ceiling, introducing it as a temporary measure in 1955, when the banks were asked to bring their advances down by an average of 10 percent. Later an attempt was made to impose a true ceiling, requiring that bank advances not exceed the average of the period October 1956 to September 1957. This was continued until July 1958. Again, in 1961, the authorities indicated the banks must aim at checking the rate of rise in bank advances; this came to be interpreted as a request that the level of advances at the end of 1961 be no higher than in the previous June. The banks also were not to encourage an increase in the volume of commercial bills. The request was modified in May 1962 and largely withdrawn in October; but it was made again in May 1965, when the clearing banks were requested not to increase their advances to the private sector, at an annual rate of more than about 5 percent, in the 12 months to mid-March 1966 (likewise with commercial bills). Other financial institutions were requested to observe a comparable degree of restraint. For 12 months after March 1966, advances and discounts, allowing for seasonal factors, were not permitted to rise above levels set for March 1966. This represented an intensification of the credit squeeze because prices were rising. The credit restriction led to a falling off in business confidence, and, consequently, toward the end of 1966 bank lending was well below the official ceiling. In April 1967 authorities announced a change in techniques, with an emphasis on making calls to special deposits, but the ceilings returned again in November 1967. There was to be no increase in bank advances to the private sector (excluding exports and shipbuilding) except for seasonal reasons. In May 1968 a new ceiling was instituted for all such lending (including that for exports and shipbuilding); the clearing banks were asked to restrict the total of this lending, after seasonal adjustment, to 104 percent of the November 1967 figure, with priority to be given to finance for exports and for activities directly related to improving the balance of payments. The restrictions also extended to other types of credit. Credit became even tighter (in March 1969) when the ceiling was reduced to 98 percent of the November 1967 level. The banks had considerable difficulty in meeting this requirement and agreed merely to "do their best." Advances increased above the ceiling, and, as a penalty, the interest paid by the Bank of England on special deposits was halved. Not until late 1969 did it become clear that the authorities were prepared to abandon their long campaign to get bank loans down to the target figure. The ceiling was subsequently replaced by minimum reserve requirements. The system of quantitative credit control requires, for its successful implementation, the full cooperation of the banking community. In the United Kingdom, where banks base much of their lending on the overdraft technique, the system was very unpopular.

In France, however, the *encadrement du crédit,* as it is called, temporarily imposed in 1958–59, was revived during the first half of 1973. Subject to certain exclusions (*e.g.,* certain investment credits, agricultural credits, export credits, the financing of energy savings and innovation, leasing transactions, and special medium-term construction loans), the mechanism chosen was to permit a certain percentage rate of growth in bank credits in relation to a particular month in the previous year, these limits being fixed quarterly and subject to variation from time to time. Subsequently, in early 1975, reference was made to a fixed base defined as equal to an index of 100, in relation to which the index might be increased (or decreased) and credit expanded (or contracted). The system was further refined to vary the rate of change of credits within different financial sectors, and over the years it has been subject in the interests of flexibility to many amendments. In effect, there has been a combination of quantitative and qualitative credit controls.

Selective controls on credit

In addition to regulating the quantity of credit, central banks have sometimes attempted to influence the directions in which the commercial banks lend. A loose system of control prevailed in the United Kingdom during World War II and afterward, based initially on directives from the Bank of England. A highly formalized technique was employed in Australia during the war and earlier postwar years; detailed and specific instructions were given to the trading banks, marginal cases being referred to the central bank. The system of Voluntary Credit Restraint in the United States in 1951 was similar. The more formal controls seemed to be no more effective than the looser system employed in the United Kingdom.

Selective controls have been imposed on consumer installment finance in the United States and elsewhere (*e.g.,* by stipulating the percentage of deposit that is required and the length of the term over which repayments may be made). Even when these are not varied in order to serve as a control over credit, there is a case for insisting on such requirements for prudential reasons. In the United States, under the Securities Exchange Act of 1934, the Federal Reserve can vary the margins that purchasers of securities must pay in cash, thereby limiting the credit available for this purpose.

The structure of modern banking systems

The banking systems of the world have many similarities, but they also differ, sometimes in quite material respects. The principal differences are in the details of organization and technique. The differences are gradually becoming less pronounced because of the growing efficiency of international communication and the tendency in each country to emulate practices that have been successful elsewhere.

Unit banking and branch banking

Banking systems may be classified in terms of their structure as unit banking, branch banking, or hybrids of the two. For example, unit banking prevails in large areas of the United States. In other countries it is more usual to find a small number of large commercial banks, each operating a highly developed network of branches. This is the system used in England and Wales, among others. Examples of hybrid systems include those of France, West Germany, and India, where banks that are national in scope are supplemented by regional or local banks. Some of these hybrid systems are slowly changing their character, the banks becoming fewer in number and individually larger, with a larger number of branches.

UNIT BANKING: THE UNITED STATES

Bank organization in the United States during the years after World War II was still passing through a phase of structural development that many other countries had completed some decades earlier. Development in the United States has been subject to constraints not found elsewhere. The federal Constitution permits both the national and state governments to regulate banking. Some states prohibit branch banking, largely because of the political influence of small local bankers, thus encouraging the establishment and retention of a large number of unit banks.

Even in its early years the United States had an unusually large number of banks. As the frontiers of settlement were pushed rapidly westward, banks sprang up across the country. One reason for this was the demand for capital in the expanding frontier economy. There was also an obvious need for a large number of banks to serve the diverse and rapidly expanding demands of a growing and constantly migrating population. It must be remembered, too, that at this time communications between the frontiers of settlement and the established centres of commerce and finance were still inadequately developed.

The U.S. concern for competition

As long as communications remained imperfect, the existence of large numbers of competing institutions is not difficult to explain. The subsequent failure of bank mergers or amalgamations to produce a concentration of financial resources in the hands of large banking units can be attributed in part to the character of the federal Constitution as noted above. Among the people, moreover, there was a widespread distrust of monopoly and a deep-rooted fear that a "money trust" might develop. This went hand in hand with a political philosophy that emphasized the virtues of individualism and free competition; restrictions on branching, merging, and on the formation of holding companies were a feature of both the state and the federal

banking laws. Where permitted, however, bank branches are numerous in the United States (especially in California and in New York); in states in which branching is prohibited, one often finds local bank monopolies in small towns. Interstate banking is prohibited by federal law, but large banking organizations have provided financial services (*e.g.*, through loan offices and offices of nonbank subsidiaries) for many years across state lines. A number of states have passed limited interstate or reciprocal banking laws, so that banks in other states with similar laws can acquire or merge with local banks. The banking system of the United States would not work without a network of correspondent bank relationships, which are more highly developed there than in any other country.

From the 1970s there was an acceleration in the evolution of U.S. banking patterns. Unregulated financial institutions (and some nonfinancial institutions) moved into traditional banking activities; at the same time, depository institutions began offering a fuller range of financial services. Money-market mutual funds, for example, secured access to open-market interest rates for investors with relatively small amounts of money. Securities firms and insurance companies moved aggressively into providing a range of liquid financial instruments. Likewise, large manufacturing and retail firms moved into the commercial and retail lending businesses—*e.g.*, by acquiring a savings and loan association, a securities brokerage house, an industrial loan company, a consumer banking business, or even a commercial bank. Meanwhile, depository institutions developed a number of new services, most notably the Negotiable Order of Withdrawal (NOW) account, an interest-bearing savings account with a near-substitute for checks. These appeared first in 1972 in New England and after 1980 spread to the whole nation; they were offered both by commercial banks and by thrift institutions. Share drafts at credit unions also became a means of payment, and after 1978 the automatic transfer services of commercial banks permitted savings-account funds to be transferred automatically to cover overdrafts in checking accounts. So-called Super-NOW accounts (with no interest rate ceilings and unlimited checking facilities with a minimum balance) were subsequently introduced, along with money-market deposit accounts, free of interest rate restrictions but with limited checking.

Rapid changes in financial structure and the supply of financial services posed a host of questions for regulators, and, after much discussion, the Depository Institutions Deregulation and Monetary Control Act was passed in 1980. The object was to change some of the rules—many of them obsolete—under which U.S. financial institutions had operated for nearly half a century. The principal objectives were to improve monetary control and equalize more nearly its cost among depository institutions; to remove impediments to competition for funds by depository institutions, while allowing the small saver a market rate of return; and to expand the availability of financial services to the public and reduce competitive inequalities among financial institutions offering them. The major changes were: (1) Uniform Federal Reserve requirements were phased in on transaction accounts (demand deposits, NOW accounts, telephone transfers, automatic transfers, and share drafts) at all depository institutions—commercial banks (whether Federal Reserve members or not), savings and loan associations, mutual savings banks, and credit unions. (2) The Federal Reserve Board was authorized to collect all data necessary for the monitoring and control of money and credit aggregates. (3) Access to the discount window at Federal Reserve banks was widened to include any depository institution issuing transaction accounts or nonpersonal time deposits. (4) The Federal Reserve was to price its services, to which all depository institutions would now have access. (5) Regulation Q, which had long set interest-rate ceilings on deposits, was to be phased out over a six-year period. (6) An attempt was made to grasp the nettle of the state usury laws. (7) NOW accounts were authorized on a nationwide basis and could be offered by all depository institutions. Other services were extended. (8) The permissible activities of thrift institutions were broadened considerably. (9) Deposit insurance at commercial banks, savings banks, savings and loan associations, and credit unions was raised from $40,000 to $100,000. (10) The "truth in lending" disclosure and financial regulations were simplified to make it easier for creditors to comply.

BRANCH BANKING: THE UNITED KINGDOM

If the United States banks can be taken as representative of a unit banking system, the British system is the prototype of branch banking. Its development was linked to the growth of transportation and communications, for otherwise banks cannot clear checks drawn on other banks and effect remittances speedily and efficiently. The Scots favoured branch banking from the very beginning (the Bank of Scotland was founded in 1695), but at first they were not very successful—largely because of poor communications and the difficulty of supplying branches with adequate amounts of coin. Not until after the Napoleonic Wars, when the road system of Scotland had been greatly improved, did branch banking begin to develop vigorously there. As the Industrial Revolution progressed and as the size of businesses increased, the structure of English banking underwent a corresponding change. Greater resources were required for lending, and banks also needed more extensive interconnections in order to provide an increasing range of services. Where banks remained small, they were frequently unable to take the strain of the larger demand; they tended to become overextended and often failed.

The growth in size of banks was also greatly encouraged by legislation that encouraged joint-stock ownership, beginning in 1826. Joint-stock ownership, which reduced the risk to any individual, must be distinguished from limited liability, which did not become widely accepted until the failure of the City of Glasgow Bank in 1878 demonstrated the need for a legal device to protect the stockholder. The early joint-stock banks tended to remain localized in their business interests; it was only gradually (with the spread of limited liability and disclosure of accounts) that amalgamations began to convert the banking system in England and Wales into its highly concentrated modern form. The main movement was completed before World War I, though there was to be a further degree of concentration in the years after World War II. By these means, British banks were able to attract deposits from all parts of the country and to spread the banking risk over a wide range of industries and areas.

HYBRID SYSTEMS

A third group of banking systems differs from the unit banking system of the United States and also from the branch banking systems of countries that have followed the British model (such as Australia, Canada, New Zealand, and South Africa). This group is characterized by the existence of a small number of banks with branches throughout the country, holding a significant part of total deposits, along with a relatively large number of smaller banks that are regional or local in emphasis. Such systems exist in France, West Germany, and India. Japan has a small number of large city banks with branch networks but a larger number of local banks.

France. Banking institutions in France were classified after World War II into three main groups: deposit banks, *banques d'affaires* (or investment banks), and institutions that were either specialized or operated mainly outside France. New banking legislation in 1966 greatly reduced the importance of the distinction between deposit banks and *banques d'affaires*. There was also (1) a further concentration of banking resources, as a result of several large mergers and also of greater financial integration through share-exchange agreements and interlocking directorates, and (2) the conversion of a number of *banques d'affaires* into deposit banks, which hived off their investment interests into separate investment or holding companies.

Integration in France

Further legislation in 1982 nationalized the remaining large and medium-sized banks (36 in all, plus two financial holding companies—those of Indosuez and Paribas); the largest deposit banks had already been nationalized after World War II. Another new law in 1984 abolished the old divisions between the several categories of banks, which

were now defined simply as *établissements de crédit,* able to receive deposits from the public, undertake credit operations (including loans), and provide means of payment. The intention was to move cautiously toward a system of "universal banking." The new law was extended to cover the Caisse Nationale de Crédit Agricole, the *banques populaires,* the *crédit mutuel,* the central organizations of the cooperatives and the savings banks (and thereby institutions affiliated with them), and semipublic institutions like the Crédit Foncier and the Crédit National, but not the Caisse des Dépôts et Consignations nor the central banking institutions.

All of the regional banks and some local banks have branches. The balanced character of the regional economies often provides these banks with a good portfolio of risks; they serve not only a prosperous agriculture but also a number of local industries. Some of the local banks are also very sound institutions, despite their small size.

The survival of a hybrid system in France, despite the long-run trend toward centralization, reflects certain characteristics of French society. These included, until recently, a strong emphasis on small business, together with a preference for individual and personal service. Particularism in some parts of France manifests itself in support for local institutions, and the local banker also often has the advantage of special knowledge of local industries and people, which makes possible the acceptance of risks that the big banks decline.

West Germany. An even more direct conflict between the forces favouring concentration and those working against it may be seen in Germany, where banking grew in the latter part of the 19th century along with industry. The banks were inclined to rely mainly on their own capital resources and did not at first try to attract deposits from the public. Not until 1874 did the Deutsche Bank A.G. begin to seek deposits through offices specially opened for the purpose. This was done to provide cheap finance for traders, the deposits being invested in mercantile bills that were regarded as both safe and liquid. In pursuit of deposits, the banks built up a widespread network of branch offices, which were also used to establish and maintain industrial contacts throughout the country. The unification of Germany in 1871 removed the political obstacles to a more integrated banking system, and the selection of Berlin as the capital made that city the country's financial centre. Four of the largest banks were already established there; the new Reichsbank was set up in 1876. In addition, the larger and more enterprising of the provincial banks were attracted to the capital. The Berlin stock exchange rapidly displaced that of Frankfurt am Main as the country's leading securities market.

The Berlin banks

The Berlin banks extended their influence by developing correspondent relationships and subsequently by acquiring a financial interest in the provincial banks and being represented on their boards. Each of the big Berlin banks came to be associated with a group of provincial banks more or less under its control. At the same time, all of the banks, Berlin and provincial alike, expanded their business by opening branches.

During World War I the degree of centralization increased; by 1918 the big Berlin banks held more than 65 percent of total deposits. In the early 1920s there were amalgamations, and branch systems became much larger. Bank failures and the financial crisis of 1931 resulted in further consolidation until the German banking system was dominated by three giants. But there were countervailing forces. Probably the most important of these was the establishment of publicly owned banking institutions, such as the communal savings banks and their central institutions, the *Girozentralen,* which became of increasing importance after World War II.

West German savings banks, which were permitted to have checks drawn on them from 1909 and which had giro clearing from the 1920s, now offer a wide range of services, especially to lower income groups and smaller businesses. The large commercial banks have concerned themselves more with big business and with wealthy individuals. The savings banks now compete in wholesale banking as well. A number of them, together with their *Girozentralen,* are to all intents and purposes "universal banks," like the Big Three and the larger regional banks. The Big Three (the Deutsche Bank, the Dresdner Bank, and the Commerzbank) remain unchallenged only in stock exchange and foreign banking business.

Of the private bankers, only about a half-dozen are of any size. The bigger private banks are important in the fields of investment and wholesale banking, while the smaller ones flourish in the leading stock-exchange cities, such as Düsseldorf and Frankfurt am Main. Many of these private bankers, however, are not bankers in the true sense; they subsist mainly on stock-exchange transactions, investment services, portfolio management, and insurance and mortgage brokerage. There are also consumer finance institutions, mortgage and other specialist banks, and a large number of cooperatives.

Regional and private banks are often within the sphere of influence of the Big Three. In some cases the latter have a financial interest in these banks, and in some cases they own them. The Big Three also have shares in certain of the private mortgage banks. There are also "cooperation agreements," and a number of mergers have taken place. In these several ways, much more integration exists than appears on the surface. While banking in West Germany remains a hybrid system, a trend toward greater concentration is evident.

India. Until the 1950s, banking in India was carried on by a large number of banks, many of them quite small. India is still primarily an agricultural country, with an economic and social structure based largely on the village. The integration of banking has been impeded by poor communications, by illiteracy, and by the barriers of language and caste. Banking and credit have remained largely in the hands of the so-called indigenous banker and the village moneylender. Although their influence has been greatly reduced in recent years, they still remain important in many an up-country area. The indigenous banker, who is also a merchant, offers genuine banking services: accepting deposits and remitting funds; making loans quickly and with a minimum of formality; and, by means of the *hundi* (a credit instrument in the form of a bill of exchange), financing a still significant, if declining, portion of India's internal trade and commerce.

Problems in the integration of banking in India

Efforts were made to eliminate the moneylender by developing a network of rural credit cooperatives. When progress proved to be slow, a more successful alternative was found in requiring banks to open "pioneer" branches in rural areas. The first branches were those of the semipublic Imperial Bank of India and its nationalized successor, the State Bank of India (and its subsidiaries). Many smaller banks began to disappear, sometimes by merger and sometimes as a result of failure. Between 1952 and 1967 the number of "reporting" banks fell from 517 to 90. Nationalized banks, including the State Bank of India and its seven subsidiaries, the 14 large commercial banks taken over in 1969, and the six additional banks nationalized in 1980, accounted for more than 90 percent of aggregate deposits in commercial banks. Banking services are also provided by chit funds, which accept and pay interest on monthly deposits against which it is possible to draw only by way of loan, and by Nidhis, mutual loan societies that have developed into semibanking institutions but deal only with their member shareholders.

The main path of banking development in India is the expansion of bank branches into the under-banked areas. The authorities have sought to expand the number of branches but to avoid their concentration in the larger towns and cities and, in particular, to provide the rural areas with adequate facilities. The ultimate objective is to encourage the mobilization of deposits on a massive scale throughout the country, a formidable challenge in a country of 575,000 villages, and a stepping up of lending to weak sectors of the economy.

Japan. Banking business in Japan is largely concentrated in the hands of the big banks (some of which are specialized), though a number of small banks still survive. The principal classes of banks are city banks and regional banks, but it should be noted that the distinction has no legal basis, though they are separately supervised.

Both belong to the Federation of Bankers' Associations of Japan. The city banks service mainly manufacturing industry and commerce, particularly the big firms, while the regional banks are based on a prefecture, though some extend their operations into neighbouring prefectures, collecting deposits and lending to local business and smaller firms. The regional banks have city bank correspondents, not only to hold surplus balances but also for assistance in investing their funds, especially in the call-money market. In addition, a city bank may introduce certain of its large customers to a regional bank (*e.g.*, a big company having a local factory). City correspondents in Japan do not, however, provide the wide range of ancillary services common in the United States.

Decline of city banks

Since World War II there has been much stability in Japanese banking, but the city banks have suffered a relative decline in the importance of their business in competition with other institutions, especially the agricultural cooperatives, which attract the larger part of the Treasury's payments on account of government purchases of the rice crop. There has also been a relative increase in the importance of the life insurance companies and the trust funds, which have attracted sizable funds from the general public.

BANKING IN PLANNED ECONOMIES

The Soviet Union. The present-day Soviet banking system was established by the credit reforms of the early 1930s, which centralized practically all short-term credit in the hands of the Gosbank (State Bank, established in 1921). There was much restructuring of banking during succeeding years, mainly to ensure that the system became an effective instrument for carrying out the national economic plan. The Gosbank's control over payments flows was also tightened in order to maintain stability of prices. The activities of the Gosbank are by no means limited to purely financial operations, and it actively controls the implementation of production and financial plans and the spending of wages funds. In addition, it has a monopoly of the note issue and is responsible for putting money into circulation.

Functions of the Gosbank

The Gosbank was originally concerned with the provision of short-term credit; a number of other banks were created to finance capital investment in the socialized economy. Even in the 1930s there was a tendency to consolidate these banking units, and this continued into the postwar period. In 1959 there were further mergers and a reallocation of activity, as a result of which the Stroybank emerged as the credit and financial institution responsible for canalizing state budgetary appropriations and also short- and long-term credits into capital investment in various sectors of the economy. The savings bank system, with 79,000 branches, became part of the Gosbank in 1963. The only other banking institution is the Vneshtorgbank (Bank for Foreign Trade), whose operations were considerably expanded in 1961. Originally concerned mainly with the provision of currency for tourists and diplomatic missions and with remittances from abroad, the Vneshtorgbank came to handle all foreign-exchange transactions, including those relating to trade.

Since the functions of the Stroybank are essentially administrative (supervising the disbursement of funds), the Gosbank is the only domestic bank servicing the cash, credit, and payment needs of a population exceeding 273,000,000. The organization of the Gosbank is as follows:

Organizational structure of the Gosbank

it has a policymaking head office and principal offices in the various Soviet republics. There are also regional offices and a network of 4,500 local branches; the latter are the bank's main points of contact with a variety of economic enterprises, collective farms, and lower level government units. The Gosbank serves the needs of the urban population through its network of branches, which collect rent, taxes, and other compulsory payments and contributions. It maintains a small number of special cash service agencies in large industrial establishments and at construction projects. Seasonal agencies are operated at remote places where large purchases of farm products are made.

In the industrial area, the Gosbank services hundreds of thousands of state enterprises that operate on the basis of cost accounting. Each of these enterprises has its own working capital and prepares a balance sheet and a statement of income; it may borrow regularly or occasionally, depending on its needs. The Gosbank has hundreds of thousands of other customers comprising collective farms, party, trade union, cultural and other organizations, and individuals. The aggregate balances maintained by Gosbank customers are small in comparison with the cash balances held by business and government accounts in the United States.

Transactions of individuals. Although consumer loans are extended by stores rather than by banks, there has also been a rapid increase in the transactions of individuals. This has resulted from rising incomes, an increase in savings, and the provision of facilities for crediting wages to savings accounts and making periodic payments from them. Housing loans and tourism have also been growing in importance. As savings bank offices handle virtually all the accounts and transactions of individuals, the considerable increase in the volume of these transactions was probably the main reason for absorbing the savings bank system into the Gosbank in 1963.

Transactions of collective farms. An even greater problem has been created by the rise in transactions involving the collective farms. Before 1953 the collectives paid in kind for the services performed by the state-owned tractor and farm-machinery stations. Since 1953 these transactions have gradually shifted to payment in cash. As a first step, the tractor and machinery stations were closed down and their equipment sold to the collectives. The farms now had to pay in cash for all machinery and fuel, as well as for building materials, fertilizers, and other supplies. Subsequently, they were enabled to sell their output to the state for money.

Farm labour likewise has come to be remunerated in cash rather than in kind. In 1953 only one-third of the "compensation" for work contributed by members of the collectives was in cash; by 1963 the proportion was nearly three-quarters. In 1965 the flow of money income to the farm population was increased further by the introduction of state pensions for collective farmers. In 1966 it was decided to make minimum monthly payments to the members of collective farms, and this resulted in an even greater use of money in the farm sector.

The growth of money flows and of bank lending in rural areas, as well as the need to service a growing clientele in the villages, greatly added to the complexity of Gosbank operations, which had been geared primarily to the needs of industry and government. The Gosbank attempted to resolve its problems by simplifying payments procedures, through its efforts to work out a system of offsetting mutual claims were not initially very successful.

Interest rates in the U.S.S.R.

After the credit reforms of 1930–32, a uniform system of interest rates was applied to all short-term credits, irrespective of the purpose of the loan or the financial condition of the borrower. Higher rates were charged as a penalty on overdue loans. In the 1960s there was a move to differentiate rates, it being accepted as a matter of principle that bank funds should be more expensive than an enterprise's own working capital and that borrowings made necessary by shortcomings of management (*e.g.*, excessive inventories or the erosion of working capital) should carry higher rates than loans to finance normal needs. Penalty rates for overdue loans and for late payments were also increased, and collection was more vigorously enforced. Today, interest rates in the Soviet economy are seen as an integral part of the relationship between the bank and the enterprises that it both services and regulates.

Yugoslavia. Yugoslavia's banking system is very different from those of non-Communist countries and also from that of the Soviet Union. It derives from basic principles incorporated in the federal constitution. For example, the monetary and credit system is based on "self-management by workers in associated labour working with social resources," and it provides for uniform currency and monetary and foreign-exchange systems.

Until 1971 the central bank, or bank of issue, was the National Bank of Yugoslavia, with its head office in Belgrade and six offices in the main cities of the federal

republics. Following constitutional amendments in 1971, separate National Banks for the republics and the two autonomous provinces were established, and in 1972 a new central system was set up consisting of the National Bank of Yugoslavia and the National Banks for the republics and the autonomous provinces. Under the federal constitution of 1974 and federal laws of 1976, the National Banks have responsibility for regulating the quantity of money, for ensuring the stability of the currency and the general liquidity of payments at home and abroad, for implementing joint monetary and foreign-exchange policy, and for providing the bases of credit policy.

The nine National Banks operate as a uniform central banking system. Integration of the system is carried out though the Council of Governors of the National Bank of Yugoslavia. The Yugoslav concept of central banking thus reflects the federal structure of the nation, accommodating the principle of a single Yugoslav market, within which, however, the republics and the autonomous provinces are responsible for their own development.

Below the level of National Banks, the banking system consists of two types of organizations—the banks proper, and savings–credit organizations. The banks are classified as internal banks, basic banks, and associated banks. The savings–credit organizations include chiefly savings banks. There is also provision for banking consortia and associations.

An internal bank provides a financial nexus between its members such that funds can be transferred from sectors that are in surplus to other sectors which have need of funds. It also attracts savings deposits from workers employed by its members and from the general public; the latter deposits may be earmarked for special purposes. In addition, it may grant consumer credits. An internal bank cannot receive sight deposits. It is managed by its members.

A basic bank may be founded by organizations of associated labour, internal banks, self-managing communities of interest, and other social–legal entities; it is managed by those parties that have entered into a self-management agreement. Basic banks are the only banking organizations that can carry out all credit and banking operations, including the acceptance of sight deposits. Subject to special conditions, these banks can carry out payments and credit operations abroad. In order to maintain adequate liquidity, every such bank must hold a reserve fund with a National Bank; an additional reserve must be held in a separate account with a National Bank. Banks may pool surpluses of liquid resources as well as part of their reserve fund, and they may obtain short-term credits from the pool in order to maintain current liquidity.

The associated banks are the third form of legally constituted banking organization in Yugoslavia. Only basic banks may be members of such a bank. The associated bank carries out operations of common interest to the members of two or more basic banks—*e.g.*, credit operations and payments abroad, and operations of exceptional significance for the development of large economic complexes. The associated bank cannot accept sight deposits in home currency nor carry out operations with the general public. If the associated bank effects payments abroad, it may keep part of the resources of its reserve fund in foreign exchange, thereby ensuring its liquidity abroad.

The Yugoslav Bank of International Economic Cooperation is a specialized banking organization set up under a federal law in 1978. The basic task of the bank is to stimulate expansion of production over the long term and financial cooperation between domestic social–legal entities and foreign parties and joint action in third markets.

Savings banks may be established by organizations of associated labour and other social–legal entities to attract savings deposits and deposits in current and giro accounts from the public, to effect payments, and to grant business, consumer, or house construction or purchase credits to the public. There is also a Post Office Savings Bank, established jointly with the associations of organized labour, that conducts the whole range of savings bank business throughout Yugoslavia. It may grant credits through basic and associated banks and the National Bank of Yugoslavia or directly on the basis of their guarantees. Other savings–credit organizations are savings–credit cooperatives (founded by citizens) and savings–credit services provided by agricultural, artisan, and other cooperatives.

BIBLIOGRAPHY

History of banking: JOHN H. CLAPHAM, *The Bank of England,* 2 vol. (1944, reprinted 1966); W.F. CRICK and J.E. WADSWORTH, *A Hundred Years of Joint Stock Banking,* 3rd ed. (1958); ALBERT E. FEAVEARYEAR, *The Pound Sterling,* 2nd ed. rev. by VICTOR MORGAN (1963); MILTON FRIEDMAN and ANNA JACOBSON SCHWARTZ, *A Monetary History of the United States, 1867–1960* (1963); T.E. GREGORY and ANNETTE HENDERSON, *The Westminster Bank Through a Century,* 2 vol. (1936); BRAY HAMMOND, *Banks and Politics in America, from the Revolution to the Civil War* (1957, reprinted 1967); LLOYD W. MINTS, *A History of Banking Theory in Great Britain and the United States* (1945); HUGH NEUBERGER, *German Banks and German Economic Growth from Unification to World War I* (1977); R.D. RICHARDS, *The Early History of Banking in England* (1929, reprinted 1965); R.S. SAYERS, *The Bank of England, 1891–1944,* 3 vol. (1976); ABBOT PAYSON USHER, *The Early History of Deposit Banking in Mediterranean Europe* (1943, reprinted 1967); J.G. VAN DILLEN, *History of the Principal Public Banks* (1934, reprinted 1965).

Principles of banking: Two books that provide good general coverage of the principles of banking and finance are JOHN G. GURLEY and EDWARD S. SHAW, *Money in a Theory of Finance* (1960); and R.S. SAYERS, *Modern Banking,* 7th ed. (1967). Much useful information on the workings of the financial system is contained in GREAT BRITAIN, COMMITTEE ON THE WORKING OF THE MONETARY SYSTEM, *Report* (1959, commonly known as the "Radcliffe Report"); DAVID R. CROOME and HARRY G. JOHNSON (eds.), *Money in Britain, 1959–1969: The Papers of the Radcliffe Report—Ten Years After Conference* (1970); and the U.S. COMMISSION ON MONEY AND CREDIT, *Money and Credit: Their Influence on Jobs, Prices, and Growth* (1961). HAROLD WALLGREN, *Principles of Bank Operations,* rev. ed. (1975), is a broad survey of commercial banking activities published by the American Institute of Banking.

Central banking: For general discussions, see M.H. DE KOCK, *Central Banking,* 4th ed. (1974); and R.S. SAYERS, *Central Banking After Bagehot* (1957, reprinted 1982). For specific countries, see the BANK OF JAPAN, *The Bank of Japan: Its Organization and Monetary Policies,* 3rd ed. (1971); BANK FOR INTERNATIONAL SETTLEMENTS, *Eight European Central Banks* (1963); H.A. DE S. GUNASEKERA, *From Dependent Currency to Central Banking in Ceylon* (1962); GERHARD DE KOCK, *A History of the South African Reserve Bank, 1920–1952* (1954); E.P. NEUFELD, *Bank of Canada Operations and Policy* (1958); RESERVE BANK OF AUSTRALIA, *Reserve Bank of Australia,* 3rd rev. ed. (1979); RESERVE BANK OF INDIA, *History of the Reserve Bank of India, 1935–51* (1970); RICHARD H. TIMBERLAKE, JR., *The Origins of Central Banking in the United States* (1978); and the U.S. BOARD OF GOVERNORS OF THE FEDERAL RESERVE SYSTEM, *The Federal Reserve System: Purposes and Functions,* 7th ed. (1984).

Banking systems: For a general survey of banking systems throughout the world, see BENJAMIN HAGGOT BECKHART (ed.), *Banking Systems* (1954). A survey of the United Kingdom, the United States, and the Commonwealth countries is J.S.G. WILSON, *Banking Policy and Structure: A Comparative Analysis* (1985). A study of U.S. experience is CHARLES R. WHITTLESEY, ARTHUR M. FREEDMAN, and EDWARD S. HERMAN, *Money and Banking: Analysis and Policy,* 2nd ed. (1968). Other titles include: H.W. ARNDT and W.J. BLACKERT, *The Australian Trading Banks,* 5th ed. (1977); BANK OF JAPAN, *Money and Banking in Japan* (1973); W.F. CRICK (ed.), *Commonwealth Banking Systems* (1965); GEORGE GARVY, *Money, Banking, and Credit in Eastern Europe* (1966); BRANKO HORVAT, *The Yugoslav Economic System: The First Labor-Managed Economy in the Making* (1976); S.A. MEENAI, *Money and Banking in Pakistan,* 3rd ed. (1984); ROBIN PRINGLE, *A Guide to Banking in Britain* (1973, reissued 1975); R.S. SAYERS (ed.), *Banking in the British Commonwealth* (1952), and *Banking in Western Europe* (1962); P. BARRETT WHALE, *Joint Stock Banking in Germany* (1930, reprinted 1968); J.S.G. WILSON, *French Banking Structure and Credit Policy* (1957); and ADAM ZWASS, *Money, Banking and Credit in the Soviet Union and Eastern Europe* (1979). For articles on banking in most countries, see the London monthly *The Banker.* See also GLENN G. MUNN, *Encyclopedia of Banking and Finance,* 8th rev. ed., rev. by F.L. GARCIA (1983).

(J.S.G.W.)

Barcelona

On his visit to the city in 1862, Hans Christian Andersen remarked that Barcelona was the "Paris of Spain." The city is, indeed, a major cultural centre with a remarkable history. The capital of the autonomous region of Catalonia, it abounds with archives, libraries, museums, and buildings of interest, and it contains superb examples of modernist and Art Nouveau decor and architecture. Since the late 1970s, with the official recognition of the Catalan language and the granting of significant levels of self-government, cultural life has been revitalized, bringing a new awareness of the depth and variety of Catalan culture. This vitality combines with the striking physical setting of Barcelona, between scenic mountains and the Mediterranean Sea, its benign climate fostering street life, and its significance as an economic power and a major port, to create a city of infinite variety.

This article is divided into the following sections:

Physical and human geography

THE LANDSCAPE

The city site. Barcelona, facing the Mediterranean to the southeast, is located on a plain generally confined by the Río Besós and Río Llobregat, the rocky outcrop of Montjuich (630 feet, or 192 metres), and the semicircle of mountains, of which Tibidabo (1,745 feet) is the highest point. Throughout its past Barcelona has had to contend with the consequences of its strategic location and political significance. The city was heavily fortified until relatively recent times and did not spread much beyond its medieval confines until the 19th century, a factor that contributed to the emergence of industrial satellite suburbs and towns around the city proper. This combination of a concentrated core with a highly developed industrial belt has made Barcelona one of the most congested cities in the world.

Climate. Although Barcelona is sometimes windy, its protective semicircle of mountains shields it from the harsh, cold winds that blow out of the north and west. The average annual temperature is 61° F (16° C); January is the coldest month, averaging 49° F (9° C), and August is the hottest, at 76° F (24° C). Precipitation amounts to about 23 inches (600 millimetres) per year.

The Barrio Gótico

The city layout. At the core of the city lies the Barrio Gótico; located between the Ramblas, a series of connected boulevards, and the Via Layetana, it is a close-packed maze of narrow streets, punctuated by magnificent medieval buildings. The cathedral, episcopal palace, and churches bear witness to Barcelona's importance as a religious centre. The government buildings, such as the Palau de la Generalitat, the Casa de la Ciudad (a 14th- and 15th-century building with Baroque and Neoclassical facades), and the Palacio Real Mayor, attest to the city's importance as an administrative capital. The Roman walls survive in places largely because stretches of them were incorporated into the medieval city, and the wall built in the 13th century along the Ramblas effectively hemmed it in. The defenses that played such a large part in the battle for Barcelona during the War of the Spanish Succession (1701–14) were augmented by the construction of a citadel after the city was taken.

City fortifications

By the mid-19th century the need for elaborate defenses had passed, and the city was bursting at the seams. Accordingly, plans were devised to extend the city. The final plans were based on geometric blocks, allowing for open spaces, greenery, and social areas. The area into which the town expanded, now called the Ensanche ("extension"), was open land left originally to give a clear field of fire from the city walls. Unfortunately, the city plan was not carried out completely, and within 30 years the open areas were being exploited, causing the density of buildings to triple. Urban sprawl and uncontrolled speculative development during the Francisco Franco era added to the congestion.

The Ramblas

For the visitor, the main attraction still tends to be in the city centre, particularly around the Ramblas. The famous promenade is separated from the Ensanche by the monumental Plaza de Cataluña, and it leads down to the port and the Plaza Puerta de la Paz, where the Columbus monument stands in commemoration of the discovery of America and the explorer's announcement of it in Barcelona. The Ramblas form one of the most delightful aspects of the city, their broad treelined centre strips given over to a series of stalls and kiosks selling items such as flowers, pets, and books and newspapers.

The skyline of the modern city inevitably reflects the style of the present age, but Barcelona has always attracted distinguished and original architects. Some people find the more modern buildings along the Diagonal quite striking, but little of it can compare with the work of the Catalan sculptor and architect Antonio Gaudí, whose huge and elaborate Templo Expiatorio de la Sagrada Familia has become a symbol of the city itself. He made a number of other notable contributions, including the multistory apartment buildings Casa Batlló and Casa Milá. Other architects, such as Luis Doménechy Montaner, produced remarkable structures in the modernist style, such as the Palacio de la Música.

THE PEOPLE

Immigration has played a key part in the economic growth of the region. Up to 30 percent of the population of modern Barcelona was born outside Catalonia, a condition that has caused some social strain, given the Catalans' firm sense of national identity and their aloof attitude, which is often displayed toward the rest of Spain as a whole. In many respects, however, the city is outward-looking, conscious of cultural trends in the rest of Europe and of its historical links with other Mediterranean countries.

The Catalonian temperament

THE ECONOMY

Industry and trade. Barcelona's industry is relatively up-to-date, and the city has long-established external markets to give it stability. Almost a quarter of Spanish exports come from Catalonia, and three-quarters of Catalan industry is concentrated in the Barcelona area, which provides some 20 percent of Spain's industrial output. The textile

Spires of the Templo Expiatorio de la Sagrada Familia, the unfinished church by Antonio Gaudí in Barcelona; begun in the 1880s.

industry, set up in the second half of the 18th century, was the driving force behind the development of other major industries, such as metals and machinery, which overtook textiles in economic importance. The textile industry, however, received some impetus to modernize as a result of the agreement in 1985 for Spain to join the European Communities. Other industrial products include chemicals, pharmaceuticals, cosmetics, and leather goods.

A thriving business community

Commerce and finance. The Catalans are renowned for their business acumen. Emphasis commercially is on small firms, and among the companies registered in Catalonia few have more than 200 people on their payrolls. Nonetheless it has become policy to attract major international investors to the region. The city has a free economic zone near the port, the *zona franca,* where distribution centres are concentrated. Barcelona is a major site for conferences, exhibitions, and trade fairs. The main event, the Feria de Barcelona, has been held annually since 1929 at the Palacio de las Naciones, the exhibition centre on Montjuich. Barcelona's stock exchange is one of the most active in Spain.

Transportation. Public transportation is provided by buses, subways, and surface railways. There are also cable cars. Freeways link Barcelona to the Catalonia highway network, which joins the service up to the Cadí mountain tunnel in the Pyrenees, providing access to the French highway network. The metropolitan subway, opened in 1924, connects with the urban railway and provides regular service to the international airport at Prat de Llobregat, about eight miles southwest of the city. Connections can be made there to major world cities. The port of Barcelona accommodates ships from all parts of the world, as well as providing ferry service to the Balearic Islands and to Genoa in Italy.

ADMINISTRATION AND SOCIAL CONDITIONS

Government. Barcelona, the traditional centre of Catalan movements for independence, is the capital of both the province of Barcelona and the autonomous region of Catalonia. The city is governed as a municipality of Spain, and its elections are held every four years. The councillors elect the mayor, who selects three deputies from their number to assist in the duties of the mayor's office.

Services and education. Under the present government services have been decentralized and made more accountable to the public. Electricity is supplied from sources in France as well as elsewhere in Spain, and some hydroelectric power comes from the Pyrenees. Nuclear power plants at Vandellós, in the province of Tarragona, are of particular importance to the city. A receiving terminal for natural gas has been installed in the port to supply a regional distribution network, but most private houses rely on bottled butane gas. With the rapid growth of the population after World War II water supply has become a problem. Local rivers cannot supply both industrial and domestic needs. Some drinking water is drawn off upstream from the Llobregat, but more is provided by the Río Ter in the province of Gerona.

Water shortage

The tradition for higher education in Barcelona goes back to 1430, the year that its major school, the Universidad de Barcelona, was founded. It is one of the country's 22 state universities and one of three universities in the city. The other two are the Universidad Autónoma de Barcelona (1968) and the Universidad Politécnica de Cataluña (1971). Courses in most of the municipality's schools are taught in Spanish and Catalan.

CULTURAL LIFE

Barcelona has long been a major cultural centre. The city abounds in archives and libraries, including dozens of specialized collections, many of which are in private hands. Barcelona is, in fact, one of the major publishing centres for the Spanish-speaking world, and the Fiesta del Libro, held on April 23, is a major event for the book trade. It is also the day of Catalonia's patron saint, St. George.

Publishing centre

The more classical forms of culture are well represented. The Gran Teatro del Liceo, founded in 1847, presents opera and ballet performances; the Teatro Romea has been a focal point of Catalan-language drama since the last century. Classical music is amply provided for by the Palacio de la Música and the city's symphony orchestra. Museums range from the monumental maritime museum, which houses a full-size replica of a galley from the Battle of Lepanto in 1571, to the waxworks museum. Art of past ages is housed in the Museo de Arte de Cataluña (Romanesque and Gothic paintings) and the Federico Marés museum (12th- to 18th-century sculpture), while the Museo de Arte Moderno displays works by contemporary Catalan artists. There are also collections dedicated to famous artists connected with Barcelona, most notably the painters Joan Miró and Pablo Picasso. The Casa de Cervantes commemorates Barcelona's association with the writings of Miguel de Cervantes. Perhaps the most striking feature of culture in Barcelona is its easy availability at many levels—from major art exhibitions at the Palacio de Pedralbes to the pavement artists in the Ramblas. There are amusement parks on Tibidabo and Montjuich and a zoo in the Parque de la Ciudadela.

Museums and art galleries

History

FOUNDATION AND MEDIEVAL GROWTH

According to tradition Barcelona was founded by either the Phoenicians or the Carthaginians, who had trading posts along the Catalonian coast. It is no longer thought, however, that the city owes its name to the family of the Carthaginian leader Hamilcar Barca. In Roman times the Colonia Faventia Julia Augusta Pia Barcino did not become a centre of any real importance until the 3rd century AD. During the three centuries of Visigothic occupation the city was known as Barcinona. It became an important religious centre before the arrival of the Islāmic Moors in AD 717. Barjelūnah, as the Moors called the city, was seen as a prime objective by the Carolingian Franks, who gained control of Barcelona in 801 and, under an appointed count, established the Río Ebro on the edge of

Historical origins

Catalonia as the southerly limit of their power. In 985 the city was sacked by the forces of al-Manṣūr, chief minister of the Umayyad caliphate of Córdoba. The counts of Barcelona consolidated their influence over Catalonia in the 10th and 11th centuries, and after the union of Catalonia and Aragon in 1137 Barcelona grew into a major trading city.

Barcelona, weakened by outbreaks of plague in the 14th century, began to decline when Naples became the capital of the Catalan-Aragonese kingdom in 1442. The advent of the Habsburg monarchy, the rise of Turkish power in the Mediterranean, and the discovery of America all furthered this decline. Relations with the court in Madrid worsened in the 17th century. After 1705, when the Catalans permitted the archduke Charles III of Austria to establish his court in Barcelona, honouring his claim to the Spanish throne during the War of the Spanish Succession, Philip V of Spain besieged Barcelona. After the city fell in 1714, Philip dismantled all forms of local self-government. Ironically, this led to a period of prosperity spurred largely by the development of the cotton industry.

Philip V's capture of the city

THE MODERN CITY

Barcelona found itself occupied again, this time by Napoleon's troops, from 1808 to 1813. The war with the French left the province ravaged, but the postwar period saw the start of modern industrialization. The growth of textiles was to have a twofold effect: it led to the development of a modern industrial sector and to the emergence of Catalonia as Spain's wealthiest region. It also led to rapid population growth through immigration, largely from outside the province, and the development of class conflict between the bourgeoisie and industrial workers. Anarchist movements flourished, and the period up to the Spanish Civil War was punctuated by unrest. Notable incidents included the uprising of 1835 in which a number of convents were burned; the riots in the mid-1850s over the introduction of automated machinery; and the Setmana Tràgica (Catalan: "Tragic Week") in 1909, which led to more church burning.

On a positive note, the exhibition of 1888 attracted 400,000 visitors, and by 1900 nearly half of Spain's imports came via Catalonia. Economic strength led to the reemergence of calls for self-rule, culminating in a period of semiautonomy from 1913 to 1923. In 1931 a Catalan republic was declared in Barcelona. The following year the region attained a significant level of self-government, and it was the main centre of Republican strength when the Civil War broke out in 1936. Its fall in January 1939 led to the final surrender of the republic. Defeat brought the loss of many rights and privileges, and even the Catalan language was prohibited for a time. Only in 1977 was the Generalitat, an autonomous Catalan government, restored. Agreements with the national government, signed in 1979, outlined new areas for self-government and encouraged a wide range of developments in Barcelona.

Autonomy restored

BIBLIOGRAPHY. Among volumes on Barcelona in the series *Enciclopèdia catalana Aedos,* written in Catalan, are AUGUSTÍN DURÁN Y SANPERE *et al., Història de Barcelona* (1975); and PAU VILA DINARÉS and LLUÍS CASSASAS SIMÓ, *Barcelona i la seva rodalia al llargo dels temps* (1974). AUGUSTÍN DURÁN Y SANPERE, *Barcelona i la seva historia,* 3 vol. (1972–75), covers the history, culture, and economic aspects of the city. Later works in English and Spanish tend to include Barcelona within the scope of Catalonia overall, or as part of a broader canvas. A good starting point for geographical research is provided in MONTSERRAT GALERA MONEGAL, *Bibliografía geografica de la ciudad de Barcelona* (1973–), a multivolume work. An informative and well-illustrated article on the city is PASQUAL MARAGALL, "Barcelona," in *Gran enciclopèdia catalana,* vol. 3, pp. 188–225 (1971). ROBERT FERRAS, *Barcelone: Croissance d'une métropole* (1977), is lavishly supplied with charts, maps, and statistical data. Serviceable guidebooks include *All Barcelona,* 4th ed. (1982). For the more informed visitor there are JOSÉ PAMIAS RUIZ (ed.), *Guía urbana de Barcelona,* 2 vol., 22nd ed. (1981–82); BERTRAM STRAUSS and FRANCES STRAUSS, *Barcelona Step by Step* (1974); JAUME FABRE and JOSEP M. HUERTAS, *Tots els barris de Barcelona,* 7 vol. (1976–77). See also ALEXANDRE CIRICI, *Barcelona, City of Art* (1975; originally published in Catalan, 1973).

(T.J.Co.)

Beethoven

A universal genius widely regarded as the greatest composer who ever lived, Ludwig van Beethoven dominates a period of musical history as no one else before or since. Rooted in the Classical traditions of Haydn and Mozart, his art reaches out to encompass the new spirit of humanism expressed in the works of Goethe and Schiller, his elder contemporaries in the world of literature, and above all in the ideals of the French Revolution, with its passionate concern for the freedom and dignity of the individual. He revealed more vividly than any of his predecessors the power of music to convey a philosophy of life without the aid of a spoken text; and in certain of his compositions is to be found the strongest assertion of the human will in all music, if not in all art. Though not himself a Romantic, he became the fountainhead of much that characterized the work of the Romantics who followed him, especially in his ideal of program or illustrative music, which he defined in connection with his *Sixth* (*Pastoral*) *Symphony* as "more an expression of emotion than painting." In musical form he was a considerable innovator, widening the scope of sonata, symphony, concerto, and quartet; while in the *Ninth Symphony* he combined the worlds of vocal and instrumental music in a manner never before attempted. His personal life was marked by a heroic struggle against encroaching deafness, and some of his most important works were composed during the last 10 years of his life when he was quite unable to hear. In an age that saw the decline of court and church patronage, he not only maintained himself from the sale and publication of his works; he was also the first musician to receive a salary with no other duties than to compose how and when he felt inclined.

LIFE AND WORK

The early years. Baptized on December 17, 1770, in Bonn, northwest Germany, Beethoven was the eldest surviving child of Johann and Maria Magdalena van Beethoven. The family was Flemish in origin and can be traced back to Malines. It was Beethoven's grandfather who had first settled in Bonn when he became a singer in the choir of the Archbishop-Elector of Cologne. He eventually rose to become Kappellmeister. His son Johann was also a singer in the electoral choir; thus, like most 18th-century musicians, Beethoven was born into the profession. Though at first quite prosperous, with the death of his grandfather in 1773 and the decline of his father into alcoholism, the Beethoven family became steadily poorer. By the age of 11 Beethoven had to leave school; at 18 he was the breadwinner of the family.

Having observed in him signs of a talent for the piano, Johann had tried to make of his son a child prodigy like Mozart but without success. It was not until his adolescence that Beethoven began to attract mild attention.

When, in 1780, Joseph II became sole ruler of the Holy Roman Empire, he appointed his brother Maximilian Francis as adjutant and successor-designate to the archbishop-elector of Cologne. Under Maximilian's rule, Bonn was transformed from a minor provincial town into a thriving and cultured capital city. A liberal Roman

Beethoven, oil painting by Ferdinand Schimon, 1819. In the Beethoven-Haus, Bonn.
By courtesy of the Beethoven-Haus, Bonn, West Germany

Catholic, he endowed Bonn with a university; limited the power of his own clergy; and opened the city to the full tide of the German literary renaissance, associated with Lessing, Klopstock, and the young Goethe and Schiller. A sign of the times was the nomination as court organist of Christian Gottlob Neefe, a Protestant from Saxony, who became Beethoven's teacher. Although somewhat limited as a musician, Neefe was nonetheless a man of high ideals and wide culture, a man of letters as well as a composer of songs and light theatrical pieces; and it was to be through Neefe that Beethoven in 1783 would have his first extant composition (*Variations on March by Dressler*) published at Mannheim. By June 1782 Beethoven had become Neefe's assistant as court organist.

In 1783 he had also been appointed continuo player to the Bonn opera. By 1787 he had made such progress that Maximilian Francis, archbishop-elector since 1784, was persuaded to send him to Vienna to study with Mozart. The visit was cut short when, after only two months, Beethoven received the news of his mother's death. According to tradition, Mozart was highly impressed with Beethoven's powers of improvisation and told some friends that "this young man will make a great name for himself in the world." For the next five years Beethoven remained at Bonn. To his other court duties was added that of playing viola in the theatre orchestra; and, although the archbishop for the time being showed him no further mark of special favour, he was beginning to make valuable acquaintances. Sometime previously he had come to know the widow of the chancellor, Joseph von Breuning, and she engaged him as music teacher to two of her four children. From then on the Breunings' house became for him a second home, far more congenial than his own. Through Mme von Breuning, Beethoven acquired a number of wealthy pupils. His most useful social contact came in 1788 with the arrival in Bonn of Count Ferdinand von Waldstein, a member of the highest Viennese aristocracy and a music lover. Waldstein became a member of the Breuning circle, where he heard Beethoven play and at once became his devoted admirer. At a fancy dress ball given in 1790, the ballet music, according to the *Almanach de Gotha* (a journal chronicling the social activities of the aristocracy), had been composed by Count Waldstein; but it was generally known that Beethoven had written it for him. The same year saw the death of the emperor Joseph II. Through Waldstein again, Beethoven was invited to compose a funeral ode for soloists, chorus, and orchestra; but the scheduled performance was cancelled because the wind players found certain passages too difficult. He then added to it a complementary piece celebrating the accession of Joseph's brother Leopold II; but there is no record that either was ever performed until the end of the 19th century, when the manuscripts were rediscovered in Vienna and pronounced authentic by Johannes Brahms. But in 1790 another great composer had seen and admired them: that year Haydn, passing through Bonn on his way to London, was feted by the elector and his musical establishment; when shown Beethoven's score, he was sufficiently impressed by it to offer to take Beethoven as a pupil when he returned from London. Beethoven accepted Haydn's offer and in the autumn of 1792, while the armies of the French Revolution were storming into the Rhineland provinces, Beethoven left Bonn, never to return. The album that he took with him (preserved in the Beethoven-Haus in Bonn) indicates the wide circle of his acquaintances and friends in Bonn. The most prophetic of the entries, written shortly after Mozart's death, runs:

> The spirit of Mozart is mourning and weeping over the death of her beloved. With the inexhaustible Haydn she found repose but no occupation. With the help of unremitting labour you shall receive Mozart's spirit from Haydn's hands. (Count Waldstein.)

The compositions belonging to the years at Bonn—excluding those probably begun at Bonn but revised and completed in Vienna—are of more interest to the Beethoven student than to the ordinary music lover. They show the influences in which his art was rooted as well as the natural difficulties that he had to overcome and that his early training was inadequate to remedy. Three piano sonatas written in 1783 demonstrate that, musically, Bonn was an outpost of Mannheim, the cradle of the modern orchestra in Germany, and the nursery of a musical style that was to make a vital contribution to the classical symphony. But at the time of Beethoven's childhood, the Mannheim school was already in decline. The once famous orchestra was, in effect, dissolved after the war of 1778 between Austria and Prussia. The Mannheim style had degenerated into mannerism, which took the form of trivial and often inappropriate experimenting with dynamic contrasts as reflected even in Mozart's *Piano Sonata in C major,* K. 309. The preoccupation with extremes of piano (soft) and forte (loud), often in contradiction to the musical phrasing, is found in Beethoven's early sonatas and in much else written by him at that time—which is not surprising since the symphonies of later Mannheim composers formed the staple fare of the Bonn court orchestra. But what was for Mozart only a deviation from his normal style was to remain a fundamental element in that of Beethoven. The sudden pianos, the unexpected outbursts, the wide leaping arpeggio figures (chord notes played rapidly up or down over several octaves) known as "Mannheim rockets"—all these are central to Beethoven's musical personality and were to help him toward the liberation of instrumental music from its dependence on vocal style. Beethoven may, indeed, be described as the last and finest flower on the Mannheim tree.

Early influences. Like other composers of his generation, Beethoven was subject to the influence of popular music and of folk music, influences particularly strong in the Waldstein ballet music of 1790 and in several of his early songs and unison choruses. Heavy Rhineland dance rhythms can be found in many of his mature compositions; but he could assimilate other local idioms as well—Italian, French, Slavic, and even Celtic. Although never a nationalist or folk composer in the 20th-century sense, he often allowed the unusual contours of folk melody to lead him away from traditional harmonic procedure.

French music impinged on him from two main directions: from Mannheim, whose artistic links with Paris had always been strong; and from the Bonn Nationaltheater, which relied mainly for its repertory on comic operas translated from the French. In fashionable Bonn society, sympathy with the French Revolution was very strong, and the flavour of the French Revolutionary march is present in many of Beethoven's symphonic allegros. The jigging rhythms to be found in several of his scherzos are also clearly of French provenance.

Like all pianists of the late 18th century, Beethoven was raised on the sonatas of Carl Philipp Emanuel Bach, the chief exponent of "expressive" music at a time when music was regarded as the art of pleasing sounds. These sonatas, with their quirks of rhythms and harmony and their occasional wordless recitative, were equally familiar to Haydn and Mozart; but in Beethoven they evoked a much readier

Study with Haydn

Study with Mozart

Influence of C.P.E. Bach

response, not only for reasons of temperament, but also because of the intellectual climate in which he himself was reared. The favourite literary fare of the Breunings and their friends was associated with the *Sturm und Drang,* a reaction against the rationalism of the early 18th century, an exaltation of feeling and instinct over reason. Its gospel was enshrined in Goethe's early novel *Die Leiden des jungen Werthers* (*The Sorrows of Werther*), the language of which finds an echo in certain of Beethoven's letters and especially in the "Heiligenstadt Testament" (see below).

In such a movement music took on a new importance as an art of feeling. The sharp conflicts of mood that characterize the sonatas of C.P.E. Bach appear much more powerfully again in Beethoven; to Beethoven "feeling" was as important in practice as it was in theory to his master Neefe, who proclaimed it the only condition of artistic value. All of this does not make Beethoven a Romantic, although Romantics attempted to claim him as one of themselves. His literary world—he read widely and voraciously despite a formal education that in arithmetic had not carried him as far as the multiplication table—was rooted in the German classics, above all Goethe and Schiller. Like them he was to achieve in music a balance of form and emotion that can only be called classical.

The Bonn compositions of most enduring interest date, as might be expected, from the last years: a *Rondino* and an *Octet,* for wind instruments, composed in 1792, probably for the elector's *harmonie* (wind band); a *Trio in G Major for Flute, Bassoon, and Piano* (1791); and the two cantatas. The songs, which were doubtless written under Neefe's inspiration, show no great feeling for the solo voice. This is strange in one whose father and grandfather both had been singers, but it remained a limitation that pursued Beethoven throughout his career. Of particular interest are 24 variations on a theme by Vincenzo Righini, an Italian composer, which, like the *String Trio in E Flat Major,* Opus 3, Beethoven revised and then published at a much later date. These variations, representing a compendium of Beethoven's piano technique, for a long time were to serve as the mainstay of his repertory in the salons of Vienna.

Beethoven as pianist

Vienna. Before Beethoven left Bonn he had acquired a very considerable reputation in northwest Germany as a piano virtuoso, with a particular talent for extemporization. Mozart had been one of the finest improvisers of his age; by all accounts Beethoven surpassed him. In the age of sensibility he could move an audience to tears more easily than any other pianist of the time. For this reason especially he was taken up by the Viennese aristocracy almost from the moment he set foot in Vienna. Count Waldstein had, of course, prepared the way with his talk of a successor to Mozart; and it is significant that Beethoven's earliest patrons in Vienna were Baron van Swieten and Prince Karl Lichnowsky, who alone among the aristocracy had remained Mozart's supporters until his death. In the Vienna of the 1790s, music had become more and more the favourite pastime of a cultured aristocracy, for whom politics under the reactionary emperor Francis II were now discreditable and dangerous and who had, moreover, never shown a like appreciation of any of the other fine arts. Many played instruments themselves well enough to be able to take their place beside professionals. Probably at no other time and in no other city was there such a high standard of amateur and semiprofessional music making as in the Vienna of Beethoven's day.

As a composer, however, Beethoven still had many technical problems to overcome, and it soon became clear that Haydn was not the best person to help him. Outwardly their relations remained cordial; but Beethoven soon began taking extra lessons in secret. One of his teachers was the organist of St. Stephen's Cathedral, Johann Georg Albrechtsberger, a learned contrapuntist of the old school who equipped him with the comprehensive technique that he needed. He also studied vocal composition with Antonio Salieri, the imperial Kappellmeister. By 1794, when Haydn had left for his second visit to London, there was no longer any question of Beethoven's returning to Bonn, which was then in French hands. The elector himself had left, and consequently Beethoven's subsidy came to an end. But he had no need to worry for, apart from what he was able to earn by teaching and playing, he received free board and lodgings from Prince Lichnowsky. The year 1795 marked Beethoven's first public appearance as a pianist in Vienna. He played a concerto (No. 2, Opus 19) of his own and one by Mozart and also took part in a benefit concert for Haydn. More important still, his *Three Trios for Piano, Violin and Cello,* Opus 1, were published with a long list of aristocratic subscribers. In the next three years he undertook concert tours in Berlin and Prague and might have travelled more widely still had the international situation permitted. In 1800 he launched a public concert on the grand scale, in which one of his own piano concerti, the *Septet* (Opus 20), and *First Symphony* were given, together with works by Haydn and Mozart. The event did much to spread Beethoven's fame abroad.

The turn of the century concluded what is generally referred to as Beethoven's first period, a period during which his art stayed within the bounds of 18th-century technique and ideas. Most of his published works during that time are for the piano, alone or with other instruments, important exceptions being the *String Trio in E Flat Major,* Opus 3; the *Three String Trios,* Opus 9; the *Six String Quartets,* Opus 18; and the *First Symphony.* Beethoven extended his range slowly and methodically, but he was still a piano composer par excellence.

The "Heiligenstadt Testament"

Approaching deafness. The change in direction occurred with Beethoven's gradual realization that he was becoming deaf. The first symptoms had appeared even before 1800, yet for a few years his life continued unchanged: he still played in the houses of the nobility, in rivalry with other pianists, and performed in public with such visiting virtuosos as violinist George Bridgetower (to whom the *Kreutzer Sonata* was originally dedicated). But by 1802 he could no longer be in doubt that his malady was both permanent and progressive. During a summer spent at the (then) country village of Heiligenstadt he wrote the "Heiligenstadt Testament." Ostensibly intended for his two brothers, the document begins:

> O ye men who think or say that I am malevolent, stubborn or misanthropic, how greatly do you wrong me. You do not know the cause of my seeming so. From childhood my heart and mind was disposed to the gentle feeling of good will. I was ever eager to accomplish great deeds, but reflect now that for six years I have been in a hopeless case, made worse by ignorant doctors, yearly betrayed in the hope of getting better, finally forced to face the prospect of a permanent malady whose cure will take years or even prove impossible.

He was tempted to take his own life, "But only Art held back; for, ah, it seemed unthinkable for me to leave the world forever before I had produced all that I felt called upon to produce. . . ." There is a Werther-like postscript:

> As the leaves of autumn wither and fall, so has my own life become barren: almost as I came, so I go hence. Even that high courage that inspired me in the fair days of summer has now vanished.

More significant, perhaps, are his words in a letter to his friend Franz Wegeler: "I will seize fate by the throat. . . ." Elsewhere he remarks, "If only I were rid of my affliction I would embrace the whole world." He was to do both, though the condition he hoped for was not fulfilled.

From then on his days as a virtuoso were numbered. Although it was not until about 1819 that his deafness became total, making necessary the use of those conversation books in which friends wrote down their questions while he replied orally, his playing degenerated as he became able to hear less and less. He continued to appear in public from time to time, but most of his energies were absorbed in composing. He would spend the months from May to October in one or another of the little villages near Vienna. Many of his musical ideas came to him on long country walks and were noted in a sketchbook.

Beethoven's sketchbooks

These sketchbooks, many of which have been preserved, reveal much about Beethoven's methods of work. The man who could improvise the most intricate fantasies on the spur of the moment took infinite pains in the shaping of a considered composition. In the sketchbooks such famous melodies as the adagio of the *Emperor Concerto* or the andante of the *Kreutzer Sonata* can be seen emerging

from a trivial and characterless beginning into their final form. It seems, too, that Beethoven worked on more than one composition at a time and that he was rarely in a hurry to finish anything that he had on hand. Early sketches for the *Fifth Symphony,* for instance, date originally from 1804, although the finished work did not appear until 1808. Sometimes the sketches are accompanied by verbal comments as a kind of *aide-mémoire.* Sometimes, as in the sketching of the *Third* (*Eroica*) *Symphony,* he would leave several bars blank, making it clear that the rhythmic scheme had preceded the melodic in his mind. Many of the sketches consist merely of a melody line and a bass—enough, in fact, to establish a continuity. But in many works, especially the later ones, the sketching process is very elaborate indeed, with revisions and alterations continuing up to the date of publication. If, in general, it is only the primitive sketches and jottings that have survived, this is because Beethoven kept them beside him as potential sources of material for later compositions. The working out of a musical composition in all its detail ceased to interest him once the piece had been completed.

Beethoven and the theatre. The next few years were those of Beethoven's short-lived connection with the theatre. In 1801 he had provided the score for the ballet *Die Geschöpfe des Prometheus* (*The Creatures of Prometheus*). Two years later he was offered a contract for an opera on a classical subject with a libretto by Emanuel Schikaneder, who had achieved fame and wealth as the librettist of Mozart's *Magic Flute* and who was then impresario of the Theater an der Wien. Two or three completed numbers show that Beethoven had already begun work on it before Schikaneder himself was ousted from the management and the contract annulled—somewhat to Beethoven's relief, as he had found Schikaneder's verses "such as could only have proceeded from the mouths of our Viennese applewomen." When the new management re-engaged Beethoven the following year, it was largely on the strength of his now almost forgotten oratorio, *Christus am Ölberg* (*Christ on the Mount of Olives*), which had been given in an all-Beethoven benefit concert, together with the first two symphonies and the *Third Piano Concerto.* The year 1804 was to see the completion of the *Third Symphony,* regarded by most biographers as a landmark in Beethoven's development. It is the answer to the "Heiligenstadt Testament": a symphony on an unprecedented scale and at the same time a prodigious assertion of the human will. The work was to have been dedicated to Napoleon, one of Beethoven's heroes, but Beethoven struck out the dedication on hearing that Napoleon had taken the title of emperor. Outraged in his republican principles, he later substituted the words "for the memory of a great man." From then on the masterworks followed hard on one another's heels: the *Piano Sonata in F Minor,* Opus 57, known as the *Appassionata;* the *Piano Concerto No. 4 in G Major,* Opus 58; the three *Razumovsky Quartets,* Opus 59; the *Fourth Symphony,* Opus 60; the *Violin Concerto,* Opus 61. To this period also belongs his one opera, *Fidelio,* commissioned for the winter season of 1805. The play concerns a wife who disguises herself as a boy in order to rescue her imprisoned husband, and, in setting this to music, Beethoven was influenced by Paer and by Luigi Cherubini, composer of similar "rescue" operas and a musician whom he greatly admired. *Fidelio* enjoyed no great success at first, partly because the presence of French troops, who had occupied Vienna after the Battle of Austerlitz, kept most of the Viennese away. With great difficulty Beethoven was persuaded to make certain changes for a revival in the following spring, with modified libretto. This time the opera survived two performances and would have run longer, but for a quarrel between Beethoven and the management, after which the composer in a fury withdrew his score. It was not until eight years later that *Fidelio,* heavily revised by Beethoven himself and a new librettist, returned to the Vienna stage, to become one of the classics of the German theatre. Beethoven later turned over many other operatic projects in his mind but without bringing any to fruition.

Mature masterpieces

The established composer. During all this time, Beethoven, like Mozart, had maintained himself without the benefit of an official position—but with far greater success insofar as he had no family to support. His reputation as a composer was steadily soaring both in Austria and abroad. The critics of the Leipzig *Allgemeine musikalische Zeitung,* the most authoritative music journal in Europe, had long since passed from carping impertinence to unqualified praise, so that, although there were as yet no copyright laws to ensure a system of royalties, Beethoven was able to drive far more favourable bargains with the publishing firms than Haydn and Mozart before him or Schubert after him. Despite the restrictions on Viennese musical life imposed by the war with France, Beethoven had no difficulty in getting his most ambitious works performed, largely because of the generosity of such patrons as Prince Lichnowsky, who at one point made him a regular allowance of 600 florins a year. Others would pay handsomely for a dedication; *e.g.,* Count Oppersdorf, for the *Fourth Symphony.* Also, Beethoven's pupils included the archduke Rudolf, youngest brother of the emperor. Consequently, poverty was never a serious threat. But, doubtless because of increasing deafness combined with a habitual readiness to take offense, Beethoven's relations with the Viennese musicians, on whose cooperation he depended, became steadily worse; and in 1808, at a benefit concert where the *Fifth* and *Sixth* symphonies were first performed, together with the *Choral Fantasia,* Opus 80, there occurred a quarrel so serious that Beethoven thought of leaving Vienna altogether. But the threat of his departure was sufficient to stir his patrons into action. The archduke Rudolf, Prince Lobkowitz, and Prince Kinsky banded together to provide him with an annuity of 4,000 florins, requiring only that he should remain in Vienna and compose. The agreement remained in force until Beethoven's death, though it was to be affected by circumstances, one of which was the devaluation of 1811; although the Archduke increased his contribution accordingly, it was some time before his partners could do the same. Nevertheless, from 1809 onward Beethoven remained adequately provided for, although his habits of life often gave visitors the impression that he was miserably poor. Inevitably, his public appearances became less frequent.

Beethoven and his patrons

Beethoven and women. In this period, too, he considered more seriously than before the idea of marriage. As early as 1801, letters to his friend Wegeler refer to "a dear sweet girl who loves me and whom I love." This is thought to have been the Countess Giulietta Guicciardi, a piano pupil and the cousin of two other pupils, Therese and Josephine, daughters of Count von Brunsvik. It was to the Countess Giulietta that he dedicated the *Piano Sonata in C Sharp Minor,* Opus 27, No. 2, known as the *Moonlight Sonata.* But Giulietta married Count Gallenberg in 1803, and in later years Beethoven seems to have remembered her only with mild contempt. It seems clear, however, that he did propose marriage to her cousin Josephine, whose elderly husband, Count von Deym, died in 1804; and the understanding appears to have continued for about three years, until it was brought to an end partly by Beethoven's own indecisiveness and partly by pressure from Josephine's family. The prospective bride of 1810 is thought to have been Therese Malfatti, daughter of one of Beethoven's doctors, but, like the other marriage projects, this, too, lapsed, and Beethoven remained a bachelor. A curious item, however, was found among his effects, locked away in a drawer, at the time of his death: three letters, written but never sent, to the "Immortal Beloved." The content, which varies from high-flown poetic sentiments to banal complaints about his health and discomfort, makes it clear that this is no literary exercise but was intended for a real person. The month and day of the week are given, but not the year. The periods 1801–02, 1806–07, and 1811–12 have been proposed, but the last is the most probable. The identity of the person addressed is uncertain.

Wider recognition. In 1810 E.T.A. Hoffmann in Berlin produced an appreciation of the *Fifth Symphony,* which undoubtedly did much to launch that work on its triumphant career throughout the world and, above all, to interest the Romantics in its composer. The same year, Beethoven made the acquaintance of the writer Bettina

Brentano, the sister of the German poet and novelist Clemens Brentano and, later, wife of Achim von Arnim, the two compilers of the famous collection of German folk poetry, *Des Knaben Wunderhorn.* Of the letters that Bettina gave out as having been written to her by Beethoven, only one can be accepted as genuine; at least one of the others, in which the composer is made to philosophize on music in the most uncharacteristically romantic terms, must be dismissed as spurious. Bettina also performed the questionable service of bringing together Beethoven and Goethe at Teplitz in 1812. The admiration had been all on Beethoven's side; to Goethe, Beethoven was little more than a famous name. The meeting was not a success. "Goethe is too fond of the atmosphere of the courts," Beethoven wrote to Breitkopf and Härtel, the music publishers, "more so than is becoming to a poet. . . ." Goethe considered Beethoven to be "an utterly untamed personality, who is not altogether in the wrong in holding the world to be detestable, but surely does not make it any the more enjoyable either for himself or for others by his attitude." He showed a certain interest in the incidental music written in 1810 for *Egmont* "out of pure love for the subject."

Meeting with Goethe

The chief compositions of 1811–12 were the *Seventh* and *Eighth* symphonies, the first of which had its premiere in 1813. Another novelty at the same concert was the so-called *Battle Symphony,* written to celebrate Arthur Wellesley's (later duke of Wellington) decisive victory over Joseph Bonaparte at Vitoria. Composed originally for a mechanical musical instrument, the Panharmonicon, invented by J.N. Maelzel, Beethoven later scored the work for orchestra. He frankly admitted it was program music of the worst kind, so different from the ideals of "mehr Ausdruck der Empfindung als Malerei" ("more as an expression of feeling than painting") expressed in his own *Pastoral Symphony;* but in view of its success he was ready enough to score it for orchestra and even to send a copy of the score to the English prince regent, who, much to Beethoven's annoyance, made no acknowledgment. The concert, profitable as it was for the composer, led to a bitter quarrel with Maelzel, from which Beethoven emerged with little credit.

Despite the difficulties over the annuity caused by the devaluation of 1811, the years 1813–14 were profitable ones for Beethoven. The first performance of the *Seventh Symphony* was a huge success, and the audience insisted on the allegretto being repeated. When the Congress of Vienna assembled in 1814, Beethoven's music was universally known, and he himself was courted by the crowned heads of Europe. *Fidelio* was revived with tumultuous success, and Beethoven celebrated the fall of France with a grand patriotic cantata, *Der glorreiche Augenblick* (*The Glorious Moment*). In 1814, after years of war, Vienna was to enjoy a brief hour of glory before the Austrian economy collapsed and the city sank into a state of dowdy provincialism that lasted for nearly 40 years.

The last years. With the start of Metternich's long reign and the so-called Biedermeier period, marked by simplicity and homeliness in art and design, Beethoven's creative life entered its third and final phase. Because of his deafness he became more of a recluse than ever. His rate of composition, too, began to decrease. The works written between 1815 and 1827 comprise a mere fraction of his output after 1792; but they have a density of musical thought far surpassing anything that he had composed before. Though he now went less into society, he concerned himself more and more with business matters, not always with happy results.

At about this time he was brought in touch with the Philharmonic Society of London. Earlier, in 1803, he had been approached by the Edinburgh publisher George Thomson with a proposal that he should write sonatas based on Scottish folk tunes. Although nothing came of this, Thomson somewhat later succeeded in contracting him to arrange national folk melodies for voice, violin, cello, and piano, each with an introduction and coda. These remained an easy and profitable source of income to Beethoven for many years. It was in 1815, however, when Beethoven's pupil Ferdinand Ries settled in London and became one of the founder-members of the Philharmonic Society, that English music lovers began to take an active interest in the promotion of Beethoven's works. Another society member, Charles Neate, visited Beethoven in Vienna and later brought about the commission of three new overtures to be performed by the society. The overtures *König Stephan, Namensfeier,* and *Die Ruinen von Athen* were, however, late in arriving, and the discovery that they were not new, after all, caused considerable bad feeling; for a time, relations became strained on both sides. Ries did much to effect a reconciliation, but a visit to London, planned as early as 1813, never materialized, though Beethoven continued to hope that it would. The Philharmonic Society never ceased to interest itself in Beethoven's music and it undoubtedly played an important part in the genesis of the *Ninth Symphony,* which in a sense it commissioned. The society's archives contain an autograph of the first movement with a dedication by the composer. The first performance of the work was not, however, given in London but in Vienna, and the printed edition was dedicated to Frederick William III, king of Prussia. Beethoven, on his deathbed, received from the society a gift of £100, which moved him profoundly.

Commissions from England

In 1815 all prospects of foreign travel were cut short for Beethoven by the death of his brother Caspar Anton Carl, who left a widow, Johanna, and a son, Karl, aged nine. The will, which appointed Beethoven and the widow as joint guardians, was contested by Beethoven on the grounds of the widow's immorality; and after three years of litigation he won his case. But, for all the affection that he lavished on young Karl, Beethoven was far from being an ideal guardian. Quarrels between uncle and nephew were frequent and bitter and came to a head in 1826 when, just before sitting for his university examination, Karl attempted suicide. He recovered in a hospital, and Beethoven, on the advice of friends, agreed reluctantly that the boy should be launched on an army career. Once away from his uncle, Karl seems to have led a successful, law-abiding life. But the events of 1826 upset Beethoven profoundly and almost certainly hastened his death.

The important compositions of the final period begin with the *Two Sonatas for Piano and Cello,* Opus 102, the *Piano Sonata in A Major,* Opus 101, and the *Piano Sonata in B Flat Major,* Opus 106, the latter known as the *Hammerklavier.* Beethoven then reverted to sketches he had begun for the *Ninth Symphony.* This was broken off when the news came that the archduke Rudolf was to be appointed archbishop of Olmütz, and Beethoven decided to write a large-scale solemn mass for the installation ceremony. Work on this progressed slowly, and, like the early cantata for Joseph II, it was not completed in time for the intended occasion. Not until 1823, three years after the enthronement, was Beethoven able to send to the new archbishop the completed manuscript of the *Missa Solemnis.*

Late masterworks

In the meantime, Beethoven had written the three final piano sonatas (1820–22) and had worked desultorily on the symphonic sketches. The mass was followed by his last important piano work (completed 1823), variations on a theme that the publisher and composer Anton Diabelli had sent to a number of composers, Beethoven among them. Most of them, including Schubert and the archduke Rudolf himself, obliged; Beethoven at first declined, then changed his mind and decided to write a complete set of 33 variations himself.

The *Ninth Symphony* had begun to take shape; by the following year (1824) it was finished and was performed, together with movements from the *Missa Solemnis* and the overture from Opus 124, with great success at the Kärntnertor Theatre. The composer, who conducted the symphony's premiere, remained unaware of the applause until one of the soloists made him turn to face the audience. The *Ninth Symphony* was Beethoven's last work for large-scale forces. His final commission came in 1823 from Prince Nikolas Galitzin, who offered 50 ducats each for three string quartets. Beethoven accepted with alacrity, though only in 1825 was the first of the three, the *String Quartet in E Flat Major,* Opus 127, completed. Not two but four more followed, including an extra movement,

which was substituted for the original fugal finale (*Grosse Fuge*) of the *String Quartet in B Flat Major,* Opus 130. The last quartet was finished in 1826, about the time of Karl's attempted suicide. Beethoven spent that summer on the estate belonging to his surviving brother, Nikolaus Johann. On his return to Vienna he contracted pneumonia, from which he never fully recovered. He remained bedridden and died from cirrhosis of the liver in Vienna on March 26, 1827. The funeral three days later was attended by 20,000 people. Pallbearers included the famous pianist Hummel; Schubert was among the torchbearers; Franz Grillparzer, Austria's greatest living dramatist, wrote the funeral oration.

Death

REPUTATION AND INFLUENCE

Beethoven's achievement. Beethoven's greatest achievement was to raise instrumental music, hitherto considered inferior to vocal, to the highest plane of art. During the 18th century, music, being nonimitative, was ranked below literature and painting. Its highest manifestations were held to be those in which it had the aid of a text—that is, cantata, opera, and oratorio—the sonata and the suite being relegated to a lower sphere. A number of factors combined to bring about a gradual change of outlook: the instrumental prowess of the Mannheim Orchestra, which made possible the development of the symphony; the reaction on the part of writers against pure rationalism in favour of feeling; and the works of Haydn and Mozart. But, above all, it was the example of Beethoven that made possible the late-Romantic dictum of the English essayist and critic Walter Pater: "All arts aspire to the condition of music."

After Beethoven it was no longer possible to speak of music merely as "the art of pleasing sounds." His instrumental works combine a forceful intensity of feeling with a hitherto unimagined perfection of design. He carried to a further point of development than his predecessors all the inherited forms of music (with the exception of opera and song), but particularly the symphony and the quartet. In this he was the heir of Haydn rather than of Mozart, whose most striking achievements lie more in opera and concerto.

Three periods of work. It was his biographer Wilhelm von Lenz who first divided Beethoven's output into three periods, omitting the years of his apprenticeship in Bonn. The first period begins with the completion of the *Three Trios for Piano, Violin and Cello,* Opus 1, in 1794, and ends about 1800, the year of the first public performance of the *First Symphony* and the *Septet.* The second period extends from 1801 to 1814, from the *Piano Sonata in C Sharp Minor* (*Moonlight*) to the *Piano Sonata in E Minor,* Opus 90. The last period runs from 1814 to 1827, the year of his death. Though the division is a useful one, it cannot be applied rigidly. A composition begun in one period may often have been completed in another, hence the existence of such transitional works as the *Third Piano Concerto* and the *Second Symphony,* which belong partly to the first period and partly to the second. Again, the tide of Beethoven's maturity advanced at a rate that varied according to his familiarity with the medium in which he happened to be writing. The piano was his home ground; therefore, it is in the piano sonatas that the middle-period characteristics first make their appearance, even before 1800. The mass, on the other hand, was unfamiliar territory, so that the *Mass in C Major,* written during the same period as the *Fourth Piano Concerto* and the Razumovsky string quartets, sounds in many ways like an early work.

First period. Apart from the *First Symphony* and first two piano concerti, the works of the first period consist entirely of chamber music, most of it based on Beethoven's own instrument, the piano. All show a preoccupation with craftsmanship in the 18th-century manner. The material, for the most part, has a family likeness to that of Haydn and Mozart but, in keeping with the contemporary style, is slightly coarser and more blunt. Beethoven's treatment of the forms in current use is usually expansive. The expositions are long and polythematic; the developments are relatively short. Slow movements are long and lyrical with copious decoration. The third movement, though sometimes called a scherzo, remains true to its minuet origins, though its surface is often disturbed by unminuet-like accents. Finales are at once high-spirited and elegant. Two characteristics, however, mark Beethoven out strongly from other composers of the time: one is an individual use of contrasted dynamics and especially the device of crescendo leading to a sudden piano; the other, most noticeable in the piano sonatas, is the gradual infiltration of techniques derived from improvisation—unexpected accents, rhythmic ambiguities designed to keep the audience guessing, and especially the use of apparently trivial, almost senseless material from which to generate a cogent musical argument.

Use of dynamics and improvisation

Second period. The second period may be said to begin in the piano music with two sonatas "quasi una fantasia," Opus 27, of 1801, but in the symphony and concerto it is not fully apparent before the *Eroica* (1804) and the *Fourth Piano Concerto* (1806). Here the use of improvisatory material is more and more marked; but, whereas in the earlier period Beethoven was more concerned to show how it could fit naturally into a traditional 18th-century framework, here he explores in greater detail the logical implication of every departure from the norm. His harmony remains basically simple—much simpler, for instance, than much of Mozart's; what is new is the way it is used in relation to the basic pulse. From this Beethoven creates in his main themes an infinite variety of stress and accent, out of which the form of each movement is generated. The result is that, of all composers, Beethoven is the least inclined to repeat himself; all his works, but especially those of the middle and late period, inhabit their own individual formal world. Other characteristics of the middle period include shorter expositions and longer developments and codas; slow movements, too, become much shorter, sometimes vanishing altogether. The third movement is now always a scherzo, not a minuet, with frequent use of unexpected accents and syncopation. Finales tend to take on much more weight than before and in certain cases become the principal movement. Decoration begins to disappear as each note becomes more functional, melodically and harmonically. Another feature of these works is their immediacy. Here Beethoven's power is most evident; and the majority of the repertory works belong to this period.

Third period. The third period is marked by a growing concentration of musical thought combined with an increasingly wider range of harmony and texture. Beethoven's enthusiasm for Handel began to bear fruit in a much more thoroughgoing use of counterpoint. But he never lost touch with the simplicity of his earliest manner, so that the range of expression and mood in these last works is something that has never been surpassed. A form to which he gave more and more attention at this time was that of the variation. As an improviser he had always found it congenial, and, though some of the sets he had published in earlier years are merely decorative, he had created such outstanding examples of the genre as the finale of the *Eroica* and the *Prometheus* variations, both on the same theme. It is this type of variation that Beethoven began to pursue in his final period. A unique feature of the sets that occur in his last string quartets and sonatas is the sense of cumulative growth, not merely from variation to variation but within each variation itself. In the quartets, everything in the composer's musical equipment is deployed—fugue; variation; dance; sonata movement; march; even modal and pentatonic, or five-tone, melody.

Use of variations

Structural innovations. Beethoven remains the supreme exponent of what may be called the architectonic use of tonality. In his greatest sonata movements, such as the first allegro of the *Eroica,* the listener's subconscious mind remains oriented to E-flat major even in the most distant keys, so that when, long before the recapitulation, the music touches on the dominant (B flat), this is immediately recognizable as being the dominant. Of his innovations in the symphony and quartet, the most notable is the replacement of the minuet by the more dynamic scherzo; he enriched both the orchestra and the quartet with a new range of sonority and variety of texture.

The same is true of the concerto, in which, strictly

speaking, he introduced no formal innovations, the entry of solo instrument before an orchestral ritornello in the *Fourth* and *Fifth* piano concerti having been already anticipated by Mozart. Although, in the finale of the *Ninth Symphony* and the *Missa Solemnis,* Beethoven shows himself a master of choral effects, the solo human voice gave him difficulty to the end. His many songs form, perhaps, the least important part of his output. His one opera, *Fidelio,* owes its pre-eminence to the excellence of the music, rather than to any real understanding of the operatic medium. But even this lack of vocal sense could be made to bear fruit, in that it set his mind free in other directions. A composer such as Mozart or Haydn, whose conception of melody remained rooted in what could be sung, could never have written anything like the opening of the *Eroica,* in which the melody takes shape from three instrumental strands each giving way to the other. Wagner was not far wrong when he hailed Beethoven as the discoverer of instrumental melody.

Influence on piano music

Beethoven holds an important place in the history of the piano. In his day, the piano sonata was the most intimate form of chamber music that existed—far more so than the string quartet, which was often performed in public. For Beethoven, the piano sonata was the vehicle for his most bold and inward thoughts. He did not anticipate the technical devices of such later composers as Chopin and Liszt, which were designed to counteract the percussiveness of the piano, partly because he himself had a pianistic ability that could make the most simply laid-out melody sing; partly, too, because the piano itself was still in a fairly early stage of development; and partly because he himself valued its percussive quality and could turn it to good account. Piano tone, caused by a hammer's striking a string, cannot move forward, as can the sustained, bowed tone of the violin, although careful phrasing on the player's part can make it seem to do so. Beethoven, however, is almost alone in writing melodies that accept this limitation, melodies of utter stillness in which each chord is like a stone dropped into a calm pool. Beethoven was less successful in combining the piano with one other instrument, and his duo sonatas remain on a slightly lower level. But it is above all in the piano sonata that the most striking use of improvisatory techniques as an element of construction is found. Among later composers it was chiefly Liszt who extended Beethoven's principle of transferring structural weight from the first movement to the finale, making it the basis of his symphonic poems as well as of his two concertos. The two works of Beethoven that undoubtedly had most influence over succeeding generations were the *Fifth* and *Ninth* symphonies, with their progression from storm and stress to triumph. Brahms' *Symphony No. 1 in C Minor,* Tchaikovsky's *Symphony No. 5 in E Minor,* César Franck's *Symphony in D Minor,* and Mahler's *Symphony No. 2 in C Minor* are all examples of Beethoven's spiritual progeny, though few will grant that they equal, let alone surpass, their models.

MAJOR WORKS

Orchestral music

SYMPHONIES: *No. 1 in C Major,* op. 21 (1800); *No. 2 in D Major,* op. 36 (1802); *No. 3 in E Flat Major,* op. 55 (*Eroica;* 1804); *No. 4 in B Flat Major,* op. 60 (1806); *No. 5 in C Minor,* op. 67 (1808); *No. 6 in F Major,* op. 68 (*Pastoral;* 1808); *No. 7 in A Major,* op. 92 (1812); *No. 8 in F Major,* op. 93 (1812); *No. 9 in D Minor,* op. 125 (*Choral;* 1824). *Wellington's Victory,* op. 91 (also known as *The Battle of Vitoria* and the *Battle Symphony;* 1813).

CONCERTI: (PIANO): *"No. 1" in C Major,* op. 15 (1798), *"No. 2" in B Flat Major,* op. 19 (in fact composed first 1795, revised 1798); *No. 3 in C Minor,* op. 37 (1800); *No. 4 in G Major,* op. 58 (1806); *No. 5 in E Flat Major,* op. 73 (*Emperor;* 1809). (VIOLIN): *Violin Concerto in D Major,* op. 61 (1806); *Triple concerto in C Major,* op. 56 (violin, cello, piano; 1804).

OTHER ORCHESTRAL COMPOSITIONS: 2 romances for violin and orchestra; various overtures, including *Coriolan,* op. 62 (1807); *Leonore No. 1,* op. 138, *2,* op. 72A and *3,* op. 72B.

Chamber music

STRING QUARTETS: No. 1–6, op. 18 (1798–1800); No. 1-3, op. 59 (*Razumovsky;* 1806); op. 74 (*Harp* 1809); op. 95 (1810); and the late quartets (1824–26); op. 127, 130, 131, 132, 133 (*Grosse Fuge,* originally the finale to 130) and op. 135.

OTHER CHAMBER WORKS: *Octet,* op. 103 (winds; 1792); *Septet* (strings and wind; 1800); *Sextet for Horns and String Quartet,* op. 81B (1795); *Quintet for Piano and Winds,* op. 16 (1796); *String Quintet in C Major,* op. 29 (1801); 7 piano trios; 5 string trios; 10 sonatas for violin and piano, including *Sonata in A Major (Kreutzer;* 1803); 5 sonatas for cello and piano; sonata for horn and piano.

Piano music

32 sonatas, including *Sonata in C Sharp Minor,* op. 27, no. 2 (*Moonlight;* 1801); and *Sonata in F Minor,* op. 57 (*Appassionata;* 1804); 3 sets of Bagatelles; 20 sets of variations; 4 rondos.

Vocal music

Missa Solemnis (mass in D major; 1823); *Mass in C Major,* op. 86 (1807); *Christus am Ölberg* (oratorio 1803); various smaller works for chorus and orchestra including *Choral Fantasia,* op. 80 for piano, chorus, and orchestra (1808); songs, including the cycle *An die ferne Geliebte,* op. 98 (1816), and Goethe and Gellert settings; Scottish, Irish, and Welsh folk-song settings.

Theatre music

One opera, *Fidelio* (1805; revised versions, 1806, 1814—the final version is the one usually heard today); one ballet, *Die Geschöpfe des Prometheus* (1801); incidental music to four plays; *Egmont,* op. 84 (1810), *Die Ruinen von Athen,* op. 113 (1811), *König Stephan,* op. 117 (1811), *Die Weihe des Hauses,* op. 124 (1822). (J.M.Bu.)

BIBLIOGRAPHY

Works: A multivolume complete edition has been begun by the Beethoven Archives in Bonn: LUDWIG VAN BEETHOVEN, *Werke,* ed. by JOSEPH SCHMIDT-GÖRG (1961–). The important works for the most part have opus numbers allocated by Beethoven himself. Lists of those of Beethoven's works without opus numbers (*Werke ohne Opuszahl*) may be found in the catalogs of Kinsky and Hess: GEORG KINSKY, *Das Werk Beethovens: Thematisch-bibliographisches Verzeichnis seiner sämtlichen vollendeten Kompositionen,* ed. by HANS HALM (1955), supplemented by KURT DORFMÜLLER (ed.), *Beiträge zur Beethoven-Bibliographie: Studien und Materialen zum Werkverzeichnis von Kinsky-Halm* (1979); and WILLY HESS, *Verzeichnis der nicht in der Gesamtausgabe veröffentlichten Werke Ludwig van Beethovens* (1957). None of the above is complete in its information. ALAN TYSON, *The Authentic English Editions of Beethoven* (1963), is a bibliographic study with facsimiles and musical illustrations; KURT E. SCHÜRMANN (ed.), *Ludwig van Beethoven: Alle vertonten und musikalisch bearbeiteten Texte* (1980), is a compilation of texts of Beethoven's vocal works; a discography, with reviews, is *The Recordings of Beethoven as Viewed by the Critics from High Fidelity* (1971, reprinted 1978); a collection of facsimiles and transcriptions in modern musical notation is presented in LUDWIG VAN BEETHOVEN, *Autograph Miscellany from Circa 1786 to 1789,* ed. by JOSEPH KERMAN (1970).

Letters and conversation books: *The Letters of Beethoven,* 3 vol., collected, trans., and ed. by EMILY ANDERSON (1961), is the standard edition of Beethoven's letters; a selection from these has been issued, with additional notes by ALAN TYSON, in *Selected Letters of Beethoven* (1967). *New Beethoven Letters,* trans. and annotated by DONALD W. MACARDLE and LUDWIG MISCH (1957); and *Beethoven Letters in America,* ed. by OSCAR G. SONNECK (1927), are other selections. Also see GEORG SCHUNEMANN, *Ludwig van Beethovens Konversationshefte,* rev. and enl. ed. by KARL-HEINZ KOHLER and GRITA HERRE, 8 vol. (1968–81). The Conversation Books represent Beethoven's only way of keeping contact with his friends after the onset of complete deafness; they represent mostly his friends' side of the conversation. A compendium is LUDWIG VAN BEETHOVEN, *Letters, Journals, and Conversations,* ed. by MICHAEL HAMBURGER (1966, reprinted 1977).

Life: ALEXANDER W. THAYER, *The Life of Ludwig van Beethoven,* 2 vol., ed. by ELLIOT FORBES (1964, reissued in 1 vol., 1973), is the standard biography, representing the third completed edition, revised and brought up-to-date, of Thayer's original work. It is, however, considerably condensed; and students are recommended to consult in addition the earlier American edition, *The Life of Ludwig van Beethoven,* ed. and trans. by HENRY E. KREHBIEL, 3 vol. (1921); and the German *Ludwig van Beethovens Leben,* 5 vol., ed. by HERMANN DEITERS and HUGO RIEMANN (1907–11, reissued 1917–23). ANTON F. SCHINDLER, *Biographie von Ludwig van Beethoven,* 3rd rev. ed. (1860; trans. into English by CONSTANCE S. JOLLY as *Beethoven As I Knew Him,* ed. by DONALD W. MACARDLE, 1966, reprinted 1972), has the value of a detailed life written by someone who knew the composer intimately, and its errors and distortions of fact are corrected in some excellent annotations. *Beethoven: Impressions of Contemporaries,* ed. by OSCAR G. SONNEK (1926, reissued 1967), is a useful anthology of opinions and accounts given by those with whom Beethoven came into contact. GEORGE R. MAREK, *Beethoven: Biography of a Genius* (1969, reissued 1972), gives a balanced, readable, and convincing ac-

count of the composer's life without going into as much detail as Thayer. EDITHA and RICHARD STERBA, *Beethoven and His Nephew: A Psychological Study of Their Relationship*, trans. by WILLARD R. TRASK (1954, reissued 1971), is a controversial exercise in posthumous psychoanalysis. H.C. ROBBINS LANDON (comp.), *Beethoven: A Documentary Study* (1970; abridged ed. 1975; originally published in German, 1970), is a commemorative scholarly study, with documents and colour illustrations; THOMAS K. SCHERMAN and LOUIS BIANCOLLI (eds.), *The Beethoven Companion* (1972), is an anthology combining biography, analysis, reminiscences, and letters; MARTIN COOPER, *Beethoven: The Last Decade 1817–1827* (1970), includes an account of his medical history; PETER PÖTSCHNER, *Das Schwarzspanierhaus: Beethovens letzte Wohnstätte* (1970), is a description, with illustrations, of the composer's last home in Vienna; JOSEPH SCHMIDT-GÖRG, *Beethoven: Die Geschichte seiner Familie* (1964); and JOSEPH SCHMIDT-GÖRG and HANS SCHMIDT (eds.), *Ludwig van Beethoven* (1974), is a commemorative pictorial biography; FRITZ ZOBELEY, *Portrait of Beethoven* (1972), is an illustrated biography based on contemporary research; MAYNARD SOLOMON, *Beethoven* (1977), is a psychoanalytical approach to Beethoven's life and music; MUNDANEUM, *Beethoven: Biographies* (1972), is one in a series of scholarly bibliographies prepared at the National Bibliographic Centre of Belgium.

Studies of the music: DONALD F. TOVEY, *Beethoven*, ed. by HUBERT J. FOSS (1944, reprinted 1975), is a series of penetrating essays on various aspects of Beethoven's work; it was intended to form the basis of a full study, which the author never lived to complete. Equally valuable is his *Companion to Beethoven's Pianoforte Sonatas* (1931, reprinted 1976), which provides a close structural analysis of all 32 works. ERIC BLOM, *Beethoven's Pianoforte Sonatas Discussed* (1938, reissued 1968), is compiled from a set of notes written for the famous recordings made by Artur Schnabel. CARL CZERNY, *On the Proper Performance of All Beethoven's Works for the Piano* (1970), is a work by a great pianist and teacher, edited by another great pianist, PAUL BADURA-SKODA; KENNETH DRAKE, *The Sonatas of Beethoven As He Played and Taught Them* (1972, reprinted 1981), is an interpretative analysis of the sonatas; WILFRID MELLERS, *Beethoven and the Voice of God* (1983), is also devoted to the piano sonatas. See also WILLIAM NEWMAN, *Performance Practices in Beethoven's Piano Sonatas* (1971); and RUDOLPH R. RETI, *Thematic Patterns in Sonatas of Beethoven* (1967). JOSEPH KERMAN, *The Beethoven Quartets* (1967, reprinted 1982), offers a comprehensive and stimulating treatment of the music.

PHILIP RADCLIFFE, *Beethoven's String Quartets*, 2nd ed. (1978), is a shorter study but very concentrated; ROBERT WINTER, *Compositional Origins of Beethoven's Opus 131* (1982), is another study of the string quartets. In the symphonic field GEORGE GROVE, *Beethoven and His Nine Symphonies*, 3rd ed. (1898, reprinted 1962), is an established classic; ANTONY HOPKINS, *The Nine Symphonies of Beethoven* (1981), is a later study that provides structural and harmonic analysis of every movement, with illustrations; LIONEL PIKE, *Beethoven, Sibelius, and the "Profound Logic"* (1978), is a comparative study. For a great musician's opinion, see HECTOR BERLIOZ, *A Critical Study of Beethoven's Nine Symphonies, with a Few Words on His Trios and Sonatas, a Criticism of Fidelio, and an Introductory Essay on Music* (1913, reprinted 1976), a translated selection from the author's *A Travers Chants* (1898). The only Beethoven opera is discussed in the English National Opera guide *Fidelio* (1980), which includes the libretto in the original German, an English translation, a critical essay, and a bibliography and discography. IRVING SINGER, *Mozart and Beethoven: The Concept of Love in Their Operas* (1977), explores moral, aesthetic, and erotic concepts in the music of the Romantic composer. ROBERT WINTER and BRUCE CARR (eds.), *Beethoven, Performers, and Critics* (1980), is the material of the International Beethoven Congress of 1977. ALAN TYSON (ed.), *Beethoven Studies* (1973), *Beethoven Studies 2* (1977), and *Beethoven Studies 3* (1982), are collections of scholarly essays on the composer's music and life. IRVING KOLODIN, *The Interior Beethoven: A Biography of the Music* (1975), explores the development of Beethoven's musical ideas and their influence on others. DAVID B. GREEN, *Temporal Processes in Beethoven's Music* (1981), examines musical form and aesthetics. DENIS ARNOLD and NIGEL FORTUNE (eds.), *The Beethoven Reader* (1971), is a collection of essays. CHARLES ROSEN, *The Classical Style*, rev. ed. (1976), is a blend of musical, literary, and art criticism; GERALD ABRAHAM (ed.), *The Age of Beethoven: 1790–1830* (1982), is a history of the music of the period. PAUL NETTL, *Beethoven Encyclopedia* (1956, reprinted 1975 as *Beethoven Handbook*), is a reference source. DONALD W. MACARDLE, *Beethoven Abstracts* (1973), provides an index to and summarizes the Beethoven literature.

Animal Behaviour

Animal behaviour includes any activity of an intact organism. A living animal behaves constantly in order to survive, and all animals must solve the same basic problems. They must, for instance, periodically replace their energy source (consume food), avoid dehydration (drink), avoid becoming another animal's energy source (avoid being eaten), maintain their body surfaces (clean and groom), and reproduce. This article discusses these and other basic behavioral activities of animals ranging from protozoans to higher vertebrates. It not only treats the behavioral patterns of individuals but also considers those of animals in groups. Coverage of the latter includes reciprocal altruism, communication, and various other factors involved in social interaction. Although references are made to human behaviour, the reader should consult the articles BEHAVIOUR, THE DEVELOPMENT OF HUMAN and BEHAVIOUR, INNATE FACTORS IN HUMAN for specific information. Likewise treated in passing are certain behavioral tendencies of plants that resemble or parallel simple unlearned behavioral responses and adaptive mechanisms of animals (see also GROWTH AND DEVELOPMENT: *Plant development*).

This article is divided into the following sections:

NATURE AND PATTERNS OF ANIMAL BEHAVIOUR

Diversity of behavioral activity

Any animal may be regarded as an agglomeration of interacting and interdependent structures and behaviours that are responses to environmental conditions. The behavioral features of modern animals are the accumulated results of millennia of selective pressures acting on small variations inherent in individuals. This selection is relentless because environments are constantly changing.

An understanding of comparative behaviour is helpful in understanding human behaviour, just as an understanding of comparative anatomy is helpful in understanding human anatomy. The reason for this is that both behaviour and anatomy have a genetic basis. All vertebrates, for example, share certain anatomical features that distinguish them as vertebrates; and smaller groupings such as fish, amphibians, reptiles, birds, and mammals may be distinguished from one another on the same basis. This type of distinction holds true between species and even between individuals within a species. Consequently, an understanding of the anatomy or behaviour of any species is helpful in understanding other species, including man himself. In general, the closer the relationship between any two species, the more similar are the structures and behaviours of the two species. The converse is also true. Exceptions, however, do exist. Human beings and chimpanzees, for instance, are closely related genetically, but, because of historic differences in environment, the behaviour of humans is, in many ways, more like that of wolves, which experience many problems similar to those of ancient man. Such convergences and divergences are commonplace in biological evolution. Convergence occurs when unrelated animals independently evolve similar responses to similar environmental conditions—*e.g.,* the similar body shapes of porpoises and sharks; the similar social behaviour of wolves and humans (see below *Behaviour of animals in groups*). Divergence occurs when closely related species are adapted to different conditions, with a resultant difference in behaviour and structure. This is the usual type of response; sometimes, however, divergence is extreme enough to obscure a close relationship. The males of many species of closely related hummingbirds, birds of paradise, pheasants, and ducks, for example, are superficially so different from one another that many of these species were formerly assigned to different genera.

Behaviour as a taxonomic aid

The study of behaviour has provided valuable information about relationships among animals. The Greek philosopher Aristotle was one of the first to use behaviour as a taxonomic aid, but only in recent times have behavioral features been important in animal taxonomy. Aristotle regarded pigeons and doves as closely related to the sand grouse, basing his view partly on their similar way of drinking. Pigeons, doves, and sand grouse, unlike most other birds, keep the bill in the water and drink with a pumping action.

Behaviour may be quite simple, as in taxis (movement toward or away from a stimulus) and kinesis (undirected response proportional to the intensity of a stimulus). These two types of behaviour—most often descriptive of invertebrates—may be further subdivided. Orthokinesis, for example, is a response that involves change in the speed of movement of the body as a whole. Klinokinesis involves changes in the rate of turning from side to side. Klinotaxis is a type of orientation to stimuli in which, in alternate body movements, external stimuli are received with equal intensity. In tropotaxis the orientation of the animal is similar to that in klinotaxis, but it depends upon stimuli acting simulaneously upon two receptors or upon two parts of one receptor. These are stimulated unequally if the animal is not oriented directly toward or away from the source of stimulation. In telotaxis the animal orients to one or the other of conflicting stimuli affecting the same sensory mechanism. In menotaxis, or light compass response, animals (*e.g.,* honeybees, ants) do not orient either directly away from or toward a source of stimulation but assume a constant angle to the direction of the stimulus. Complex behaviours such as nest building, courting, and fighting do not lend themselves to the simple labelling of taxes or kineses and are classified according to other systems, which are dealt with below.

Classification of behaviour

TYPES

An animal's behaviour may be described according to the nature of the muscular contractions involved or in terms of the consequences of the behaviour. Because the muscle contractions involved in specific behaviours are often complex, a kind of shorthand is commonly used to describe them. Terms such as tail flick, head bob, and threat posture are of this nature. In describing behaviour in terms of its consequences, terms such as avoidance, courtship, nest building, and burrowing are often employed. Sometimes a behaviour must be described both in terms of the type of muscle contraction and in terms of the consequences of the behaviour.

Types of classification

Three types of behavioral classification are used most commonly: (1) on the basis of immediate causation, (2) on the similarity of evolutionary history, and (3) on the similarity of function.

Immediate causation. Classification based on immediate causation requires that the causal factors first be identified. All behaviours triggered by similar causal factors are then grouped together. Whether or not behaviours share the same causal factor can be determined by either of two methods. One method is to administer the causal factor and see if all of the behaviours are elicited and affected similarly. The other method, used when the causal factor is not known, is to examine the chronological correlations between the activities in question. Two activities that consistently occur together are likely to be causally related. This method is often used in studies of agonistic (attack–escape) courtship, feeding, egg laying, and similar complex behaviours.

Evolutionary history. Types of behavioral evolution are often classified according to similarities in biological evolution. Similarities between patterns of muscular contractions are compared among species believed to be related. The degree of difference between presumably homologous behaviours provides a criterion for measuring the degree of evolutionary affinity and for determining the direction of evolutionary changes. Data useful in classification may sometimes also be obtained by observing the appearance or disappearance of behaviours during ontogeny, or individual development.

Function. Classification based on similarity of function depends on the identification of behaviours with similar evolutionarily adaptive values. Such behaviours may or may not be homologous. The flying behaviour of a bird, for example, is not homologous with that of a butterfly, but both have similar functional and adaptive significance and may be classified together on this basis. A functional classification often closely agrees with a causal classification; this is because evolutionarily functional and causal mechanisms are commonly associated. The more distantly animals are related, the less likely it is that their functional and causal classifications will overlap; the converse also tends to be true: the more closely animals are related, the more likely it is that the two classifications will overlap.

In the case of behaviour in juveniles and adults of the same species, causal and functional categories may sometimes not overlap; an example is penis erection in juvenile mammals, in which reproduction is not a causal factor, compared with that in adults, in which it is.

THE INFLUENCE OF GENETICS AND EXPERIENCE

nnate versus learned ehaviour

Behaviour is sometimes classified according to the nature of the changes occurring during evolution or ontogeny. From these bases groupings such as learned, innate (*i.e.,* unlearned, or instinctive), and ritualized behaviour are derived. This classification scheme is useful in studies such as those dealing with the acquisition of communication behaviour.

Every aspect of an animal—behavioral as well as structural—is influenced to some degree by heredity as well as by experience. Although a muscle can become larger, harder, and more vascularized (*i.e.,* supplied with blood vessels) or smaller, softer, and less vascularized depending upon the nature of the experience to which it is subjected, the possibility of such modification is not infinite because of genetic limitations. Some behaviours are inherited in the same sense that an organ is inherited, and they exhibit little or no capacity to be modified by experience. Other types of behaviour achieve definitive form and employment only with experience.

Various intermediate conditions exist. If the environment is predictable relative to the appropriateness of a given behaviour in an animal, it is biologically advantageous for this behaviour to be innate, because presumably there is no need for a variety of response. Variability would, in fact, be disadvantageous, because an inappropriate response could threaten the animal's chance of survival. On the other hand, if the environment is unpredictable in certain respects, it is biologically advantageous to have some degree of inherent variability based upon response to experience. In this case, however, it is likely that the animal would sometimes respond with the wrong behaviour, resulting in a lowered survival value. In behaviours that may be modified by experience, the nature and extent of the modification are not limitless. The nature, intensity, timing, and duration of experience that may influence a given behaviour vary with the type of behaviour involved. The kind of response to these aspects of experience is determined genetically and has evolved to the extent that the maximum possible survival value is achieved.

Some behaviours are innate in the same sense that an organ of the body is innate. Their functions are relatively fixed and predictable. Other behaviours have an inherent resiliency and may respond radically to changes in the environment. Genetically determined limits, however, are always imposed upon even the most variable behaviours. An animal would otherwise be capable of limitless modifications of its behaviour, within its structural limitations, and be able to associate arbitrarily any environmental stimulus with any behaviour.

A general evolutionary trend exists for more and more behaviours to be modified by environmental stimuli as the phylogenetic scale ascends (*i.e.,* as the animal becomes more complex and "advanced"). This tendency has reached its extreme expression in vertebrates, particularly mammals, and especially in humans.

Components of behaviour

FIXED ACTION PATTERNS

A behaviour that is independent of environmental stimuli for its form is known as a fixed action pattern (FAP). An environmental stimulus may, however, be responsible for the elicitation and proper orientation of the FAP and may have an influence on the completeness of the response. Common examples of FAP's include displays (visible and audible signals), nest-building movements, various food-gathering and food-preparation movements, thermoregulatory movements, and attack and escape movements.

Species specificity of FAP's

Because FAP's are often specific for particular species, they are frequently useful in taxonomic and evolutionary studies. The homologous relationship among FAP's of related species is often easily determined, and their qualitative and quantitative differences can be evaluated.

FAP's are ordinarily quite constant in form, but this stereotypy is not a defining characteristic. Some vary considerably in the degree of completeness, even though the proportions of the components of the response may remain quite constant, relative to one another. Some FAP's share a single environmental stimulus. Sight of a rival male, for example, may elicit flight, attack, or any of a variety of agonistic (attack–escape) displays.

It is sometimes difficult to distinguish an FAP from its orienting movements. In such cases the eliciting stimulus can be manipulated, and the subsequent effects on the behaviour can be observed. The graylag goose (*Anser anser*), for example, retrieves eggs displaced from the nest by means of a highly stereotyped behaviour. While sitting on the nest, the goose extends its head beyond the egg so that the undersurface of the bill is against it. The head is then pulled toward the body until the egg is again safely in the nest. While the head is being pulled toward the body, it makes balancing adjustments that compensate for the tendency of the egg to roll to either side. If the egg is removed during retrieval, the head continues its movement toward the nest, but compensatory movements to counter the erratic roll cease. The FAP is elicited by the egg-out-of-nest stimulus and, once triggered, goes to completion whether or not the egg is still balanced against the underside of the bill. The balancing movements, however, depend upon continuing stimulation by the egg itself and are not part of the head-withdrawing FAP. The balancing movements also cease if a smoothly rolling cylinder is substituted for the egg.

This egg-retrieving FAP illustrates other features of FAP's in general. Every species can perform a finite number of FAP's and have a limited capacity, or none at all, for developing new ones. The limitations may be determined by physical structure: a woodchuck cannot fly, nor can a pelican burrow. Yet neural organization may provide restrictions as ubiquitous and rigid as anatomical ones. The egg-retrieving graylag goose is physically capable of retrieving an egg with its broadly webbed feet or with

a wing used as a kind of broom. This never happens, however, because the goose's neural organization permits egg retrieving only in the manner described. Alternative behaviours can sometimes be employed. Parrots of the genus *Agapornis,* for instance, always scratch the head by bringing a foot forward over the wing on the same side; however, when a foot is brought forward for cleaning the bill, it always passes below the wing on the same side. The bird has the physical structure as well as the neural organization necessary for both movements, but each method is specific for its particular stimulus.

KEY STIMULI

An animal reacts to relatively few of the stimuli present in its environment. This is a basic characteristic of behaviour. It is seldom known whether an animal actually perceives but does not react to the various available stimuli or whether it fails to perceive stimuli of no significance to it—at least of no significance within a given context. Both situations are probably true, at least for some species, under given circumstances. The human eye, for instance, focusses on the retina a detailed picture of all that the eye is directed toward. The human observer, however, is unaware of much within his field of vision. The same is true of other sensory modalities. At some point, insignificant stimulation is filtered out. An animal could not function if it had to respond to all of the stimuli its sensory organs are capable of receiving at any moment. Each species has, therefore, evolved responses only to those stimuli significant to itself and at such times that responses to such stimuli are relevant. This simpler world that actually falls within the animal's perception at any particular moment is termed its *Umwelt.*

Umwelt

The tick's response to its relatively simple *Umwelt* graphically illustrates how an animal selectively responds to only those stimuli pertinent to its immediate requirements. The mature female tick responds to light falling on her photosensitive body surface by moving to the tip of a twig or some suitable substitute, where she waits for a mammal to approach. The tick is capable of waiting for years until that occurs. When a mammal passes close by, she releases her grip on the twig and, if successful, falls onto the body of the mammal. The key stimulus causing her to release her grip is the scent of butyric acid, which naturally issues from the body of any mammal. When the female tick is on the host's body, she reacts to its body heat by inserting her proboscis (feeding organ) into the mammal's skin and sucking herself full of blood. This behavioral sequence depends on three stimuli: light, the odour of butyric acid, and warmth. The tick will relinquish its perch for anything with the scent of butyric acid, sink its proboscis into any warm surface—even a balloon filled with warm water—and fill herself with whatever fluid is within. The taste of blood and mammalian characteristics are not meaningful to the tick.

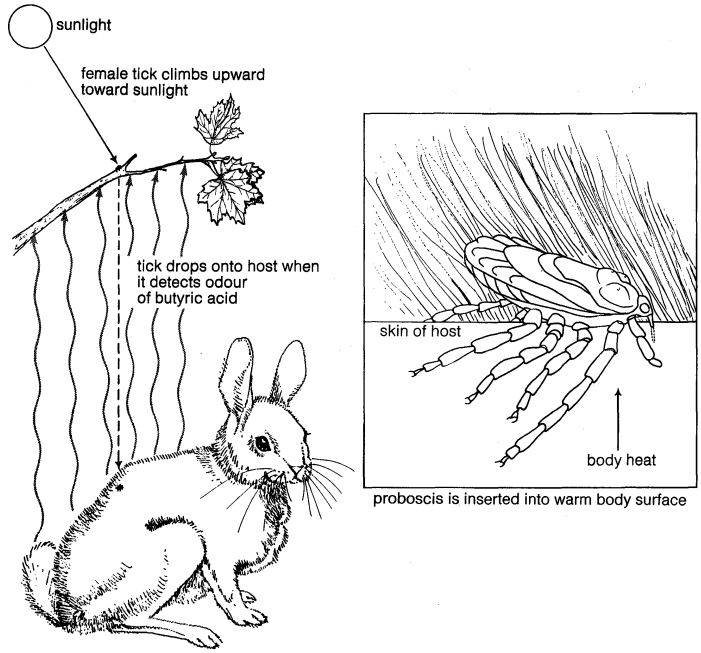


Figure 1: Response of the tick to environmental stimuli. Before obtaining a blood meal from the proper host, it must successively react to the stimuli of light, butyric acid, and body warmth.

Aggressive behaviour in the male stickleback fish (*Gasterosteus aculeatus*) occurs when the fish sees the red belly of other males. Crude dummies are attacked as long as they have red bellies; realistic models or actual fish without red bellies are not attacked.

Various species of predacious animals utilize lures that simulate the natural foods of prey species. The alligator snapping turtle (*Macrochelys temmincki*) has two slender, reddish structures projecting from the tip of its tongue that move like small worms. The turtle, with its mouth open, rests on a river bottom. A fish attracted by the false worm is snapped up by the turtle. Several species of deep-sea anglerfish have a long slender projection from the forepart of the dorsal fin. This "fishing rod" terminates in a wriggling wormlike structure that is dangled close to the anglerfish's mouth. When another fish investigates the lure, it is easily snapped up by the anglerfish. Different species of anglerfish possess different lures that are specific for different prey species.

The male swordtail characin fish attracts female mates by means of an organ resembling a daphnia, a favourite prey. The gill covers of the male are modified into a long, slender projection terminating in this "daphnia," which is moved within view of the female. When she has been lured close enough, the male copulates with her.

Female fireflies of the genus *Photurus* lure males of other genera by imitating their flashing codes, then seize and eat them. A certain petal in various species of fly orchids is modified to resemble the females of various wasp species, and the odour of these female wasps is also imitated. When a male wasp is attracted to a model female of the same species and attempts to copulate with it, the pollen of the flower adheres to the male and is, in turn, carried to the next flower that attracts the wasp.

Fertilization behaviour in cichlid fish

Males of certain African cichlid fish (genus *Haplochromis*) deceive females by means of a body pattern that resembles cichlid eggs. The females, which are mouth breeders, take into their mouths the eggs they have laid before the male fertilizes them. Visual models of these eggs are part of the pattern of the ventral fins of the male. After the female has placed the eggs in her mouth, the male spreads his ventral fin before her. The male emits sperm as the female snaps at the false eggs, thus permitting the real eggs in her mouth to be fertilized.

Another mouth-breeder fish, *Tilapia macrochir,* ensures fertilization by means of a different deception stimulus. The male produces sperm in filament-like packets (spermatophores), which are shed into the water. Later, they are picked up by the female. She may not always find them, however, and the male has evolved long filament-like spermatophore models that project from his genital region. For some reason, these models exert a stronger stimulus than the real spermatophores. The female takes them into her mouth and at the same time receives real spermatophores that have been placed among the models.

Visual behaviour in the frog

The eyes of the meadow frog, *Rana pipiens,* have five types of cells, each of which responds to a different kind of stimulus. One type responds briefly when a light is turned on or off; it also responds to the passing of the leading and trailing edges of an image moving across the retina. It does not respond to a stationary image, however. A second type of cell responds to the passing of straight or curved edges. A third type does not respond to changes in light intensity but to the passage of the image of a small object, in contrast to its backround, across the retina. The fourth type of cell measures a decrease in illumination, and the fifth type measures light intensity.

The frog is capable, therefore, of receiving information about the size, shape, movement, and illumination of objects and is particularly well equipped to perceive small

moving objects—its normal food. Much of the stimulus filtering in the frog's vision takes place peripherally at the retinal level. Studies of other subjects such as cats, rabbits, and moths reveal that the processing and integrating of sensory data occurs at various levels in the central nervous system (brain and spinal cord).

Releasing mechanisms. Hundreds of types of responses to a few key stimuli have been identified in various animals. These responses are mediated in the central nervous system by a so-called releasing mechanism (RM) and are responsible for triggering the specific motor response appropriate to the stimulus. If the releasing mechanism is innate, it is termed an IRM. If it is acquired through individual experience, it is termed an ARM. Some innate releasing mechanisms may be modified as a result of individual experience; such a mechanism is termed an IRME.

A given stimulus does not always prompt the same response in the same individual. Such differences are due to internal factors. Some changes are seasonal and are brought about by internal conditions that may, for example, be related to reproduction and associated aggressive behaviour. In the spring, a male wood thrush (*Hylocichla mustelina*) responds to a female with courtship behaviour and to another male with aggression. In the winter the same thrush fails to respond in this way. Relatively short-term changes in responsiveness also occur. An animal that has just fed, for example, shows no further interest for a time in food. In such an instance a short-term internal change has taken place.

The strength of a stimulus necessary to evoke a response of standard intensity also varies with time. The longer an animal is deprived of food, for example, the more unappetizing the food can be and still be accepted. The converse is also true. The more recently an animal has eaten, the more appetizing the food must be to be accepted.

The intensity of a noxious stimulus or the degree of difficulty of an obstacle that an animal will attempt to surmount varies with time. The longer an animal has been without food (short of physical debilitation), the more difficult an obstacle can be and still be surmounted by the hungry animal. Again, the converse is true.

DRIVE AND MOTIVATION

Internal changes that initiate behavioral changes are commonly termed drive or motivation. These terms are usually applied to short-term reversible changes in response to a constant stimulus. They are not applied to long-term changes that are the result of learning or to short-term changes that result from muscular fatigue, sensory accommodation, and sensory adaptation.

Searching behaviour

When internal conditions are intense enough to initiate a particular drive, an animal commonly behaves as if it were searching for the correct environmental stimulus necessary to trigger the appropriate response. Such searching, or appetitive, behaviour is often highly variable. The true nature of a particular appetitive behaviour can, as a rule, be ascertained only as the act progresses. The drive that motivates a robin to search a lawn, for example, cannot be determined until the search nears culmination. If the robin seizes an earthworm, it is evident that hunger was the activating drive. If it picks up mud or grass, the appetitive phases of nest-material gathering are apparent. Appetitive behaviour tends to become less and les variable as the appropriate terminating situation becomes more and more likely to occur. A hunting falcon flies a search pattern until a potential prey is sighted. The bird dives upon it, after which the exact flight path is determined by whatever evasive action the prey may take. If the falcon is successful, the prey is struck and killed, carried to a perch, and systematically pulled apart and eaten. When the hunger is satisfied, the drive state no longer exists, and some other activity follows. The closer the appetitive sequence is to termination, the more stereotyped the falcon's behaviour becomes. The consummatory act is the most stereotyped behaviour of all.

Courtship behaviour culminating in successful copulation provides many examples of characteristic appetitive–consummatory chains of behaviour. In such cases copulation itself is the terminal appetitive behaviour and is highly stereotyped. Ejaculation, or discharge of sperm by the male, in certain species is completely stereotyped and is followed by temporary cessation of the sex drive.

Learned behaviour is important in appetitive sequences in many animals. A jay, for instance, quickly discovers the best places to find particular foods, and it learns to begin its search in such places rather than search at random until a suitable forage area is found.

"SUPERNORMAL" STIMULI

A major subject of investigation in animal behaviour has been the determination of key stimuli necessary to trigger particular behaviours. In order to determine those characteristics of an egg by which an incubating bird identifies it as such, a selection of model eggs can be presented to the bird, each model differing in one respect from a normal egg. The reactions to variations in colour, pattern, shape, size, and texture vary according to the species. Generally, differences in shape do not seem important to an incubating bird; but models with more rounded contours appear to be favoured. Differences in colour, pattern, and size are important, but differences in texture do not seem to be.

A model in which the key stimulus has been exaggerated to an extreme degree may be chosen in preference to a normal model. The oystercatcher, for example, prefers a "supernormal" egg, several times the usual size. It also prefers an abnormal clutch of five eggs to the normal clutch of three.

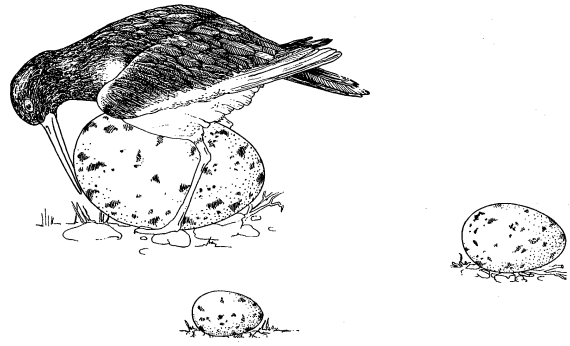


Figure 2: Oystercatcher (*Haematopus ostralegus*) reacting to giant egg in preference to normal egg (foreground) and herring gull's egg (right).

Feeding behaviour of herring gull chicks

Chicks of the herring gull (*Larus argentatus*) are stimulated by a red spot on the lower bill of the adult. When the chick pecks at this spot, the adult regurgitates food for it. By presenting the chick with various models of beaks, it has been found that differences in the colour of the head and bill are not significant; but the red spot, narrowness of the bill, movement, low position of the head, and a downward pointing of the bill are all important in eliciting a response. A thin rod with a red band near the tip moved in a low position provides a supernormal set of stimuli, which elicit a positive response.

When more than one stimulus elicits a given response, the stimuli may supplement one another. If two or more stimuli are required to evoke a response, a weakness of one stimulus may be counterbalanced by the strength of another. Such a compensatory effect is termed the law of heterogenous summation. In higher animals, learned behaviour may play an important role in this phenomenon. A response may originally depend upon one or a few key stimuli, but, as a result of experience, an animal may come to regard previously irrelevant conditions as among those stimuli necessary for the response, resulting in a kind of gestalt response, in which several stimuli are perceived as an integrated whole. Animals lower on the phylogenetic scale may be more apt to respond to heterogeneous summation, and those higher on the scale may be more apt to respond to a gestalt, however acquired.

The supernormal stimuli discussed above have been observed experimentally. The tendency of an animal to respond more vigorously to enhanced stimulation may be of evolutionary significance, because animals with a geneti-

cally based tendency to prefer more advantageous variants of a stimulus would tend to have a higher probability of survival. In social situations in which the relevant stimuli are part of an animal's body, those structures offering a favourable departure from the normal would enhance the probability of survival in the offspring.

MOVEMENT

Form. The form of movement of a response is not determined by either the eliciting stimulus or by the properties of the musculature involved. It is possible, however, that the form of movement may be determined by either of two other factors. First, the sequence in which muscles contract to produce a movement may be determined solely by the properties of central nervous system mechanisms responsible for the movement. In this case the movement, once initiated, is independent of further sensory stimulation.

Second, the muscle contractions near the completion of a movement may also be influenced through a feedback mechanism provided by the earlier contractions. The form of the movement would thus be continuously monitored by sensory control. Fixed action patterns, although independent of further stimulation once elicited, may depend upon such internal feedback mechanisms.

Distinction between external and internal movement. The way by which animals are able to distinguish between movements of the environment and movements of the sense organs is not fully understood. When the human eye views a moving object, the object appears to move. If the eye is moved while looking at a stationary object, the object appears stationary, though in both cases the image moves across the retina. But if a stationary object is viewed while the eye is displaced slightly by pushing with a fingertip, the object appears to move.

Equilibration in fishes

If a resting fish is tilted to one side, the statolith (organ of equilibrium) on that side shifts position, thereby activating sensory endings; these set in motion muscular action that restores the fish to an upright position. A fish often deliberately tilts sideways, however, and, in this case, the automatic reflex does not pull the fish upright. It was formerly believed that the righting reflex is blocked during spontaneous movement. Studies have shown, however, that such blocking does not occur. If a fish is whirled in a centrifuge, the deliberate tilting movements made by the free-swimming fish are of lower intensity. The tilting movements become less because the statoliths are made heavier. The righting reflex is not blocked during deliberate tilts, therefore, but is dependent upon the feedback caused by the tilts.

During spontaneous movement, the stimuli that otherwise release postural reflexes are not inactivated but must be neutralized in another way. The principle of reafference has been hypothesized to account for this. By this hypothesis the functional system is visualized as a feedback loop, whereby afferent nerves carry impulses toward the central nervous system and efferent ones carry impulses away from the central nervous system to the motor areas. Afferences can be divided into receptor excitations caused by internal changes in the musculature (reafference) and those produced passively by external stimulation (exafference). Reafference and exafference are integrated in some manner in the higher centres of the nervous system. The reafference hypothesis postulates that, with each voluntary movement, a copy of the efferent motor impulse is stored in a subordinate nervous centre. The efferent impulse continues to the effector, and movement results. The sense organs then report the result of this movement as a reafference—a feedback of information. This reafference is matched with the efferent copy and is cancelled. If the total afference is too much or too little, as the result of external stimulation, there remains a plus or minus value as compared to the efferent copy stored in the subordinate centre. The discrepancy is reported to the higher centre, which then strengthens or weakens the initial command.

BEHAVIORAL CHAINS

When an animal responds to a stimulus, the releasing situation is often altered because the animal has progressed to a new position, in which other stimuli are effective. For example, when a female three-spined stickleback enters the territory of a male, he performs a zigzag dance. She responds to this with a signal of her own, which, in turn, releases a behaviour in the male that causes her to follow him. The male shows her the nest opening, which she enters. The male trembles with his snout against her tail, stimulating her to spawn, after which she leaves the nest. The male then fertilizes the eggs. Each of these behaviours depends upon the appropriate stimulus. If one is omitted, the chain terminates without a productive conclusion.

The behaviour of the bee-hunting wasp *Philanthus triangulum* illustrates another such chain. This wasp flies from flower to flower as it searches for bees. It responds initially to the visual stimulus afforded by any moving bee-size object; during this time it is indifferent to bee scent. After the wasp perceives the visual stimulus, it hovers about 10 to 15 centimetres (four to six inches) downwind of the bee and then is sensitive to bee scent; if the scent is appropriate, the wasp attacks the bee and seizes it. Following seizure, bee scent is no longer an effective stimulus. Moving models of the appropriate size attract the wasp, but they will not be seized unless they have bee scent. The behaviour depends upon a succession of stimuli that must occur in a precise sequence.

Scent and vision in wasps

Many such behavioral chains are known for vertebrates as well as invertebrates. They are not always precisely ordered, and variations may occur. Many such behavioral chains do not exhibit a succession of stimuli made available as a result of the responses. Single causal factors may stimulate several responses. Activities occurring near the end of a chain may require a higher intensity of stimulus than earlier ones. If a causal factor proves inadequate at

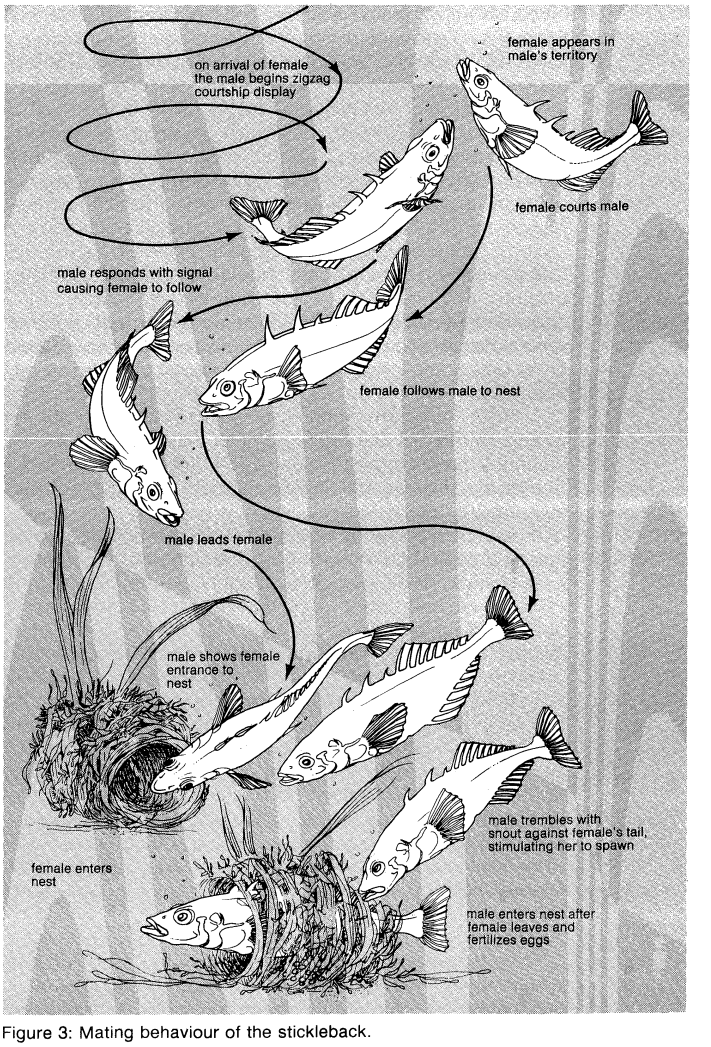


Figure 3: Mating behaviour of the stickleback.

some point, the behaviour reverts to an earlier stage in the sequence. This is probably true in the courtship of the three-spined stickleback in which all the activities of both sexes depend upon common endocrine factors (*i.e.*, hormones), on short-term states of heightened responsiveness, and on the nature of the external stimuli.

CONFLICT RESOLUTION

Simultaneous stimulation. Causal factors for many types of behaviour are usually present at any given moment. A male stickleback may be simultaneously confronted with several stimuli: a ripe female, food, a rival male, and a predator. Animals usually respond to one stimulus at a time and according to a certain priority of sequence. In the male stickleback, escape from a predator almost always takes priority over concurrent stimuli.

More than one drive is often activated simultaneously by the same situation. A conflict between the drives occurs, and the situation must be resolved. The resolution may occur in any one of several ways. Sometimes two drives may be expressed simultaneously. Pecking and head turning, when activated together, often occur simultaneously in chickens. Chickens that are in conflict between watching out with an elongated neck and making wide sweeping movements of the head as they peck, elongate the neck even further but reduce the extent of the head sweeping. Such conflicts can also be resolved by alternately performing the activities appropriate to the conflicting drives. The separate activities are often incomplete. A housewife, for example, may perform in this manner if the telephone begins ringing just as something begins to boil over on the stove.

Redirection and displacement. Sometimes an animal has a drive to perform a particular behaviour but is prevented from doing so and directs the behaviour to another object. If an animal is prompted to attack another but is prevented by fear of the opponent or by a reluctance to leave its territory, it might attack a harmless companion, the ground, vegetation, or even itself. Such behaviour is termed redirection. Displacement is the resolution of a conflict situation in which a seemingly irrelevant activity is performed. When an animal is obviously in conflict between, for example, sex and aggression or between aggression and fear, it will often perform an apparently irrelevant activity such as grooming, feeding, or scratching, or the animal may go to sleep. It is as if the two activated drives neutralize one another and the surplus energy is fed into another system. In disinhibition, two drives that appear to independently inhibit a third mutually inhibit each other in a conflict situation and lose their inhibiting effect on the third, which then becomes free to activate its own behaviour.

Disinhibition hypothesis

The probable reason that displacement activities are so often comfort behaviours (*e.g.*, preening) is that such behaviours have a typically low threshold of performance; ordinarily, they do not have a particularly high priority and are, therefore, easily elicited when more urgent demands are not being made upon the animal. This explanation is consistent with the fact that, though the frequent performance of these displacement activities is essential to survival, they are seldom associated with any condition of urgency.

The nature of the displacement activity may also depend upon the immediate environment or upon the effects of autonomic nervous activation on the animal; sometimes the activity depends on both factors. A wood thrush caught, while sitting in a horizontal position, by a strong attack–escape conflict will wipe its bill on its perch. The same bird caught by the same conflict, while sitting almost vertically, will make perfunctory preening movements directed at its upper breast. The relationship of the bird to its immediate environment makes it easier to perform a particular comfort activity. Autonomic responses—the result of fear or aggression in man, for example—may cause an expansion or constriction of blood vessels, a tingling sensation in the skin, a tendency to urinate or defecate, and so on. Scratching and temperature-adjustment behaviours may sometimes be prompted by such physiological changes.

Transitional activity. In transitional activity, another type of conflict resolution, the animal is stimulated to perform a particular behaviour, but the required environmental stimulus becomes unavailable during the course of the response. The animal discontinues its initial behaviour and substitutes another behaviour that it initially had not "intended" to perform.

The common grackle demonstrates a transitional activity in the form of a threat behaviour termed a spread–squeak; the bird ruffles its plumage, then utters a squeak as it compresses its feathers. If a rival approaches, prompting a spread–squeak, and turns away only at the plumage-ruffling stage, the bird will then, instead of squeaking, often shake its body—a normal comfort activity customarily preceded by plumage ruffling. When a man offers his hand to be shaken and it is not accepted, he often behaves as if he had really intended to gesture or perform a similar action. Transitional activities may be cases of displacement in which the nature of the behaviour is determined by the immediate environment. A grackle involved in the ruffling stage of a spread–squeak, after the rival leaves, may simply shake its body because, in the absence of further stimuli for fear and aggression, it is left with the stimulus for shaking, which may be ruffling. Conflict situations are of great interest from an evolutionary standpoint, because they are often the raw material from which signals have evolved.

Behavioral evolution and development

Behaviours are believed to evolve in the same way as structures. The recombination of genes afforded by sexual reproduction ensures that each individual differs in some degree from all others of its species. Even slight variations from the norm increase or decrease the probability of survival. Advantageous features tend to be conserved and disadvantageous ones eliminated. Marked changes are the result of the slow accumulation of small variations over long periods of time, representing many generations. A species never becomes totally adapted to its environment because the environment is constantly changing. As a result, selective pressures always exist, and the process of evolution continues.

SELECTION IN DOMESTIC ANIMALS

Animals can be selectively bred for specific behavioral changes. Many domestic animals differ markedly in behaviour from their wild progenitors. Domestic breeds such as fighting cocks and Siamese fighting fish are hyperaggressive, but most domestic animals tolerate greater crowding and are more docile than their wild ancestors. House mice have been selectively bred in the laboratory to produce unusually aggressive, as well as unusually timid, strains. Mating selection in domestic animals is usually less restrictive than in their wild counterparts. Reciprocal signalling systems between animals become less precise with domestication, and behavioral components may be omitted or lost altogether. Promiscuity may replace pairing in certain animals.

Courtship in pigeons

Marked differences in courtship displays occur between wild pigeons (*Columba livia*) and domestic breeds. Wild males loudly clap the wings over the back in flight and then glide with the wings held well above the horizontal position. In pouter pigeons (a breed of *C. livia*), the wings are clapped so frequently that two-thirds of the length of the primary wing feathers may thus be worn off, with the result that flying becomes very difficult. The elevation of the wings during the glide phase becomes so exaggerated that the wing tips touch, and the bird quickly loses altitude. In roller pigeons, another breed, the gliding flight has become a series of backward somersaults. On the ground the male wild pigeon, while cooing, twirls and makes a small hop when the female walks away. The German ringbeater breed elaborates on this behaviour by performing a wing-clapping flight around the female, who remains on the ground.

Domesticated zebra finches (*Poephila guttata*) show marked loss of specificity in their mating interactions and in care of the young, when compared with their wild counterparts. Wild chickens will kill their own chicks that

lack specific colour patterns. Domestic chickens, on the other hand, will care for almost any chick regardless of colour and pattern; yet, they retain specific reactions to the species-specific calls of chicks so that they do not ordinarily accept other young birds, such as ducklings. Highly domesticated chicken breeds, such as Plymouth Rocks and barred Plymouth Rocks, however, will rear even ducklings. The least domesticated breeds, such as certain game bantams, still show much of the specificity characteristic of wild birds. Wild graylag geese form pairs only after a very long courtship period and remain monogamous. Domestic derivatives, on the other hand, pair quickly with any member of the opposite sex and are not monogamous. All of these differences between wild animals and their domesticated derivatives have a genetic basis.

BEHAVIOUR IN HYBRIDS

Two closely related species of small African parrots, the peach-faced lovebird (*Agapornis roseicollis*) and Fischer's lovebird (*A. personata fischeri*), have completely different methods of carrying nesting material. The females of both species prepare nesting material by cutting long, narrow strips of bark, leaves, or paper. The peach-faced lovebird tucks each strip, after she cuts it, into the feathers of the lower back, or rump. When she has accumulated about six strips, she flies to the nest cavity, retrieves the strips, and places them in her nest. Fischer's lovebirds carry each strip in the bill, one at a time, to the nest cavity.

Female hybrids between these two species initially tuck nest material into their rump feathers, but the strips fall out before the birds reach the nest. The birds gradually develop, through learning, an increased tendency to carry each strip singly in the bill. About four months after the onset of the tucking behaviour, they are utilizing both behaviours about equally. Although the tendency to carry in the bill continues to increase after this point and the tendency to tuck continues to decrease, the rate of divergence between the two methods becomes much slower. By the end of the third year the hybrids carry all strips in the mouth, but they make small intention movements to tuck. These intention movements consist of little tic-like, side movements of the head just before the bird flies off to the nest.

The courtship behaviour of male hybrids, paired with female hybrids of this same cross, is intermediate between that of the two parental-species males. When the hybrid males are paired with parental-species females, their courtship behaviour, in most cases, is closer to that of the parental species of the female, although it sometimes remains intermediate. The species-typical behaviour of the females is thus seen to influence the pattern of male courtship. The courtship behaviours of some bird hybrids are not so greatly modified; for them, no permissible variability has been inherited.

Behaviour inheritance of crickets

Two cricket species, *Gryllus campestris* and *G. bimacularis,* are so similar morphologically that they can be distinguished from one another only with great difficulty. Their behaviours on the other hand, differ markedly. If the two species are crossed, however, the inheritance patterns may be traced by means of behaviour. Four behaviour patterns—antennal vibration in the post-courtship period, pendulum movements of the thorax, stridulation (rubbing one body part against another to produce sound), and fighting by young adults—have been investigated in particular detail. It has been found that antennal vibration and juvenile fighting in the hybrids have a monofactorial inheritance (*i.e.,* are caused by a single gene). The pendulum-like movement of the thorax during mating is found only in *G. campestris* and has a polygenic basis (*i.e.,* is caused by more than one gene). The stridulating sounds preceding courtship, performed only by *G. bimacularis,* are seemingly based on one pair of alleles (*i.e.,* different forms of a single gene).

Two races of honeybees are distinguished from one another by the presence or absence of hygienic behaviour. The race exhibiting hygienic behaviour opens comb cells containing dead pupae, which are removed. The nonhygienic race leaves dead pupae in their cells. The first generation (F_1) of hybrids contain only nonhygienic bees. One F_1 queen produces four kinds of drones, or males. When the F_1 is backcrossed with the hygienic form, a second generation (F_2) is obtained, which is made up of four different types of bees. One group is hygienic; one group opens the cells of dead pupae but does not remove the dead pupae; one group does not open the cells of dead pupae but removes the dead pupae if the cells are open; and the remaining group is nonhygienic.

THE INFLUENCE OF EXPERIENCE ON BEHAVIOUR

The adaptive change of behaviour as the result of experi-

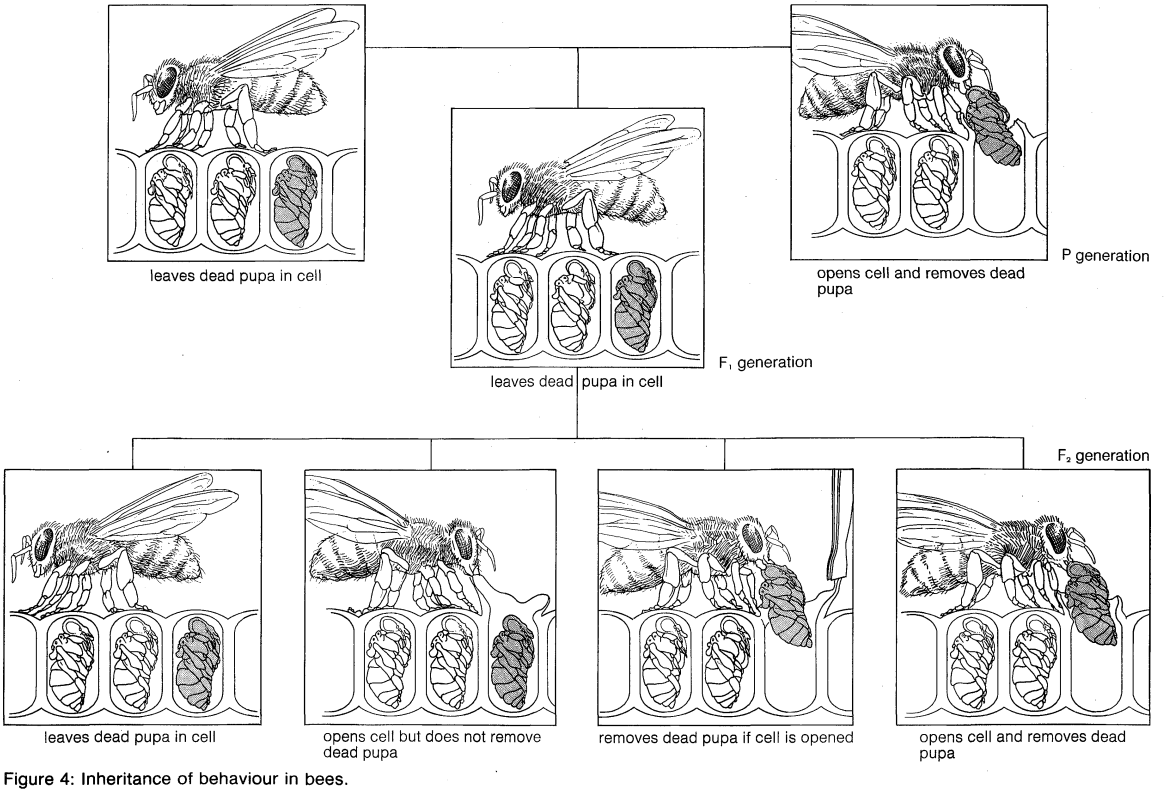


Figure 4: Inheritance of behaviour in bees.

ence—usually known as learned behaviour or experience-dependent behaviour—may be observed in all higher vertebrate forms. Several types of such learning in animals are recognized: habituation, classical conditioning (CR Type I), trial and error (CR Type II, instrumental conditioning, or operant conditioning), latent learning, insight learning, and imprinting.

Habituation. Habituation is learning to disregard stimuli that are without significance to the animal. In many respects it is the simplest form of learning, and it is sometimes regarded as a fundamental property of all living matter. Most animals inherit a response to be frightened by sudden and strong stimuli such as loud sounds, flashes of light, and the sudden intrusion of anything foreign into the animal's sensory field. Yet, if an animal reacts nonselectively to all phenomena such as rustling leaves, thunder, snapping twigs, and the sudden appearance of harmless animals, those phenomena that are significant to the animal's well-being will not receive the intensity of response that they often require. All animals, then, quickly habituate to such harmless stimuli, but the adaptation is highly specific. An animal that habituates to one type of sound does not, as a consequence of this habituation, become habituated to other sounds. Habituation is distinct from failing to respond to stimulation as a result of fatigue, sensory adaptation, or injury. The effects of habituation are generally long lasting. If an animal is repeatedly exposed to a potentially harmful stimulus (such as to a predator) without being harmed, habituation does not generally occur. Responses to dangerous stimuli often seem to have an inherited resistance to habituation—a mechanism of obvious survival value.

alivation xperi- nents of avlov

Classical conditioning. Classical conditioning was studied early in the 20th century by the Russian physiologist Ivan Pavlov, who observed that dogs salivate when food is placed in their mouths. He gave dogs food and, at the same time, provided another stimulus such as a flashing light or the sound of a bell. After a few such pairings of stimuli, a dog would salivate upon seeing the light or hearing the bell but without the presence of food. The dog had learned to associate the flashing light or the sound of a bell with food. A previously irrelevant stimulus that assumes significance as a result of association with a relevant stimulus is called a conditioned stimulus (CS). Salivation in response to such a stimulus is termed a conditioned response (CR). Prior to the learning experience, only the food (unconditioned stimulus) was effective in producing salivation (unconditioned response).

Such conditioned responses have been observed in a wide variety of animals, from lower invertebrates to man. Birds learn to avoid noxious insects in this manner; the distasteful monarch butterfly (*Danaus plexippus*) or a species of stinging wasp provide effective stimuli that quickly become associated with the appearance of such insects. This kind of association together with the conditioned response relevant to it is also the basis for Müllerian mimicry, in which palatable insects and other animals evolve to resemble noxious ones, thus enhancing their chances of survival.

Trial and error. In trial-and-error learning, an animal learns to behave in a particular way by associating something it does with a desired effect. If a dog's foot is lifted by the experimenter and food then given, the dog, after a few such trials, will spontaneously lift its foot in anticipation of food. Both classical conditioning (CR Type I) and trial-and-error learning (CR Type II) are termed associative learning, because in both cases an unconditioned response is associated with a conditioned stimulus. In natural situations an animal probably learns to associate certain spontaneous activity of its own with certain desired results, thus fixing the conditioned stimulus and response. Learning of this type often occurs when animals modify their behaviour during appetitive sequences such as those involving feeding and mating.

Latent learning. Latent learning is the association of indifferent stimuli or situations with one another without reward. The phenomenon is clearly exemplified in exploratory behaviour. Animals finding themselves in unfamiliar environments or among unfamiliar objects, but in familiar surroundings, show exploratory behaviour. The animal uses its sense organs on all that is novel, shows much ambivalence between approach and avoidance, and, finally, as the hesitancy to approach wanes, will test the novelty. A mammal may, at this point, sniff, nudge, or handle a strange object; a bird may peck at it. The first contact often results in abrupt avoidance, but, typically, the object or situation is, at last, thoroughly explored and finally abandoned if neutral. A mouse sniffs and pokes about most of a new environment. A bird, having sight and hearing, rather than smell and touch, as dominant sensory modalities does not have to occupy physically as much of a novel environment. Instead, it places itself successively at several vantage points and then carefully peers about and listens.

Subsequent behaviours of an animal can reveal that it has learned much about its environment during such an exploratory phase. It may learn, for example, the physical features of the environment and their spatial relationships to one another, the location of food and water, and the location of places safe from predators. Animals apparently learn all of these things early in their exposure to an environment, even though the information acquired may become significant only at some time in the future. The learning takes place without being associated with immediate reward (unlike that in conditioned response Types I and II) unless a need to know is postulated as a kind of immediate self-reward.

Insight learning. Insight learning is believed to be an advanced type of learning. Insight involves the spontaneous combination of a number of isolated experiences; the result is a new experience that is effective for gaining a desired result. Humans are able to exercise insight; it is extremely difficult, however, to identify such behaviour in most other animals. Animals are believed to use insight when they solve a problem too rapidly for normal trial and error to occur. It is possible that such an animal is carrying out trials in its brain; this implies reasoning ability. The higher primates are probably capable of insight learning at times, but further down the phylogenetic scale the evidence of such learning becomes progressively less conclusive.

Chimpanzees, to get food out of reach, will pile boxes to make a stand for themselves or will fit sticks together to knock the food down. These solutions may come quickly, an obvious result of prior experience (latent learning). Much trial-and-error learning is demonstrated, however, when they actually pile boxes or fit sticks together. In humans, insight is probably often aided by latent learning and trial and error.

Insight in chimpanzees

Imprinting. Imprinting, a learning process observed in young birds and mammals, is the identification of an animal with another animal. Normally, it is a relationship between members of the same species, but it can occur, for example, between a bird and a human. Imprinting can take place only during a particular period of the animal's development—a time span that is specific for each species.

In 1935 the Austrian ethologist Konrad Lorenz first observed the process in ducklings and goslings. After goslings hatch and become dry, they follow their parents. The adults provide warmth, safety, and shelter and bring the goslings food. The more uncomfortable a gosling becomes (*e.g.*, cold, frightened, hungry), the more intensely it follows. If goslings are reared by a human, they become imprinted to humans; thus, they ignore geese.

The development of this response occurs during a sensitive period, before and after which the response cannot be learned; if the response is not acquired during the sensitive period, it will never occur. Zebra finches that are isolated from their own species before they are 35 days old are never able to distinguish males and females of their own species. This is because their sensitive period for imprinting occurs before they are 35 days old.

The duration and time of onset of the sensitive period depend on the species and on the type of behaviour involved. Some animals imprinted to animals of another species will mate with members of their own species but, if given a choice, will prefer the animal to which they have been imprinted. Many species refuse social contact with any animal except the one to which they are imprinted.

Male golden pheasants (*Chrysolophus pictus*) imprinted to humans will court females of their own species but immediately transfer this behaviour to a human, should one appear. The same is true for budgerigars (*Melopsittacus undulatus*) and turkeys (*Meleagris gallopavo*). Mallard ducks imprinted to humans, on the other hand, will not associate with members of their own species (conspecifics) and will continue throughout their lives to treat humans as conspecifics. Imprinting is fixed for life, in contrast to other types of learning, in which forgetting is common. Imprinting of motor patterns, such as birdsong, also occurs. Exposure to a particular birdsong may be relatively brief and still be permanently fixed in the bird's memory. Chaffinches (*Fringilla coelebs*) learn their songs during the first 13 months of life, although they do not sing until nearly a year later. A 12-day-old nightingale (*Erithacus megarhynchos*) was kept in the same room with a singing black-capped warbler (*Sylvia atricapilla*) for about one week. The following spring the nightingale sang a typical black-capped warbler song.

Little is known about imprinting in mammals, but hoofed animals, such as sheep and horses, that are imprinted to man express this response by following him about. Dogs from four to six weeks old develop normal social responses to dogs or to another species such as man.

Imprinting in man

Imprinting apparently also occurs in humans. An infant deprived of its mother for a short period during its first year may develop serious mental retardation. A separation of several months—particularly during the seventh to the 12th month—will frequently result in irreparable damage; under such conditions, death may result.

PLAY BEHAVIOUR AND CURIOSITY BEHAVIOUR

Play and curiosity are exhibited by many mammals and by some birds and figure importantly in the learning of numerous activities. Play is especially characteristic of young animals, but the adults of many species also engage in it. Spontaneous curiosity, in which the animal actively seeks out novel situations for exploration, is exhibited by the young of mammals and some birds; indeed, they seem to be under the compulsion of some drive to do so. Carnivores and primates exhibit more curiosity than rodents, which gnaw novel objects and may hoard them. Monkeys inspect and manipulate such objects.

The curiosity drive implements the development of new motor skills and ensures the acquisition of new perceptual impressions, thereby resulting in new knowledge. The only reward, however, seems to be the performance of the activities themselves. Rhesus monkeys will learn a puzzle game without any reward except its successful solution. Among rats, it has been observed that the nerve cells in the lateral hypothalamus and preoptic regions of the brain are more active in those rats that explore. Electrical stimulation of these brain areas is rewarding to rats, and they will learn to press a lever that activates this stimulation.

Such curiosity behaviour seems linked to play behaviour. Play is difficult to define; it is usually easy, however, to distinguish a playing animal from one that is seriously occupied. An animal plays only when it is satiated and not preoccupied with other tasks. Play seems not to be dictated by immediate need but is extremely important in behavioral development. Only animals that spontaneously seek new situations on their own initiative play in the true sense. Invertebrates, fish, and amphibians do not seem to play. The taxonomic distribution of play among mammals and birds suggests that play is related to learning. Play involves interactions with the environment; this leads to the acquisition of knowledge about environmental features, including information about conspecifics and the animal's own possibilities of movement. Play behaviour occurs only at particular times; progression to a second play activity takes place only after a certain level of skill has been achieved in the first.

Mock fighting and fleeing

Much play appears to be fighting or fleeing behaviour, and usually it is easily identified as such. An animal that is play escaping or play attacking does not actually escape or attack. A rodent play fleeing into a hole, for example, quickly reappears. If a rodent's flight is truly an effort to escape, it reappears only after a much longer interval. Play-fleeing animals often reverse roles quickly, and the pursuer becomes the pursued. Threat behaviour that is associated with real attack is missing, and there is strong reluctance to bite. Play tends to be highly repetitive. A dog may retrieve a stick many times or play fight until it is exhausted or until a more interesting activity distracts it.

Bernhard Grzimek

Figure 5: Young chimpanzees exhibiting mock fighting behaviour.

Such play behaviour could mistakenly be postulated as the performance of immature instinctive activities. In many instances, however, this is known not to be the case. Much playful behaviour occurs at a time in an animal's life when it is fully capable of serious activity. Play also involves the use of species-typical patterns of behaviour in various sequences that do not occur in serious activity.

MODIFICATION OF INSTINCTIVE BEHAVIOUR BY EXPERIENCE

Behaviours based on both instinct and learning are commonly intercalated into functional wholes. In the peach-faced lovebird, for example, the cutting of nest-material strips is partly instinctive and partly the result of experience. The propensity for cutting is instinctive. This includes punching holes in the sheet of material from which the strips are fashioned and a "knowledge" of the proper width, length, and straightness of strips. The spacing of punch holes in such a manner as to form a strip is learned through experience. It is as if the animal has an instinctive picture in its central nervous system and persistently tries punching holes, in various relationships to one another, until the right pattern is made. The bird tends to repeat punching patterns that most closely approach the ideal and, thus, gradually, through progressively more satisfying feedbacks, approach the definitive functional technique of cutting strips. Idiosyncratic techniques develop—some birds stand on the sheet while cutting, and others stand off the sheet; some cut to the left, others to the right, and still others cut in various combinations of these directions. Birds developing their techniques try all of the directions and places of standing but gradually act with less and less variation. They do not need to observe experienced birds in order to develop a technique. The stimuli necessary for learning response are intrinsic to the birds' individual activities.

Song learning

If a meadowlark (*Sturnella magna*) is exposed to an alien song during its sensitive period for learning song, it will learn the alien song. Meadowlarks, then, learn their species-typical songs rather than inherit the capacity for particular melodies. But, if they are exposed during the sensitive period to alien songs along with meadowlark songs, they will learn only the normal species-typical song. Although the song must be learned, the bird instinctively learns the species-typical song if there is a choice. The bullfinch (*Pyrrhula pyrrhula*) instinctively learns the male parent's song rather than the species-typical song. Bullfinches raised by foster parents of another species will learn the song of the male foster parent, even though the

normal bullfinch song is also audible to it during the same period. In both types of song acquisition, learned and instinctive elements combine in the development of the species-typical song. Some species, however, do not learn their songs and do not need the experience of hearing others sing during their development. Different vocalizations in the same species are commonly acquired in more than one way. Some are purely instinctive; others are learned.

The behaviour of an individual animal is the result of a genotype that has developed over millions of years of evolution—a genotype that also permits a certain degree of variability through experience.

Hormonal and nervous control of behaviour

INTERACTION OF ENDOCRINE AND NERVOUS SYSTEMS

Physiological changes within an animal are largely the direct or indirect result of nervous and endocrine (hormonal) changes and their interactions. These changing states are also responsible for a changing responsiveness to internal and external environmental stimuli.

Neurosecretion

The endocrine system and the nervous system are probably of equal evolutionary age; the two systems may be evolutionarily linked. Certain nerve cells, or neurons, that are highly modified produce substances that pass through the axons (threadlike extensions of neurons) and into the bloodstream. These neurosecretory cells are sometimes clustered, forming glands that have connections with both the nervous system and the bloodstream. The endocrine glands are similarly distributed. Some may have evolved from clusters of neurosecretory cells. The vertebrate pituitary gland has evolved as a fusion between neural tissue and epithelial tissue (the lining, or covering, of organs) and is intimately associated with the hypothalamus, a ventral portion of the brain. The pituitary may be regarded as a master gland that regulates, with its secretions, all the other endocrine glands. The nervous system, in turn, has a regulatory effect upon the pituitary, which may also be influenced by feedback effects stemming from the secretions of other endocrine glands. This relationship permits the outside and inside environments to exert influences on the endocrine system. Seasonal changes in day length, for example, influence the nervous system by means of visual stimuli. The endocrine system is then activated through stimulation of the pituitary gland by the hypothalamus; the pituitary, in turn, secretes hormones appropriate to the most adaptive response. The ultimate effect of a seasonal change in day length may be migration or reproduction (see ENDOCRINE SYSTEMS).

The nervous system and the endocrine system interact with and complement each other. The nervous system sends information with great speed but of short duration along its pathways. Its messages can change rapidly. Hormones, secreted by the endocrine system into the bloodstream, travel much more slowly than nervous inpulses. The endocrine system can keep a message constantly available for many months if necessary. The nervous system generally affects only muscles and glands, but hormones can reach every cell in the body. Adrenaline is a hormone that acts with relative speed. It is secreted by the two adrenal glands, which are attached to the kidneys. The adrenals consist of two portions that differ from one another both in origin and in function. The inner portion is the adrenal medulla, and the outer portion is the adrenal cortex. During stress, such as occurs in fighting, mating, and fear, the adrenal medulla is stimulated by the autonomic nervous system to release adrenaline into the bloodstream. Some of the changes that occur throughout the body under the stimulus of adrenaline include hair erection, sweating, and acceleration of heartbeat and breathing; adrenaline also causes the blood to be diverted from the digestive tract to the muscles. All of these changes, and others, help prepare the animal for extreme effort. During brief stressful periods the bloodstream is quickly flushed with adrenaline, but the hormone is quickly dissipated. If the stress situation persists, other events take place—the adrenal cortex becomes involved, and its hormones are released. The cortex, unlike the medulla, is not under direct nervous control but is stimulated by another hormone, the adrenocorticotropic hormone, or ACTH, produced by the pituitary gland. Prolonged stress stimulates cells in the hypothalamus, which, in turn, stimulates the pituitary gland to produce ACTH. The cortical hormones in turn stimulate various responses to prolonged stress. Some of these responses are concerned with the metabolism of glucose, a sugar that may be associated with the utilization of food reserves. The effects of cortical hormones are, in any case, profound, and continuing stress will result in enlargement of the adrenal cortex, which leads to increased production of the cortical hormones. Chronic stress may cause severe illness and even death. It has been shown that rats confined to the territories of other rats will die as the result of overproduction of cortical hormones in response to the stressful situation. Stress as a result of overcrowding may also cause death in animals. Such stress may, in fact, be the cause of a decline in numbers of mice after they have reached a certain population level in a given area.

ACTH

SEX HORMONES

The pituitary hormones that have a direct effect upon reproductive behaviour are the follicle-stimulating hormone (FSH), the luteinizing hormone (LH), and prolactin (lactogenic hormone or luteotropic hormone). FSH and LH are called gonadotropins because they stimulate the gonads (ovaries and testes) to produce germ cells and gonadal tissue; gonadal tissue, in turn, secretes other hormones. Prolactin has a variety of effects. In different species of vertebrates, prolactin affects different target organs. In female mammals, for example, it stimulates growth of the mammary glands and the secretion of milk. It also stimulates the corpora lutea—glandular bodies of the ovary—causing them to produce another hormone, progesterone. In pigeons and doves, prolactin causes the characteristic modification of the crop (stomach) associated with the production of so-called pigeon's milk—a soft white substance that is passed from the mouth of the adult pigeon to that of the young. A slow change of colour in some fish is caused by the influence of prolactin on pigment cells. (Rapid colour changes are under nervous control.)

The gonads secrete hormones from special cells when stimulated by FSH and LH from the pituitary gland. Collectively, the female hormones are termed estrogens, and the male hormones are called androgens. Both androgens and estrogens belong to a chemical group known as steroids. All steroids have closely related chemical structures; the different vertebrate groups have slightly different steroid hormones that seem to be largely interchangeable in function. Removal of the testes or ovaries (castration) in vertebrates causes profound changes in behaviour and structure, especially if done early in life. Among many invertebrates castration has no such profound consequences. Apparently, therefore, only in the vertebrates are the gonads important endocrine organs.

Effect of castration

Androgens and estrogens control the development of the secondary sexual characteristics; they also effect the production of eggs and sperm from the gonads. Secondary sexual characteristics tend to be relatively permanent throughout life, but many are temporary, occurring only during the breeding season. Examples of temporary characteristics include the special breeding plumages and songs of many birds, the antlers of deer, colour changes in some fish, and all the behaviour associated with the formation of fertilized eggs (zygotes).

Progesterone, the production of which is stimulated by prolactin, is produced in mammals by the ovary. After an egg is shed, the empty follicle enlarges, forming the corpus luteum, a conspicuous yellowish structure, which secretes progesterone; progesterone, in turn, stimulates changes in the uterus preparatory to its receiving the fertilized egg. Progesterone also inhibits the contraction of uterine muscles. The females of other vertebrate groups produce structures similar to the corpora lutea of mammals. Birds, which have rather inconspicuous corpora lutea, also produce progesterone.

The amount of any hormone normally present in the bloodstream is minute. Artificially high levels often have marked effects; the introduction of massive doses of testosterone into females, for example, causes male sexual

behaviour. It will also cause female sexual behaviour in both sexes, in which case testosterone may be acting as a general stimulant.

Sometimes a massive dose of hormone has an effect opposite to the one expected. Female canaries (*Serinus canarius*) given overdoses of estrogen would be expected to show enhanced sensitivity of their brood patches (highly vascularized areas of skin in close contact with the eggs during incubation); instead, overdoses of estrogen may result in some desensitization of these areas.

It is tempting to conclude that the sex hormones have a direct effect on all the structures and behaviours concerned with the formation and nourishment of zygotes. Much is not known, however, about the hormonal control of behaviour. There may be feedback effects after the hormones have initiated a particular response; for instance, estrogen in conjunction with progesterone in some birds causes increased thickening and vascularization of the brood-patch area. The target tissues in this case (skin, blood vessels, and feather follicles) all have a nerve supply, and feedback through them to other parts of the nervous system could initiate further behavioral modifications. Physiological changes resulting from hormonal action may render an animal more or less responsive to its environment and thus modify its behaviour.

There is evidence that hormones directly affect behaviour by acting directly on neurons in the hypothalamus. If estrogen is injected into certain areas in the hypothalamus of castrated female cats, they develop strong estrous behaviour, even though the reproductive system remains underdeveloped.

Behavioral stimulation of hormone production

Behaviour may stimulate hormone production. Female cats, rabbits, and some other mammals are "induced ovulators." In other words, copulation stimulates the hypothalamus via the nervous system, and the pituitary gland is then stimulated to produce luteinizing hormones (LH), which in turn affects the ovaries. A few hours after copulation ovulation occurs at about the time the sperm have reached the upper levels of the reproductive tract; the germ cells meet, and fertilization takes place. In deer, sheep, and weasels, stimulation of the pituitary gland (resulting in the production of follicle-stimulating hormones [FSH]) is achieved when the animal senses a change in day length. The females of other mammal species (*e.g.*, the house mouse [*Mus musculus*]) seem to have an internally based clock that periodically triggers the release of FSH regardless of environmental changes. They may still be influenced, however, by the presence of a male.

Complex interactions between hormones and behaviour are known to occur in birds. The sight of a courting male stimulates the release of FSH and LH in a female dove (*Streptopelia risoria*). Physical contact between the birds is not necessary for this response to take place.

NERVOUS SYSTEM AND BEHAVIOUR

The nervous system receives information about the external environment through sensory receptors and about the internal environment through hormones, internal neural responses, and other physiological events. This information, regardless of the source, is processed in the brain or spinal cord, and appropriate responses are initiated by outgoing (efferent) nerve impulses leading to muscles or glands.

There is much evidence to support the view that the central nervous system (CNS) is hierarchically organized. Its organization is thought to consist of a system of centres, each with the function of collecting stimuli and appropriately redispatching them. Reproduction in the peach-faced lovebird depends first upon the activation in the proper sequence of a number of subordinate centres. These involve formation of a pair bond between a male and female; selection of a nest site; conduction of courtship; laying and incubating of eggs; and caring for the young. Nest building occurs throughout courtship, egg laying, incubation, and care of the young.

Subordinate events such as those mentioned above control, in turn, other neurally controlled events. Nest building, for example, consists of various subordinate activities that have already been described. Each of these subordinate activities also has subordinate activities. Tucking a strip of nesting material consists of several activities: simultaneously turning the head back over the rump, lowering the unfolded wing on the same side, and erecting the rump feathers; pushing the strip into the feathers, performing rapid hooking movements, which seem to function as an anchoring behaviour; releasing the strip; and, finally, simultaneously bringing the head, wing, and rump feathers back to the normal position.

It has been proposed that the smallest irreducible neuromuscular coordinations of an activity be thought of as acts. Each species is capable of performing a finite number of acts, and these are combined in various ways to produce all the behaviours of which an animal is capable. Each act is thought to have an act centre in the CNS. The act centres are subordinate to the behavioral centres, which coordinate them; behavioral centres are, in turn, subordinate to their initiating and coordinating centres and so on. The term centre in each case refers to a functional rather than to an anatomical locus. Portions of the CNS responsible for mediating a given response may be quite diffuse anatomically; typically, there may be much redundancy—a condition that frequently enables an animal to regain normal behaviour following damage to the CNS. Brain tissue, however, does not regenerate.

Act centres in central nervous system

Detailed implementation of various subordinate activities may depend upon the details of environmental feedback, as in the taxis components of many fixed action patterns (see above *Fixed action patterns*). Such detailed implementation may also depend upon long-term changes in behaviour that result from experience. Various releasing mechanisms may also be altered as a result of experience, as may the significance of various environmental stimuli. Experience may therefore exert modifying effects upon the input of information, its mediation in the CNS, and the details of its implementation through muscle coordination and glandular activity. These experiential effects may be of long or brief duration.

For many years the classical reflex was thought to explain adequately the mechanism of various behaviours. Such a reflex consists, in its simplest form, of an afferent neuron carrying information to the CNS, excitation then being carried to an effector (muscle or gland) via an efferent neuron. Intermediate neurons often occur between the afferent and efferent neurons. This entire mechanism is termed a reflex arc. In certain reflexes, the excitations activate the same muscles or glands from which the stimuli originate (monosynaptic reflexes). Most reflexes have intermediate neurons, and the stimulation of a single receptor may activate many effectors; similarly, the stimulation of many receptors may activate but a single effector. Chain reflexes occur as a result of one reflex triggering another. Additional reflex arcs may become established as a result of experience (conditioned reflexes), and some reflexes are thought to have facilitatory or inhibitory effects on others.

Reflex arc

It is now known that not all behaviours are the result of afferent impulses. Early in the 20th century it was found that cats' leg muscles with all afferent nerves removed still demonstrate rhythmic movement. Afferent impulses are not necessary for coordinated response; the swimming movements of eels and other fish, the crawling of earthworms, and the flying movements of grasshoppers are some examples. It is now known that spontaneously generated stimulation of the central nervous system initiates and controls much behaviour.

The evolutionarily older portions of the CNS, such as the spinal cord and the medulla and hypothalamus of the brain, seem to be concerned mainly with inborn behaviour such as heartbeat, breathing, and reflexes and with instinctive behaviour. The evolutionarily newer portions of the brain, such as the cerebrum of mammals, seem to be concerned either with new behaviours resulting from experience or with the modification of inborn behaviours. The conspicuous cerebrum of mammals often comprises a major portion of the brain. Although it is generally accepted that mammals, as a class of vertebrates, are the most intelligent of all animals, many birds seem to be capable of greater modification of behaviour through experience than are some mammals. Many other vertebrates, such

as certain reptiles and fish, are capable of learning new behaviours rather easily. It was long believed that, because birds and other nonmammalian vertebrates have little or no discernible cerebral tissue, their behaviours were instinctive. It is now known, however, that such animals are often capable of much behavioral modification as a result of experience and that some other portion of the brain must be involved. Different portions of the brain in different animal groups seem to have been selected for the development of learning ability. (W.C.Di.)

UNLEARNED BEHAVIORAL REACTIONS

Stereotyped response

A stereotyped response may be defined as an unlearned behavioral reaction of an organism to some environmental stimulus. It is an adaptive mechanism and may be expressed in a variety of ways. All living organisms exhibit one or more types of stereotyped response.

GENERAL CONSIDERATIONS

The capacity for unlearned behaviour is genetically determined in much the same sense as are the position, size, shape, and function of organs. Like structural features, stereotyped responses are the result of a continuing process of evolutionary modification and refinement. Those actions that most successfully aid the animal or plant in its basic drives (*e.g.*, reproduction, search for nourishment, escape from predators) are the ones most likely to be retained in succeeding generations. As environmental conditions change, inherently determined responses also become modified in order to ensure continuation of the species.

Problems of identification and analysis

The problems that arise in the study of stereotyped responses are many and varied. Particular responses in animals do not readily lend themselves to identification in highly evolved forms because learned behaviour patterns obscure the underlying unlearned behaviour; in addition, stereotyped responses provide the building blocks of instinctive behaviour, the complexity of which may obscure the integral parts (see below *Instinctive behaviour*). In lower animals, just as in plants, in which learned behaviour is absent or nil, the analysis of behavioral mechanisms is limited by the fact that many of the most fundamental cell processes are not well understood.

Animal behaviour, as a branch of psychology, represents a confluence between the disciplines of ethology and comparative psychology. Most of the pioneer work in stereotyped responses of animals was done by ethologists. During the first half of the 20th century, when much of the groundwork in experimental psychology was laid, ethologists (who were for the most part European) concerned themselves with behaviour in insects, fishes, and birds and were particularly interested in the evolution of instinct. The comparative psychologists during this formative period were mostly Americans. They studied primarily behaviour in common laboratory animals such as guinea pigs, mice, rats, and monkeys, and their interest tended to focus on environmental influences on behaviour as opposed to genetic influences. Since the 1950s, psychologists in general have recognized that both environmental and genetic factors play essential roles in any biological phenomenon. As a consequence of the separate development of ethology and comparative psychology, however, some difficulties have arisen in the use of terminology. The German-American biologist Jacques Loeb applied the term tropism to all oriented movements of organisms, and he proposed that all behaviour is composed of tropisms. Subsequently, to avoid confusion, the terms taxes (singular: taxis) and kineses were introduced by other investigators to refer to animal responses other than those of sedentary, plantlike forms. The terms also have been applied to certain plant movements. Although a variety of discrete stereotyped response movements occur in plants, particularly in higher forms such as flowering plants, these autonomous movements usually occur too slowly to permit detection by casual observation. That movements of plants or plant organs actually take place can be strikingly demonstrated by time-lapse photography, in which single photographs are taken at regular intervals as brief as seconds or as long as several days or more. The photographs are then compared or viewed in rapid sequence as a motion picture.

TYPES OF STEREOTYPED RESPONSES

Stereotyped response in animals may be separated into the following four categories: unorganized or poorly organized response, reflex movements of a particular part of an organism, reflex-like activity of an entire organism, and instinct.

Unorganized or poorly organized responses are given by early embryos or by animals (such as sponges) that lack nervous systems.

Reflex. Reflexes proper, or reflex-arc movements, include responses such as the immediate withdrawal of the hand on touching a hot surface. The basic components of the reflex arc are the receptor, or sensory-nerve cell, which senses the stimulus, and the affector, the nerve cell that directly activates the muscle. These are a theoretical minimum rather than an observed functional arrangement of cells in the body of an animal (see below *Varieties of instinctive behaviour*).

Reflex-like activities. Reflex-like activities of entire organisms may be unoriented or oriented. Unoriented responses include kineses—undirected speeding or slowing of the rate of locomotion or frequency of change from rest to movement (orthokinesis) or of frequency or amount of turning of the whole animal (klinokinesis), the speed of frequency depending on the intensity of stimulation. Examples of orthokinesis are seen in lampreys, which are

Orthokinesis and klinokinesis

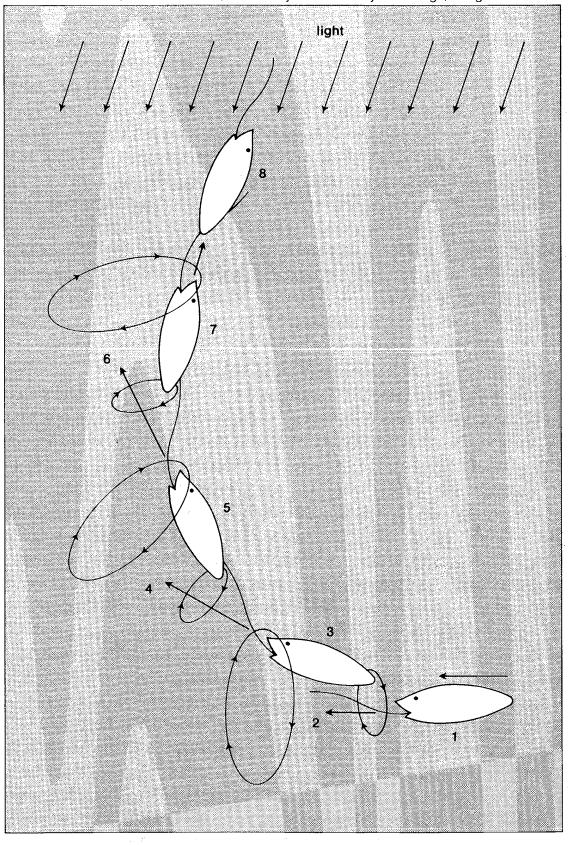


Figure 6: Klinotaxis in *Euglena*. As the animal swims in a spiral, the single eyespot at its front end swings in a circle. *Euglena* gradually orients itself so that the eyespot continues to receive even illumination.

more active in high intensities of light, and in cockroaches, which are more active in low intensities; flatworms and many fly larvae, among other invertebrates, show orthokinesis. Klinokinesis is well demonstrated by the movements of the wood louse (*Porcellio scaber*). When wood lice are placed in dry air, they crawl about actively but without direction until they become gradually dehydrated. When the lice are placed in humid air, they move at first, but any activity they exhibited soon ceases and they become quiet. Wood lice placed in a container with dry air at one end and humid air at the other gradually congregate at the humid end. This transfer is achieved through apparently random rather than directed movements.

Oriented reflex activities of entire organisms include tropisms, taxes, and orientations at an angle. Tropisms in animals are those directed growth-curvature movements of sessile (*i.e.*, sedentary) forms that lead to equal intensities of stimulation of symmetrically placed body parts. These movements are demonstrated by hydroid animals such as *Eudendrium.*

Taxes. Taxes may be described as oriented locomotory reactions of motile organisms. They exist in purest form as oriented, forced movements; that is, as reflex actions of entire organisms. When exposed to a single source of stimulation, the body is oriented in line with the source. Movement toward the source is said to be positive; that away from it is negative.

Klinotaxis is the achievement of orientation by alternate lateral movements of part or all of a body; there appears to occur a comparison of intensities of stimulation between one position and another and a "choice" between them. Klinotaxis is shown by animals with a single intensity receptor such as the protozoan *Euglena,* earthworms, and fly larvae. For several days before going into the pupal (or resting) state, the blow fly maggot tends to move away from a light source. As it crawls, it swings its head alternately left and right. Presumably a light receptor is located on the maggot's head, and differences in intensity between successive light stimuli as it moves its head determine the direction in which the maggot travels. This type of response is given more commonly to chemical stimuli than to light.

Blow fly maggot response to light

From N. Tinbergen, *The Study of Instinct;* The Clarendon Press, Oxford

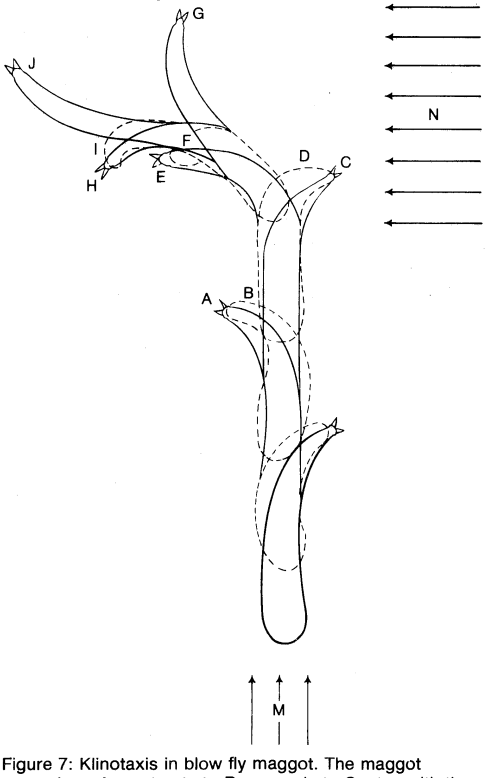


Figure 7: Klinotaxis in blow fly maggot. The maggot expands to A, contracts to B, expands to C, etc., with the light M on. At D, M is put out and N put on. The lateral movement to E is somewhat larger than A; the next one, G, increases the light falling on the receptors and is corrected by a swing over to H.

In tropotaxis, attainment of orientation is direct, resulting from turning toward the less stimulated (negative) or more stimulated (positive) side as simultaneous, automatic comparisons of intensities on two sides of the body are made. No deviations (trial movements) are required. Tropotaxis is shown by animals with paired intensity receptors. If exposed to stimulation from two sources, orientation is to some intermediate point and is determined by the relative intensity of the sources. If one receptor is effectively covered, the animal moves in spirals (circus movement). Tropotaxis is shown by many arthropods, especially insects.

In telotaxis, known only for responses to light, attainment of orientation is direct and without trial movements. When between lights from two sources, the animal orients to one light, rather than to some intermediate point. The animal switches orientation from one source to the other at unpredictable intervals and consequently follows a zigzag course. The response is given to the source as though it were a goal. Bilateral balance is not necessary, and circus movements, if they occur, demonstrate that the animal is reacting tropotactically rather than telotactically. Honeybees (*Apis*) and hermit crabs (*Eupagurus*), among others, show telotaxis.

Orientations at an angle (transverse orientations) may or may not be accompanied by locomotion. They include the light-compass reaction (menotaxis) and dorsal (or ventral) transverse reaction. Menotaxis is shown by foraging insects such as ants and bees that return to a fixed nest. It has been demonstrated experimentally by covering for 2½ hours an ant returning to its nest. After being uncovered, the ant proceeds not toward the nest but at the same angle to the Sun that it had been moving at the time it was hidden from the light.

In another demonstration of menotaxis, the sea slug *Elysia viridis* has been shown to move at angles of from 45° to 135° in relation to a steady source of light. No satisfactory explanation for this type of response in the sea slug is known.

Dorsal transverse action

Dorsal (or ventral) transverse reaction is demonstrated when the impact of the stimulus is kept at right angles to both longitudinal and transverse axes of the body. Locomotion need not occur. This reaction is given to light by various aquatic crustaceans—*Argulus,* the fish louse, and *Artemia,* the brine shrimp—and is given to gravity by crayfish. (Ed.)

Instinctive behaviour

The concept of instinct has come to refer to complex unlearned behaviour that is recognizable and predictable in at least one sex of a species.

CHARACTERISTICS OF INSTINCTIVE BEHAVIOUR

Heritability. Instinctive behaviour is largely heritable. Many of the activities of a species of animal are sufficiently constant and predictable to serve as specific characteristics in the same way, and often to the same degree, as do bodily structures. Examples of such heritable traits include the display movements of birds (*e.g.*, peacocks), the web-spinning movements of spiders, the burrowing habits of marine worms, the prey-catching techniques of weasels or wolves, the food-hoarding movements of squirrels, and the browsing methods of antelope.

The genetic or inherited nature of instinctive behaviour is particularly evident in aggressive and submissive sexual behaviour and in fighting of various kinds. Other instinctive activities of this kind serve nutrition, including methods of obtaining and eating food; care of the body surface, including cleaning, grooming, and scratching movements; escape from predators, including methods of concealment, freezing or "playing dead," and taking flight; social behaviour, including ways of responding to others both sexually and regardless of sex; and sleep, including the rhythms of rest and wakefulness and bodily positions assumed in sleep.

Many of these relatively fixed, species-characteristic types

of behaviour appear to be primarily inherited; at first sight, at least, they may seem little influenced by the particular experiences of the individual animal. But much instinctive behaviour as, for example, playing or exploring, is, nevertheless, modifiable, and the detailed form of the action taken may vary according to the circumstances of the moment and the individual experience previously encountered.

Complexity of pattern. Instinctive activity is not usually a limited response to a simple stimulus but rather is a sequence of behaviour that runs a predictable course; for example, nest-building behaviour that shows a patterned sequence of acts among many birds and some fishes. Many of these actions are far from being either simple or brief. Extraordinary elaboration may be found, and, although some instinctive behaviours may be complete in seconds, others may take minutes, hours, or days.

Adaptive function. Since instinctive behaviour is assumed to be genetically based and therefore shaped by the pressures of natural selection, it follows that most of the consequences of instinctive activity contribute to the preservation of an individual or to the continuity of the species; that is, instinctive activity tends to be adaptive, contributing to the animal's ability to reach maturity and to breed.

Stability under external change. Ideally, instinctive behaviour seems not to depend on learning or practice (see LEARNING, ANIMAL) but to emerge in full complexity without rehearsal when appropriate stimuli or circumstances are encountered. Often, such stimuli do not guide or mold the instinctive behaviour but seem simply to trigger or release it. This characteristic gives instinct the appearance of driving the animal endogenously (from within); the quality of instinctive activity thus appears to depend only secondarily on exogenous (external) stimulation.

Instinct and learning

Experiments in which an animal is raised in a very limited environment (perhaps in isolation, as when a songbird is kept alone in a soundproofed chamber from the moment of hatching) may indicate the extent to which behaviour is spontaneous (or endogenous) or is governed (or triggered) by external circumstances. Rather than seeking to distinguish sharply between instinct and learning, however, it may be more useful to assess the degree to which a given item of behaviour is environmentally stable or unstable (labile). Indeed, aspects of behaviour of a single species may differ greatly in this respect; thus, the nesting activity, calls of alarm, and courting behaviour of birds may be extraordinarily constant under varying conditions. On the other hand, food seeking and feeding may be extremely labile (or susceptible to learning), so that geographically disparate groups of individuals belonging to the same species may have very different feeding habits.

VARIETIES OF INSTINCTIVE BEHAVIOUR

No animal is ever completely isolated from some kind of environment; even in the egg or in the womb it is exposed to environmental variations just as it is after hatching or birth. There is, nevertheless, a sharp distinction between the egg or uterus phase and the free-living, active, sensing animal exposed to environmental stimuli of great diversity. In the case of the egg of an insect enclosed in a largely impervious shell or of a parasitic worm similarly isolated by an impermeable cyst wall, isolation from the environment may be so great that the animal undergoes little more than changes of temperature and oxygen supply. Thus, when such an animal emerges from its egg or cyst, the details of its structure and behaviour would seem maximally attributable to the genetic proteins in its cells; that is, its behaviour appears to be as much inherited as is its bodily structure. This sort of behaviour is environmentally so stable that it is naturally and quite reasonably given the label instinctive.

Reflex activity. A variety of what may be called simple instinctive behaviour has long been known as reflex action. When this term was introduced, it meant the simple and almost invariable response of a simple organ system (*e.g.*, a single muscle) to a simple stimulus, such as a touch or a flash of light. In its most elementary versions, this activity has been seen as the function of an idealized mechanism that has been called the reflex arc. The primary components of the reflex arc have been identified as the sensory-nerve cell (or receptor) that receives the stimulation, in turn connecting (hence the term arc) to another nerve cell that activates the muscle cell (or effector).

Although such a reflex arc might be the simplest imaginable mechanism for inflexibly automatic behaviour, it is, as noted above, a theoretical minimum rather than an actually observed functional arrangement of cells in the body of the animal; nevertheless, a mechanism but little more complicated than this helps to account for the locomotion of such animals as millipedes. In some insects, for example, the stepping movement of one limb or muscle provides stimuli that set off another limb or muscle on a similar course of movement, providing a kind of feedback system or chain of reflex arc activities. In most cases of this sort, however, the basic physiological mechanism is more complicated than the simple arc theory would suggest. Additional nerve cells capable of communicating with other parts of the body (beyond the receptor and effector) are invariably present in reflex circuits. Such connections are what make possible the conditioning of reflex responses.

Among higher animals, and perhaps many others (such as insects), what once were thought to be chain reflexes are not systems simply linked or chained together; they are systems under the precise control of coordinated complexes of nerve cells in particular parts of the nervous system, such as the spinal cord and brain. Even without evidence of a chain of feedback (or reafferent) stimuli, performance may be smoothly integrated. This is well illustrated by the complex movements of swallowing in mammals; in the dog, for example, 11 separate muscles or muscular systems are found to discharge one after the other, precisely timed to a matter of milliseconds, and all under the control of the central nervous system (CNS: brain and spinal cord). Such complexes of precisely controlled movement, known as fixed action patterns (FAP's), are thought to form the hard core of the inborn movement forms of instincts. When such sequences of uniform stereotyped responses seem to constitute an end point or goal-directed climax of some sort, they are known as consummatory acts.

Consummatory behaviour

Fixed action patterns. Some male spiders perform elaborate courtship actions that affect selectively females that are ready to respond sexually. The male, in testing for a receptive female, first stands out of reach and goes through elaborate precise gestures with limbs, pedipalps, and other body parts that are distinctively shaped or patterned in a manner characteristic of the species. Perhaps even more remarkable FAP's are found in the displays of male fiddler crabs of the genus *Uca*, about 40 species of which are distributed over the Earth's tropical ocean beaches. One of the two claws is enormously enlarged, seemingly having been evolved primarily for sexual and aggressive displays, and its ritualistic gestures are quite characteristic

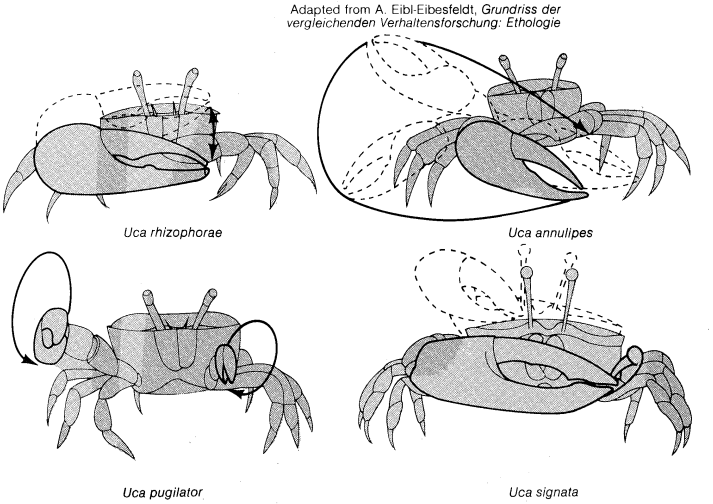


Figure 8: Species-specific instinctive movements of the claws in courtship behaviour movements of fiddler crabs (*Uca*).

of the species. The timing and form of movement are so invariable from one individual to another that an expert can distinguish by this alone the species of crab, whether it comes from the shores of Panama, Tahiti, or Bali.

There must be some very precise, built-in mechanism, presumably some integration of hormonal and nervous-system controls, which ensures that each individual in a species exhibits the distinctive FAP at the right speed, amplitude, and intensity. Beyond those noted for fiddler crabs and spiders, innumerable examples of such instinctive patterns are found among the elaborate display movements of many other animals, such as insects, fishes, and birds. Among the latter three groups, these movements also are part of the species communications mechanism, serving to permit members of a species to distinguish the signals of their fellows and prospective mates from all other visual stimuli. Thus, there is a system of instinctive perceptual abilities at least as complex as the inborn tendencies to exhibit patterns of motor behaviour. Likewise, the perceptual instincts are of primary importance for the survival of the species.

Modifiable action patterns. Some types of instinctive behaviour, while showing a rigid core of fixed action pattern, are still modifiable by conditioning and other learning processes (see above). A good example is provided by the nestbuilding behaviour of many birds: after the breeding female has chosen a nest site, she finds and deposits sticks or twigs or pieces of grass there. A jackdaw (*Corvus monedula*) or rook (*C. frugilegus*) standing on a potential nest locality with twigs held in its beak performs a downward and sideward sweeping movement that brings the material into contact with the ledge or the branches on which the nest is to be built. The moment the twig or branch carried by the bird meets resistance, the sideways movements become more vigorous and merge into a series of quick trembling thrusts (so-called tremble shoving). When the twig is in a position that offers even more resistance, the efforts become more intense until the twig wedges fast. After this consummatory achievement the bird apparently loses interest in his activities for the moment.

Tremble shoving

An inexperienced jackdaw at first will try any objects small enough to be handled, even pieces of ice and the metal ends of small electric bulbs. None of these ever becomes lodged firmly enough by the tremble shoving to result in a stimulus that is sufficiently consummatory to ensure successful nest building. Such failure quickly extinguishes the bird's tendency to fetch inadequate objects; equally rapid positive conditioning is effected, and the jackdaw learns to be a twig connoisseur, coming to use only those that are just right in shape and flexibility. Indeed, it has been observed that the nests of entire groups of such birds are predominantly constructed of twigs and other pieces that are taken from only one kind of tree, even though there are other building materials that are readily available.

In contrast to jackdaws, many small songbirds do apparently have an inborn tendency to select the kinds of materials that are appropriate for different phases of nest construction. This innate predilection for suitable building materials has been shown dramatically among canaries reared in man-made nests of felt. Even though female canaries thus reared have never encountered anything long and flexible before, when nest-building time comes, they can select materials appropriately. As soon as pieces of grass, bits of string, cotton, or any long flexible objects are placed in the cage, these female canaries display interest and, within seconds, carry the objects to the nest place and commence weaving movements. Once the proper state of construction has been reached, and not before, the birds display an innate tendency to line their nests with feathers, plucking out their own when no others are to be found.

A caged female canary long deprived of nesting material may take hold of one of its own feathers in its beak and, without detaching it, go through the motions of lining the nest with it again and again without, of course, actually lining the nest. Another striking example of complex nest building is found in the extraordinarily complicated movements and responses by which weaver birds build their elaborate hanging nests with such architectural features as roof, egg chamber, antechamber, and entrance tunnel. Research has identified an elaborate system of relations among external stimuli, internal hormonal conditions, neural function, and reproductive development in the behaviour of female canaries during the course of their breeding cycle. Evidence of an elaborate combination of innate physiological activity and individual experience also comes from studies concerning the development of songs and call notes among birds.

A young bird isolated from other members of its kind, or even more rigorously from all patterned auditory stimulation, produces an extremely limited, basic sound pattern. If the young bird is allowed to develop and sing along with other members of the species, however, the instinctive tendencies seem to be more fully realized through fine adjustments added by imitative learning.

By courtesy of N.E. Collias and E.C. Collias

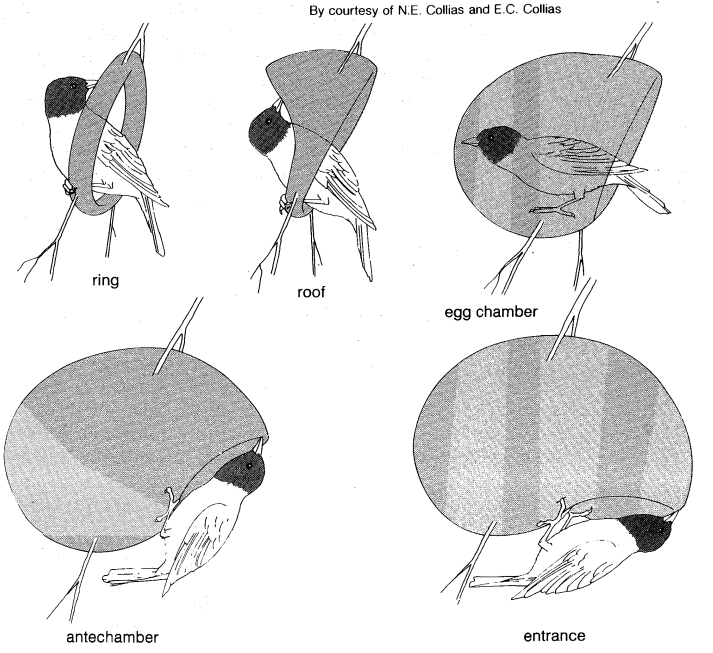


Figure 9: Five stages in the construction of a nest by the male village weaver (*Ploceus cucullatus*).

With more complexity in brain structure and function as is found in many mammals, behaviour is relatively flexible, and fixed action patterns tend to be overlaid by learned patterns and to that extent obscured; nevertheless, an inexperienced brown rat deprived of nest material tends to use its own tail instead, carrying it about in behaviour that is reminiscent of that shown by a feather-deprived canary. After gathering materials, brown rats typically heap them up in a more or less circular wall and then begin to tap down and smooth the inner surface of the nest cavity. When a very inexperienced rat is offered paper strips or other soft material for the first time, it goes into a random frenzy in which the sequence (gathering, heaping, and smoothing) is confused. Yet, each of the three phases is performed to perfection, not differing even on analysis by slow-motion films from those of an experienced rat. After having placed two or three paper strips together flat on the ground, however, the novice rat will perform heaping-up movements in the empty air above them, then apparently pat down a nest wall that is not yet in existence! Only later, after the animal has had more experience as a builder, does the rat seem to learn to inhibit the instinctive heaping and patting movements until the appropriate stages of construction have been reached.

Nesting rats

The catalog of instinctive behaviour among animals is much richer than the few examples offered here. Despite the clear differences observable in the detailed manifestations of instincts in comparing frog, goldfish, pigeon, cat, rabbit, and man, for example, all of these behaviours may be seen to hinge on genetically transmitted physiological structures and functions. Indeed, many animal instincts may profitably be compared with some of the built-in forms of behavioural tendency among plants. (W.H.T.)

Periodic biological phenomena

Continuous change characterizes both living organisms and their physical environment. Many of these changes occur irregularly and are termed aperiodic. Examples include the irregular variations in temperature, light, humidity, and other physical factors associated with the passage of weather systems; aperiodic, also, are the biological fluctuations in response to them. In contrast, are periodic fluctuations, relatively regular changes repeated at constant intervals of time. Clearly periodic, for example, are the natural fluctuations in daylight and in tides. Similarly periodic are the daily and tidal variations in both animals and plants as they respond and adjust to these periodic changes. Phenomena that recur with such regularity are termed rhythmic. Fluctuations of organisms in direct responses to long term environmental fluctuations are termed phenological. Examples of phenological phenomena include changes such as migratory movements and breeding periods.

The range of biological rhythms

Rhythmic variations in living systems are commonplace and span a very wide gamut of intervals. Among those possessing short periods and hence a high frequency of recurrence are wing beats of insects (20 to 2,000 cycles per second), brain waves (1 to 60 c/sec), walking and chewing rhythms (1 to 8 c/sec), heartbeats (20 to 1,000 c/min), and respiratory rhythms (4 to 250 c/min).

Other biological rhythms have only a few cycles per day. Examples are the variations in activity of very young infants, in depth of sleep of adults, in metabolism of chick embryos, in spiralling of growing plant tips, and in a number of bodily processes of the human during certain illnesses. Longer periods include the well-known several-day recurrences of fever in malaria and the few-day to several-month reproductive rhythms of mammals. Still lower frequency rhythms include the several- to many-year cycles of reproduction in the periodic cicadas, of abundance in animal populations, of prosperity in business, and of outbreaks of wars.

Fluctuations in living creatures occur at every level from the cell to activities of the organism as a whole and are evident in altered chemistry, physiology, and behaviour, including changing responsiveness to all factors of the environment.

Periodic biological phenomena fall into two more or less distinct different categories: those correlated with periodic changes of the planet (geophysical correlates) and those having no such correlates. Biological rhythms correlated with ocean tides, days, months, and years clearly relate the activities of organisms to the fluctuations of their physical environment as a consequence of the movements of the Earth. These rhythms possess properties that clearly set them apart from rhythms without external correlates, such as heartbeat and respiratory rate. The frequencies of these latter rhythms can be altered in response to changes in the immediate environment. Their frequencies also are altered readily by drugs and other chemicals and, as with metabolic rate itself, altered greatly as body-temperature changes. On the contrary, rhythmic variations with geophysical correlates have relatively fixed periods and are resistant to alteration in their frequency by either temperature or chemicals.

This extraordinary stability of the periods of the biological rhythms with geophysical correlates is not dependent upon direct responses to light and temperature rhythms. It was demonstrated more than two centuries ago that the capacity of plants to display their daily rhythm in sleep movements, the daytime raising and nighttime lowering of their leaves, persisted even when the plants were shielded from light changes and maintained in a relatively constant temperature. More recently, daily rhythms in a wide variety of living things, from single-celled forms to mammals and flowering plants, have been shown similarly to continue under constant conditions of light and temperature. Tidal rhythms in many shore organisms persist even after the organism is removed from the tidal environment. Comparably, semimonthly, monthly, and annual rhythmic variations continue even when organisms possessing them are deprived of every obvious periodic cue. Such organisms behave as if they possess the ability independently to time all Earth's major natural periods.

Adaptive significance of periodicity

A timing capacity, geared to the rhythmic variations of the natural environment, is indeed a most useful attribute. A time sense enables organisms to anticipate and exploit the most favourable portions of the day, tide, month, etc. It can be used by members of a species to synchronize their urge to breed. A time sense may serve, in conjunction with the Sun, Moon, and stars, as a means of geographic orientation for the migratory and homing activities of many organisms, just as a chronometer and a sextant enable a navigator to find his position or to plot a course.

BIOLOGICAL RHYTHMS AND NATURAL GEOPHYSICAL CYCLES

Living things are extraordinarily well adapted to their rhythmic environments and have become periodic in diverse aspects of their physiology and behaviour. The biological rhythms that have resulted are in part direct responses to the environment and in part indirect ones via the complex clock-timed organization.

Solar-day rhythms. Twenty-four-hour rhythms of organisms in nature are the rule; they are variously called solar day, circadian, diel, daily, diurnal, and nycthemeral. The evolution of life on Earth has proceeded in the direction of maximal utilization of every habitable niche, with species specialized to deal with virtually every kind of environmental site. This specialization is related not only to space but also to time. Some organisms have become specialized for the darker, cooler, more humid nighttime; others for the lighter, warmer, drier daytime; and still others for the transitional twilight periods. The environment is in use full-time, but for maximal efficiency organisms

Adapted from F.A. Brown, "Living Clocks," *Science*, vol. 130, pp. 1535–44 (December 4, 1959), in R.L. Smith, *Ecology and Field Biology* (1966); Harper & Row

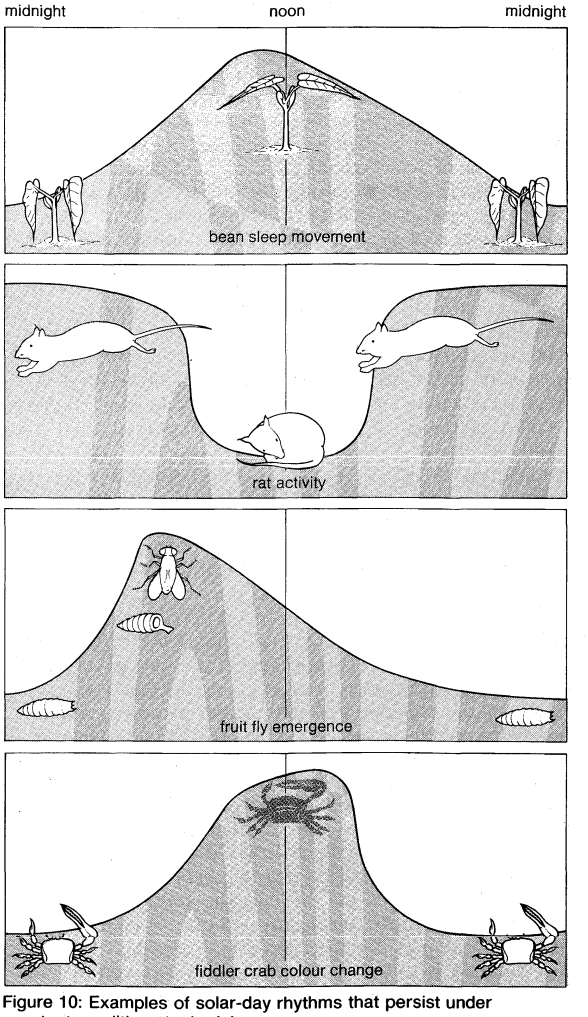


Figure 10: Examples of solar-day rhythms that persist under constant conditions in the laboratory.

are organized to work largely in "shifts." The night shift includes animals such as mice, cockroaches, scorpions, moths, owls, bats, skunks, opossums, and raccoon; the day shift includes songbirds, hawks, lizards, butterflies, and honeybees. Green plants alternate between the daytime use of sunlight in photosynthesis and the nighttime emphasis on growth and assimilation. Many plants undergo a daily sleep rhythm, their leaves drooping at night and rising by day. Some plants open their blossoms only at night, others only by day, with times commonly synchronized to the activity schedules of animals that pollinate them.

In each organism's daily pattern, a multiplicity of supporting chemical and physiological variations occur as the organism's whole being is thrown periodically into activity. In man, as an example, the daily alternation of sleep and wakefulness is accompanied by many changes, including activities of the nervous and endocrine systems and the liver and kidneys. Daily variations occur in body temperature, in heart and respiratory rates, and in pressure and composition of the blood. There are daily variations, also, in chemical syntheses and in cell divisions. Such rhythmically recurring phenomena are each set to the most efficient time of day for the organism and collectively contribute to the characteristic 24-hour phase map, or activity record, of that individual (Figure 10).

Daily variations in responsiveness

One of the very striking manners in which the presence of fundamental 24-hour variations is reflected in man and other animals is in changing responsiveness or susceptibility to physical and chemical insults. The following are a few examples. Doses of X-rays that at one time of day can kill every individual exposed will at another time kill only a few individuals or none. Doses of powerful poisons, fully lethal at one time of day, leave animals unharmed at another. Noise that can induce fatal convulsions in mice at one time of day fail to do so at another. The ability of histamines to induce inflammatory response, and the effectiveness of alcohol, insulin, and even aspirin vary systematically with time of day. In insects the degree of resistance to insecticides or to exposures to high temperatures also depends upon the time of day.

Lunar tidal rhythms. Another major environmental period is the lunar tidal rhythm, the regular ebb and flow of the ocean tides, a cycle that subjects seashore plants and animals to a rhythmic change with typically two high and two low tides occurring each lunar day of about 24 hours' 50 minutes' duration. Some intertidal organisms, such as barnacles, green crabs, snails, clams, and oysters, are most active when submerged by the rising tide. Others, such as shorebirds and fiddler crabs, are especially adapted to feed on beaches exposed at ebb tide. Some single-celled algae called diatoms move down into the sand and mud when the beaches are submerged and rise onto the surface to permit photosynthesis when the tide recedes. The diatoms and many of the shorebirds are active only when the beaches are exposed during daylight. Such organisms, therefore, display simultaneously 24-hour and tidal rhythms, which steadily undergo changing phase relations with respect to one another, coming into the same relationships exactly once each half-month. Indeed, all the organisms of the intertidal region must deal with both the daily light and temperature changes and the tidal ones. Fiddler crabs, for example, show concurrently a daily rhythm of skin colour change adaptively set to the times of day and night and a tidal rhythm of locomotor activity adjusted to the local tidal schedule.

Although the ocean tides are affected chiefly by the Moon, the gravitational attraction of the Sun also has an influence. The Sun and Moon "cooperate" at new and full moons to produce the greatest tidal ranges of each month, the so-called spring tides. At the Moon's quarters the Moon and Sun maximally oppose one another, resulting in the smallest tidal ranges. As a consequence, inhabitants of the seashore that occupy the highest and the lowest levels of the intertidal zone experience periodic tidal submergences or exposures only for brief intervals once each half-month. One species of marine snail that inhabits the upper level of the beaches has been reported to show a semimonthly activity rhythm.

Any living creature that possesses both tidal and daily variations will exhibit day-by-day patterns of variation that systematically change from one day to the next as the peaks and troughs in the tidal cycles, like the tides themselves, occur at progressively later times of day. The specific pattern of organismic variation through any given day will repeat itself exactly a half month later.

Animal navigation

An interesting expression of these rhythmicities is evident in the reported phenomenon of celestial navigation by animals. Such animals as birds, fishes, turtles, spiders, insects, and crustaceans have been variously described to navigate or to "home" by using the relative positions of the Sun or the Moon or, in the case of birds, even the constellations as geographic navigational guides. Such Sun-compass, Moon-compass, or star-compass navigation requires that the animals be able systematically to alter their orientational angle at rates to compensate exactly for the Earth's rotation relative to each of these celestial references.

Monthly rhythms. The synodic month, averaging 29.5 days, and the semimonth can be sensed by organisms, as noted earlier, through "beats" between their lunar tidal and circadian rhythms. Another kind of environmental monthly variation involves nighttime moonlight. Synodic monthly variations occur in such unlikely phenomena as rat and hamster running activity, light responsiveness of guppies, human colour vision, potato metabolic rate, and light responses of freshwater flatworms, even when shielded from all moonlight.

Among the biological activities linked to the synodic month are reproductive cycles of many different kinds of marine animals and plants. Best known and most studied of these are the monthly reproductive rhythms of palolo worms of the southwest Pacific, the fireworms of Bermuda, the grunions of the California coast, and the brown marine algae *Dictyota.* These rhythms are sufficiently precise in their relation to Moon phase and season of the year as to be predictable. The changing relations of the solar-day and lunar-day physical rhythms appear to time these breeding cycles, since the phase of Moon at which, for example, *Dictyota* breeds is related to the times of occurrence within the lunar day of the local high tides, which differ from place to place. The grunion's breeding is timed to within a few minutes of high tide on the local beach on the nights of its breeding, and the breeding times of palolo worms and fireworms are very accurately timed to within a few minutes in the day.

Precision timing of sea lily reproduction

The most spectacular precision in timing of a month- and year-related reproductive rhythm has been described for a deep-sea sea lily near Japan. This echinoderm liberates its sex cells once each year in October at about 3 PM on the day of one of the Moon's quarters. In succeeding years the time of sex cell release changes, among the Moon's two quarters, first-third-first, to progressively slightly earlier dates in October. The triplets are repeated until about the first of the month whereupon the following year it jumps abruptly to near the end of the month to start the advancing triplet progression again. The result is an 18-year cycle, which is essentially the period of regression of the Moon's orbital plane. There have been reports of monthly menstrual cycles in some primates with initiation of menstruation linked to the new moon. Reproductive swarming of mayflies has also been related to Moon phase.

Little is known about the timing of such cycles as the human menstrual cycle, the approximately semimonthly estrus cycles of sheep, the three-week cycles of cattle and pigs, and the six-month cycles of dogs. Three kinds of hypotheses have been offered for their timing: (1) they are timed by fully independent periodic oscillations within the organism; (2) they are dependent for their timing on periodic interference between 24-hour and lunar-day clocks, since most of them appear to be simple multiples of a quarter month; and (3) they depend for timing on monthly variations in moonlight (directed chiefly at the primate menstrual cycle). Some recent studies have given strong support for a long-held belief that the menstrual period is in some manner regulated by the Moon ("menses" means lunar month). In a study of a number of women with variable onset of menstrual periods, artificial illumination of the bedroom through the 14th to 17th nights following

the onset of menstruation resulted in a regularization of the period, with the period length coming very close to 29.5 days, the natural synodic month. That this period is a biologically significant one for the human species is further suggested by the fact that the average duration of pregnancy (from ovulation to birth) in the human is rather precisely nine 29.53-day synodic months.

Associated with the human menstrual cycle are related variations in almost every measurable bodily factor: temperature, blood chemistry, weight, heart rate, blood-cell counts, and pain thresholds are among the described variations. These factors pass through maxima and minima at different times but in such characteristic time relationships to one another that phase maps for the menstrual rhythm, comparable to those for the circadian, can be determined.

Annual rhythm. Another major rhythmic period for living things is the year. In most parts of the world substantial seasonal changes in light and temperature are correlated with rhythmic biological phenomena predominantly related to activity, reproduction, and growth. For example, in higher plants in the temperate zones, growth, flowering, and seed production typically occur over the course of the warmer months. In the oceans and fresh waters there are annual rhythms in plankton abundance. Annual reproductive rhythms in animals are typically set to the time of year that provides the optimum opportunity for birth and rearing of young. Annual rhythms include also hibernation to bridge the coldest period of the year and estivation to avoid hot-summer droughts.

The readily observable systematic annual variations in the activities of the plants and animals are accompanied by innumerable less obvious changes within their bodies, variations that pass through their peaks and depressions with such regularity that annual phase maps can be made.

Seasonal variations in human biology

Seasonal rhythms occur in humans and are evident in such phenomena as birthrate and weight and in growth rate of children. There are annual variations also in innumerable other factors, among them body temperature, heart rate, basal metabolic rate, blood-cell counts, and blood chemistry. Not surprising, also, are reports of annual variations for many systemic noninfectious diseases, and even for mortality. Many human infectious diseases are well known to exhibit peaks at characteristic times of year, while diseases without such evident annual fluctuations could perhaps be expected to display an annual variation in degrees of complications and mortality as a consequence of a probable annual variation in susceptibility related to the rhythmically changing biochemical and physiological states.

Too little investigated for man but difficult to treat because of the long period involved is the degree of phase shifting possible for the annual rhythms by manipulation of such physical factors as photoperiod and temperature.

Epochal rhythms. There are innumerable natural geophysical periods both shorter and longer than a year resulting from periodic activities in the Sun and from relative movements of the Earth, Sun, Moon, planets, and stars. Among the well-known shorter periods are the Sun-generated 10-cycles per second electromagnetic fields, suggestively similar to the 10-cycles per second alpha rhythm of the human brain. Variations in climate, rainfall, temperature, and light have been claimed to be correlated with many longer periods. Among the periods, and perhaps the best known, are the 11- and 22-year sunspot cycles, the 18.6-year period of lunar "wobble" (nutation), and the 26,000-year precession of the Earth's rotational axis. Beyond the solar system are the fluctuating galactic fields. Probably the longest physical period that has been suggested is the approximately 250,000,000-year period of rotation of the galaxy, which some believe to be correlated with terrestrial glacial periods. All these periods are accompanied by subtle alterations in the gravitational and electromagnetic fields of the terrestrial environment and so influence all living things.

Numerous several- to many-year biological periodicities involving a wide variety of phenomena have been described. Perhaps most familiar are the three–four-year cycles of lemming migrations and the eight–10-year cycles in abundance of the Canadian lynx, varying hare, and ruffed grouse. Still other biological periodicities described from tree rings, shell deposition in mollusks, fossil depositions, business, and wars have been claimed. Very commonly reported are cycles of about nine, 11, 18, and 22 years. Although it seems plausible that such geophysically correlated biological periodicities are real, their long-period durations, together with their cycle irregularities, tend to make difficult the critical establishment of their exact lengths or causal associations with environment periodicities.

THE BIOLOGICAL CLOCK

Natural time. Many organisms, man included, when removed from their natural surroundings and placed in a laboratory under constant conditions, can often repeat a rhythmic pattern that they had experienced before such isolation. It is concluded from such behaviour that the organism can time the length of the natural day and even determine the points and measure durations of relatively detailed periods within a single day. This timing capacity is essentially uninfluenced by temperature changes and all ordinary drugs and chemicals. Every living system studied behaves as if it contains a highly dependable clock. The specialized properties of this clock suggest that it involves some unique biological mechanism indissociable from life itself; however, the clock can be investigated only indirectly, through observing the more conventional biochemical, physiological, and behavioral processes timed by it.

Biological clocks that time the solar day, the period of the Earth's rotation relative to the Sun, have been the most extensively studied. Except possibly for the bacteria, the same kind of clock exists in members of every major group of living creatures. Single-celled animals and plants appear to possess just as accurate and effective clocks as do the most highly differentiated multicellular organisms.

A number of general properties of the clock-timed solar-day rhythms have been discovered. One such property is that under constant conditions in the laboratory the rhythms usually change slightly, becoming instead a little longer or shorter than 24 hours. This commonly observed deviation from the 24-hour period under these conditions led to the concept of circadian (about a day) rhythms. (The explanation for such variance is discussed below.) In day-active (diurnal) vertebrates circadian rhythms, in constant light and temperature, are typically shorter than 24 hours, while in night-active (nocturnal) ones they are typically longer. The periods of circadian rhythms under these conditions rarely exceed the range of 20 to 28 hours and usually are within an hour or so of 24 hours.

The basic circadian rhythm

When an organism is returned to its normal 24-hour light–dark environment, it soon resumes an accurate solar-day rhythm. The transitions from light to darkness and darkness to light appear to be the signals for readjusting the organism's daily physiological cycles. A nocturnal mammal, for example, adjusts its time of activity to the time of darkness, a diurnal songbird to the time of daylight. Such resetting or phase-adjusting factors for the rhythms are termed daily clues, phase setters, or synchronizers.

Involved in adjusting an organism's daily activity pattern to the appropriate time of day is a rhythmic variation in the influence of the light as a resetting factor. This variation has been termed a response curve. A light change near the time of the beginning of an organism's activity pattern advances slightly the phases of the cycle to an earlier time of day; a comparable light change near the end, on the other hand, extends it. At other times during the cycle comparable light changes have little or no phase-shifting influence. Although response curves appear to possess the same general form regardless of the kind of organism, each individual possesses its own detailed form of response curve, which adjusts its activity pattern to the light–dark cycles in its own characteristic manner.

Whereas light appears to be the dominant phase setter in adjusting clock-timed rhythms to natural 24-hour cycles, temperature is also effective. Many organisms are able to adjust their activity patterns to 24-hour rhythms in temperature in the absence of light variations. Still other phase setters have been reported, including sound, feeding time, and social interactions.

Free running rhythms

Modifiable clock-timed rhythms. The circadian rhythms that continue under experimental conditions of constant light and temperature are termed free-running; they reflect the underlying natural rhythms of an organism. The free-running period can usually be altered, but only very slightly, by changing the intensity of the illumination or the temperature. A 10° C (18° F) temperature increase seldom changes a circadian period more than 10 percent, and in most organisms the period is shortened. For fiddler crabs in constant darkness the free-running period, of exactly 24 hours, remains unchanged over a 20° C (36° F) range. In contrast to these relatively resistant rhythms, the rates for ordinary biochemical or physiological processes typically double or triple for each 10° C rise in temperature. Free-running rhythms sometimes cease spontaneously, only to resume at a later time as if the rhythms had been continuing undetected during the intervening time. In other words, a clock-timed phenomenon within the organism may become uncoupled from the underlying clock. Circadian rhythms sometimes fade out over the course of a few cycles under constant conditions; in other cases they appear to continue indefinitely, sometimes even with increasing amplitude and sharpness.

Although chemical inhibitors or stimulants of general metabolic processes can influence rhythmic processes, they appear not to affect the underlying clock in any way. Some powerful metabolic inhibitors such as cyanide will eliminate all overt rhythmic variation in an organism, but upon removal of the inhibitor, the rhythm returns as if it had been uninterrupted. On the other hand, exposure to near-freezing temperatures inhibits rhythmic changes in such organisms as fiddler crabs and cockroaches; but when the temperature is elevated again, the restored rhythms have been phase shifted by the approximate duration of the time of chilling.

There appears to be a genetic component to circadian organization. Organisms reared from eggs in constant light and temperature may display circadian rhythms spontaneously or may be induced to do so by a single abrupt light or temperature change, a signal that itself conveys no information as to period length. Furthermore, bean plants selected and interbred for longer and for shorter circadian periods under a given set of constant environmental conditions will beget offspring with, respectively, longer and shorter circadian periods under these same experimental conditions. The length of the free-running period is clearly genetically determined in part.

The influence of heavy water, which is D_2O, on circadian rhythms is of particular interest. Treatment of organisms as different as algae and mice results in a lengthening of the free-running periods. Under a 24-hour, light–dark regime, heavy water appears to alter the phase relations of the activity patterns of organisms to the light cycles.

Clock-timed circadian rhythms have been observed during rapid geographic translocations of organisms eastward or westward by airplane to other time belts. Investigations of this type, performed with fiddler crabs and honeybees, confirmed that the organisms carry circadian rhythms still adjusted to the light cycles of their former location (Figure 11). To reset them to the new local time requires several days of exposure to the local phase setters. This phenomenon is quite familiar to persons who have travelled long distances rapidly by jet airplane. The resultant rhythmic dislocation and the need for gradual adjustment over two to ten days at the end of such a trip is often referred to as the "jet syndrome."

Time dislocation: The "jet syndrome"

It is well known that, among the innumerable periodic changes that underlie and support the overt rhythms of behaviour of organisms, the peak values occur in a characteristic sequence over the day. A description of these comprises what is called a phase map. Such a sequence and spacing reflects the order and temporal relationships of cause–effect in the normal interactions of the various bodily processes. Phase maps may undergo transitory disruptions when an organism is compelled to make a rapid phase adjustment as, for example, after a rapid move to a new geographic longitude. Under such circumstances the various individual 24-hour components comprising the circadian phase map do not reset their phases to the new environmental times at the same rate. They may become somewhat displaced in their relations to one another. This dissociation of rhythmic components is probably the major cause of the fatigue and the lowered efficiency that characterizes the jet syndrome in human beings.

Another remarkable property of free-running rhythms is their capacity sometimes to maintain well beyond chance, and for extended periods, precise 24-hour periods and, in some cases, even to set their active period to their normal time of day without any obvious daily clues. The former has been described for fiddler crabs, gila monsters, mice, and kangaroo rats. The latter capacity has been observed in chicks and some lizards. These animals are apparently able to set their newly developed behavioral system to times of outside day and night by the use of subtle and as yet not fully identified signals.

The biological clock also times other natural environmental periods. Many plants and animals whose activities vary rhythmically with the tides have been found, when taken into the laboratory, to continue rather accurately their tidal rhythms. The period of these rhythms is about 12 hours 25 minutes, or 24 hours 50 minutes for the double-cycle, which is called a lunar day. The period of the tidal changes is correlated with the rotation of the Earth relative to the Moon. The tide, as does the Moon, rises on the average of 50 minutes later each day.

The properties of the tidal, clock-timed rhythms differ in

From F.A. Brown, Jr., *Biological Clocks*, BSCS-Educational Programs Improvement Corp.

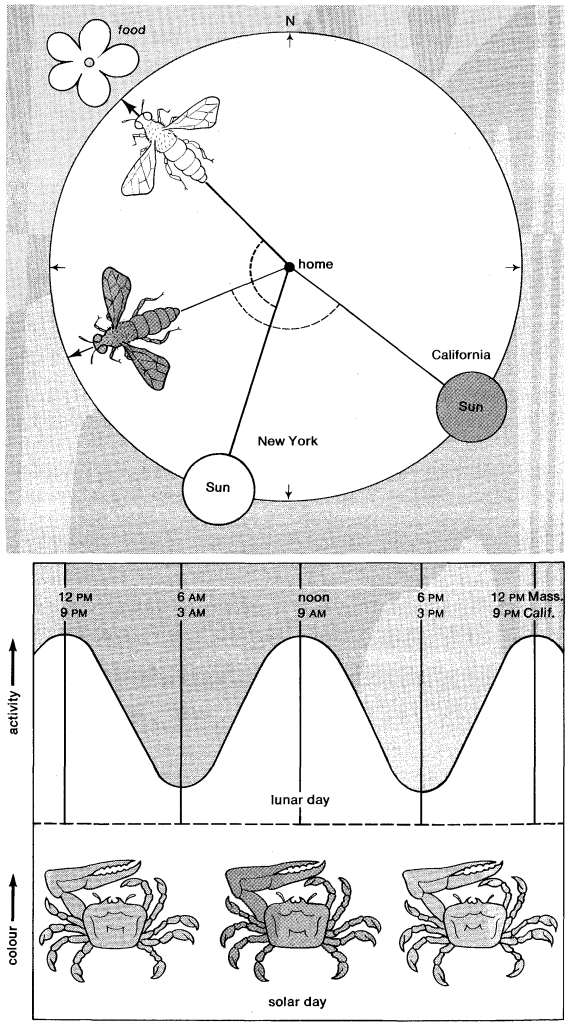


Figure 11: *The persistence of natural rhythms.* (Above) Bees trained in New York to fly northwest to a feeding station at 1 PM fly southwest at 10 AM when transported to California, following the sun's position as learned in New York. (Below) Fiddler crabs, moved from Massachusetts to California and kept in unvarying illumination, show activity rhythms and colour changes identical to their East Coast counterparts, regardless of the local time.

significant ways from the circadian ones. Although they possess the same high degree of temperature and drug independence of their period lengths, they will not synchronize to a 24-hour light–dark environment. Instead, tidal rhythms continue to move over the day at the tidal rate. This failure to follow the light cycles is obviously adaptive, since in nature the phase relationship of the tides to time of day systematically changes.

The clock-timed tidal rhythm can be adaptively adjusted to the tidal schedule of a specific beach. Fiddler crabs transported to a foreign beach with different tidal times gradually, over a few days, reset their tidal rhythms accordingly. The phase-resetting clues for the tidal rhythms remain unknown, though a number of factors such as mechanical wave action and periodic drying have been suggested.

Drifting of tidal rhythms

Clock-timed tidal rhythms commonly exhibit a tendency, under constant conditions of the laboratory, to drift away from their tidal setting until they approximate the actual time of lunar day. Oysters and fiddler crabs taken from their home beaches and kept in the laboratory gradually, during a week or so, reset their tidal rhythms to display their maximum activity at the times of the upper and lower transits of the Moon, as if some subtile synchronizers associated with the lunar tides of the atmosphere were now replacing the unknown local tidal ones.

There are clocked-timed synodic monthly rhythms of 29.5 days, the period from one new moon to the next. This is the period produced by "beats" between the solar days and lunar days—in other words, the interval between the times when the solar day and lunar day begin together. The Moon rises close to the same time of day once each synodic month. The organismic clock-timed monthly rhythms in some instances appear to depend upon a comparable interaction between tidal or lunar-day rhythms and solar-day ones. Interactions between 12.4-hour tidal rhythms and 24-hour circadian rhythms would yield semimonthly "beats," and between 24.8 hour double tidal cycles and 24-hour ones monthly "beats." Evidence for such clock-timed rhythms is seen in the occurrence of semimonthly colour changes in crabs kept under constant conditions. These exhibit phase differences among populations in their semimonthly rhythms accountable in terms of the differing phase relationships between their local tides and time of day.

The clock that times monthly rhythms appears to be just as dependable as those timing circadian and tidal rhythms. Monthly rhythms can, comparably, be phase shifted relative to the actual Moon phases. In addition to the phase setting by changes in some factors associated with local tidal movement, moonlight variations through the month have been suggested as a phase-setting factor for semimonthly and monthly rhythms, particularly the breeding rhythms of sea worms and the menstrual cycle for primates, including women.

Clock-timed annual rhythms, persisting in unvarying regimes of illumination and temperature and every other obvious clue as to the yearly period, have been reported for widely different kinds of organisms and phenomena. An alga, on the one hand, and the laboratory rat, on the other, have displayed annual variations in enzymatic activities. Seeds show annual variations in capacity to sprout, and newly germinated seedlings exhibit annual variations in their metabolic rate. Some birds show rather precise annual, and others *circa*-annual, reproductive cycles. Food intake of woodchucks has been reported to vary systematically with time of year. As for the other periods, the clock that times annual rhythms appears to be independent of temperature and uninfluenced by all attempts to alter its period chemically. The normal patterns of annual variation of many plants and animals can, however, be altered in response to changes in relative lengths of light and darkness (photoperiodism), much in the manner analogous to the phase resetting of the clock-timed circadian rhythm by times of light onset and termination. On the other hand, annual rhythms such as that of woodchuck food intake appear to be unalterable in phase by manipulation of either temperature or light.

Persistent clock-timed rhythms. There is another class of periodicities in organisms, rhythms that persist under the most carefully controlled laboratory conditions but that possess properties different from those described in the preceding account. These have precise solar-day, lunar-day, monthly, and annual periods. These have been discovered for respiratory or activity rates in a wide variety of animals and plants. The activity of an organism, other factors equal, varies systematically with local time of day, time of lunar day, time of month, and time of year. The potato, for example, shows a clear-cut sidereal-day (23-hour 56-minute period of rotation of the Earth relative to the stars) variation in its metabolic processes. These periodisms can not be reset or phase shifted. Moreover, there is no evidence to suggest that they are adaptive. The range of their variations is least for the sidereal-day and most for the annual period.

Fluctuating activity. Aperiodic, hour-by-hour fluctuations in metabolic rate or activity of organisms under presumably constant conditions continue to show correlations with weather changes, indicating beyond reasonable doubt that changing physical factors somehow continue to reach and influence the organisms, even in their experimental isolation. How do these external factors pervade the controlled environments of the laboratory? The answer perhaps may be found in a very highly specialized sensitivity of organisms to weak geoelectromagnetic fields that pervade the universe. Aperiodic biological fluctuations do coincide with fluctuations in such factors as background radiation, terrestrial magnetism, and primary cosmic radiation.

Influence of magnetism and radiation

The exogenous derivation (external to the organism) of these variations is indicated by the universality of response: the rhythmic patterns of these geophysically dependent fluctuations tend to show, among organisms spanning the gamut of kinds of living creatures, either the same general form or mirror images of parts or the whole of them, suggesting that all the organisms are responding to the same specific geophysical changes and that they have freedom only to determine the sign, positive or negative, of their responses to them.

The nature of the biological clock. As more information becomes available, it seems clearer that there are two kinds of rhythms, one resulting from response to the subtle geophysical environment, the other being rhythmic organismic responses to more obvious changes.

The timer for the first of these kinds of rhythms obviously lies outside the organism, and hence these rhythms are termed geophysically dependent. The detailed character of the biological cyclic pattern, however, is just as obviously determined by the nature of the biological response. The widespread general uniformity of the pattern for all organisms suggests that the response is generated at a very fundamental level of cellular organization. It must still depend upon the organism as an organized continuum whose composition and reactivity at any given point in time are related to past and future. The organism is, therefore, not a fully passive system in the generation of its exogenous metabolic periodisms. It may be an effective stimulus-filtering system.

The clock that times the overt, adaptive circadian, tidal, monthly, and annual rhythms is far less evident. Two kinds of hypotheses have been advanced. One, the internal-timer hypothesis, is that living systems can oscillate, fully independent of all environmental periodisms with close to, but not exactly, all major geophysical frequencies. Free-running rhythms are assumed to reflect directly the periods of the somewhat inaccurate internal clocks. Exact organismic solar-day, tidal, monthly, and annual periods are assumed to depend upon continuing corrections of the poorly accurate clocks by the environmental light, temperature, and tidal rhythms. The rates at which the clocks run are believed to be speeded or slowed to a small degree by changed levels of temperature and illumination as well as by certain chemical factors. The clock periods are assumed to be inherited, the approximation to the geophysical periods having resulted from natural selection.

The internal-timer hypothesis

By describing the rhythms as oscillations resembling in properties those of physical systems that become entrained or synchronized to other slightly different imposed

frequencies and treating them mathematically as such, surprisingly satisfactory descriptions of the observed dynamics of biological rhythms and phase setting of the rhythms can be obtained, useful predictions made, and new experiments suggested. Despite intensive research, progress has not yet been made in elucidating an internal oscillator system with the clock properties. The problem is compounded by the inability to differentiate between a biochemical or biophysical reaction that is in reality a clock element and one that is merely timed by the clock.

The external-timer hypothesis

Another hypothesis is that living things have evolved the capacity to use the subtle rhythmic input of geophysical periods to time their own processes to the most adaptive rhythmic patterns. In terms of this hypothesis the clocks are not inaccurate but quite precise (24-hour, 24.8-hour, 29.5-day, and 365 1/4-day clocks, for example). As long as the organism remains in its natural rhythmic environment, the clocks and the environmental rhythms of light, temperature, and tides have the same frequencies. When, however, the organism is placed under constant conditions, its circadian rhythm gradually drifts off phase. There will be phase advances alternating with phase delays in a cyclic manner as the constant light and temperature react at different points along the response curve. The recurring patterns will gradually shift to earlier hours if the advancing response is stronger than the delaying one or shift toward later hours if the delaying one is the stronger. This corresponds well with the observation that circadian periods deviate slightly in practice from 24 hours.

The self-setting to specific clock hour of solar day or lunar day or time of month or of year sometimes observed in constant conditions, as well as the very existence of the geophysically dependent periodisms, would find consistent explanation. The accuracy of the clocks during rapid east–west translocations would require either influence of a universal time-varying geophysical factor, phase-synchronized at all longitudes, as, for example, ionospheric charge, or a capacity to rephase its own rhythmic pattern in relation to a rapidly shifted exogenous local-time cycle effected by the metabolic inertia inherent in a continuing functional system.

In brief, some of the properties of biological periodicities definitely require for their explanation that timing information continuously flows inward from the environment, even in what have commonly been deemed constant conditions. Many other phenomena are equally compatible with either hypothesis, internal or external timing. Possibly both kinds of clocks are utilized by organisms. An environmentally dependent clock element may be responsible for the remarkable resistance to timing interference by temperature and chemical agents and for the long-term, relatively precise persistence of daily rhythms and particularly for the stability of the longer period monthly and annual ones. An environmentally independent clock element, on the other hand, may be essential for bridging shorter intervals of time within cycles but with perhaps a precision by itself inadequate to account for the observed regularity and dependability of clock-timed rhythms.

FACTORS AFFECTING BIOLOGICAL PERIODICITIES

The rhythmic patterns of variation of organisms differ from one species to another and also from one individual to another, even in response to essentially the same external conditions. The nervous and endocrine systems in interaction with the parts of the body that are coordinated and integrated by them contribute to these patterns. Under ordinary circumstances, activities comprising the patterns are arranged to meet an organism's demands anticipated over the period. Pathological or otherwise induced alterations in the organism alter the cycle forms. The periods, following those of the physical environment, usually remain unchanged. Reported, however, have been instances in blinded mice of circadian periods becoming free-running ones differing from 24 hours, resembling the state of affairs commonly seen when comparable mice are held in continuous darkness. Light, operating through the eyes, would appear in such instances to be the essential phase-setting agent.

The role played by the animals' total functional complex in determining the cycle form has been illustrated in respiratory patterns of unhatched chick embryos and in the activity of human infants. In conditions of constant temperature and light, chick embryos up to six days old show a daily activity variation with several peaks each day. By the seventh day, when they have just completed the development of a functioning sensory-neuro-motor system, the daily variation changes to one with a broad daytime higher activity and lower nighttime one. Well known, too, is the change in sleep-activity rhythms of infants. Up to about 15 weeks there are several cycles a day, but afterward the cycle simplifies to essentially a single longer period of wakefulness by day and sleep by night. It might be argued that the infant simply adopts the social pattern of the household, but in the chick the change in pattern with time seems clearly genetically determined.

The sleep-activity cycle of infants

It was mentioned earlier that evolution, involving natural selection, has suited species to fit each niche at a particular time within a cycle. In the continuing course of events a species is usually rigorously held to its rhythmic schedule by the competition of other species present at other times. It is an interesting observation, however, that some large day-active game animals in Africa, whose populations had been decimated by hunters, altered their natural patterns to become night active but resumed diurnal habits when their numbers again had increased. It is likely that comparable adaptive phenomena may be more widespread in nature than has been reported. (F.A.B.)

PHOTOPERIODISM

Photoperiodism is the behavioral response of an organism to changes of duration in daily, seasonal, or yearly periods of light and darkness. It was first reported in 1920 when the flowering of certain plants was demonstrated to be controlled by the daily duration of light.

The photoperiodic response of plants actually depends on the duration of the darkness, and not on light, as the earlier experiments seemed to indicate.

The most conspicuous activities of animals closely correlated with certain seasons of the year, and hence with changes in day length, are bird migration, reproduction, and changes in coat and plumage. Each of these occurs with marked regularity at a particular time each year. By retaining animals in captivity and subjecting them to artificial increases and decreases in day length, bird migration, reproduction, and other activities have been induced out of season. It has been concluded, therefore, that in nature the changing daily periods of light and darkness determine to a great extent when many seasonal activities occur.

Before their northward migration to breeding grounds in the spring, birds undergo a marked change in their physiological state. The change is shown by increased activity of the reproductive organs and deposition of large amounts of fat. When these same changes were induced experimentally in winter by exposing captive birds to artificial increases in day length, the birds migrated northward in winter when released. By employing different schedules of day length at various seasons of the year, it has been demonstrated that the rate of change in physiological state depends on the daily amount of illumination. As little as nine hours of light per day is stimulating. The response is rapid under long days and slow under short days. A gradual increase in the amount of light per day is not necessary. Constant day lengths are effective, as are decreasing day lengths, provided that the daily amount of light is stimulating. Therefore, migratory birds that winter on or near the Equator, where the day lengths are constant, or in the Southern Hemisphere, where the day lengths decrease after December 21, could have the onset of their spring migration regulated by day length. Prerequisite to the response to light that occurs in winter and spring is exposure to short days or long nights in autumn. This phase of the annual cycle is called the preparatory phase, since failure to complete it precludes the subsequent response. Under experimental conditions completion occurs uniformly and more rapidly with longer nights. Day lengths in autumn wind the "clock" within the bird, so to speak, while the day lengths of winter and spring make it run and determine how fast it runs.

Regulation of reproduction

In most birds, mammals, and other vertebrates, breeding is seasonal. By experimental manipulation of day length, reproductive activity has been induced out of season. For the species that normally breed beginning in autumn (*e.g.*, brook trout, deer, sheep, and goats), decreasing daylight or short days following long days will be effective. For species that breed in the spring (*e.g.*, starling, junco, ferret, and mink), increasing daylight or long days following short days will be effective. Changes in breeding season occur in some birds and mammals when they are transported from the Southern to the Northern Hemisphere or from the region of the Equator to temperate latitudes. Although day lengths are constant on the Equator and change but little in the tropics, there is some evidence that day length does regulate the breeding and plumage cycles of tropical birds. Experimental modification of the breeding cycle of invertebrate animals (*e.g.*, snails, crustaceans, and insects) by means of day length has also been demonstrated.

Attempts to discover how the length of daylight affects the breeding cycle have shown that the light stimulates the eye or part of the brain near the eye. The light stimulates the release of special hormones from the brain, which eventually reach the pituitary gland (the master endocrine gland located at the base of the brain). The pituitary, in turn, secretes the hormones that, among other functions, control the growth of the reproductive organs. There is good evidence that the length of the dark period and the ratio of light to darkness are important factors in determining the reproductive response. The effects of day length on the migratory response seem to apply as well to the reproductive response in some birds. Response to light can be modified by temperature, nutrition, and inherent internal factors. Important among the last is a daily rhythm of about 24 hours' duration, which makes it appear as though the organism were keeping track of each day. The response to light–dark cycles seems to depend in part on the time that the light and dark periods occur within these 24-hour periods. A few species of birds and mammals studied did not show any influence of light on the reproductive cycle.

The most conspicuous seasonal change in coat and plumage occurs in those mammals and birds that are white in autumn and winter and brown in spring and summer (*e.g.*, varying hare, ermine, and ptarmigan). By experimental manipulation of light, either of these colours can be produced out of season. Molt into white plumage or coat can be induced by short days. Temperature is not a factor. Day length also influences the time of annual molt in birds, growth of wool in sheep, and production of the winter coat in mink and ferret.

A number of less obvious seasonal activities also occur in response to day lengths. In some aphids, the production of winged individuals (as opposed to wingless individuals) normally occurs in autumn and is dependent on a period of darkness of 12 to 14 hours. In certain other insects the occurrence of sexually reproducing individuals, which occurs normally in autumn, can be produced out of season by shortening the days.

Practical applications of knowledge of photoperiodism are common in poultry management, for the length of daylight affects laying, sperm production, and body weight of the fowl. (A.Wo./Ed.)

BASIC BEHAVIORAL ACTIVITIES OF INDIVIDUALS

Feeding behaviour

The living cell depends on a virtually uninterrupted supply of materials for its metabolism. In multicellular animals the body fluids surrounding each cell are the immediate source of nutrients. The contents of these fluids are kept at a relatively constant level in spite of tolls taken by the cells, primarily by mobilization of nutrients stored in the body; in vertebrates, for example, glucose is stored in the liver, fats in the fat tissues, calcium in the bones. These stores, however, will become exhausted sooner or later, unless the animal takes up nutrients from outside. Movements performed for this purpose are termed feeding behaviour.

NUTRITIONAL REQUIREMENTS OF HIGHER ANIMALS

Catabolism and anabolism

Cells use nutrients as fuel for energy production (catabolism) and as material for processes of maintenance and growth (anabolism). Multicellular animals derive energy solely from the breakdown of complex organic molecules, mainly carbohydrates and fats. Because the fuel for the maintenance of animal life comes only from other living organisms or their remains, animals are known as heterotrophic organisms. All animal life depends ultimately on the existence of organisms (largely green plants) that can use inorganic sources of energy, of which solar radiation is by far the most important; some microorganisms, however, obtain energy from oxidation of simple inorganic compounds.

For anabolic purposes, food must provide adequate amounts of all chemical elements needed by the cells. Of the approximately 35 elements now known to occur in animal cells, four (oxygen, carbon, hydrogen, and nitrogen) make up about 95 percent of the cell weight; another nine (calcium, phosphorus, chlorine, sulfur, potassium, sodium, magnesium, iodine, and iron) contribute about 4 percent. All of these elements have indispensable functions. The remaining 20-odd, together constituting less than 1 percent of cell weight, are called trace elements, because they occur in minute quantities. Although some of them may become incorporated into cells by accident, many fulfill vital functions (see NUTRITION).

It is important to note that animal cells cannot synthesize from simple compounds certain necessary complex molecules. Instead, certain large organic molecules must serve as building blocks; such so-called essential dietary components include the vitamins, some amino acids, and certain fatty substances. In general, higher animals appear to have more restricted synthetic powers than lower ones and to require a correspondingly greater number of essential foodstuffs. Microorganisms in the intestines of vertebrates may synthesize materials essential for the host, so that the food of the latter need not contain these substances.

TYPES OF FOOD PROCUREMENT

Because much of animal evolution, as regards behaviour, as well as anatomy and physiology, involves adaptation for the procurement of food, the extent of the meaning of the term feeding behaviour is not clear. Migratory habits of birds, for instance, no doubt evolved in part as a result of seasonal food shortages; individual birds now, however, start migration before food becomes scarce. Migration, therefore, important though it may be in the feeding ecology of a species, is not considered in this section, which concentrates on food-directed activities that are enhanced by a need for nutrients in the body of an individual. For similar reasons, activities such as host finding and acceptance by internal parasites for themselves or their offspring also are excluded.

A classification of mechanisms of food procurement

Even with these restrictions, the diversity of feeding patterns is bewildering. A useful classification has been put forward by British zoologists Sir Maurice Yonge and J.A.C. Nicol, based on the structural mechanisms utilized, although, as Nicol observed, "many animals make use of a variety of feeding mechanisms, conjointly, or separately as occasion demands":

I. Mechanisms for dealing with small particles.

A. Pseudopodial (*e.g.*, many protozoans). Pseudopods consist of fingerlike projections of the cell membrane and its contents (cytoplasm) that surround and engulf food.

B. Ciliary (*e.g.*, sponges, bivalve mollusks). Cilia are minute hairlike projections of cell membranes that, by concerted beating in wave rhythm, set up water currents or physically move food particles.

C. Tentacular (*e.g.*, certain sea cucumbers). Tentacles are slender, flexible organs on the head. They may function in sensory perception and in actually securing food.

D. Mucoid (*e.g.*, many snails, such as *Vermetus*). In this case, the food particles become attached to a sticky mucous sheet secreted by special cells.

E. Muscular (*e.g.*, certain coelenterates). In the jellyfish *Rhizostoma,* pulsations of the bell-shaped body draw water and food in through perforations in the arms, then expel the water after the food is removed.

F. Setous (*e.g.*, many small crustaceans, such as copepods). Setae are bristlelike projections of the cuticle and are found on the appendages of many invertebrates.

II. Mechanisms for dealing with large particles or masses.

A. For swallowing inactive food, such as bottom deposits (*e.g.*, many polychaete worms, some fishes).

B. For scraping and boring (*e.g.*, some gastropod and bivalve mollusks).

C. For seizing prey.

1. For seizing and swallowing only (*e.g., Hydra,* many polychaete worms, many lower vertebrates).

2. For seizing and masticating (*e.g.*, Crustacea, mammals).

3. For seizing followed by external digestion (*e.g.*, some starfishes, spiders). (In such cases, the secretory and absorptive surfaces of the digestive system may be applied to the food by everting (*i.e.*, turning inside out) the stomach, a method employed by starfish. Alternatively, digestive enzymes may be injected into the prey, liquefying the tissues, which may then be ingested by the predator. This mechanism is found in spiders.)

II. Mechanisms for taking in fluid or soft tissues.

A. For piercing and sucking (*e.g.*, leeches, mosquitoes).

B. For sucking only (*e.g.*, many flies, butterflies).

For absorption through surface of body (*e.g.*, various invertebrates feeding on decaying organic matter, internal parasites such as tapeworms, which lack a digestive tract).

A different classification, often used, rests on the nature of the behaviour for procuring food:

A. Filter feeders strain food from the surrounding medium more or less indiscriminately.

B. Selective feeders analyze the environment with their sense organs before aiming feeding responses at chosen items.

Some feeding patterns, however, cannot be easily fitted into either of these classes alone; spiders, for example, sieve prey from the air with webs but perform directed responses to the insects trapped. Class I of the Yonge–Nicol system comprises mainly filter feeders; most members of classes II, IIIA, and IIIB are selective feeders. Selective feeding requires good sensory and nervous equipment and, in most cases, considerable mobility. It is therefore found mainly among higher animals. Yet the primitive sea anemones are selective feeders in that, capable of paralyzing relatively large prey with their stinging cells, they do not discharge them until informed by chemical and tactile senses that prey is present. At the other extreme, whalebone whales are filter feeders, even though they are highly evolved mammals. Swimming at the surface with mouth open, they filter off large plankton (krill) using several hundred horny plates with hairlike fringes hanging down from the roof of the mouth; availability of a rich food source has caused the evolution of their feeding patterns to diverge widely from that of most other mammals.

Evolutionary forces and limits on feeding pattern

In all cases, the feeding patterns adopted by species are the result of evolutionary interplay between (1) structural properties inherent in their phylogenetic line and (2) the ecological situations to which they have been exposed. These interactions are too complex to make generalizations profitable. The best approach is to study each species as a separate case in the light of its entire biology. A few examples are given below.

Filter feeders occur among sponges, coelenterates, polychaete worms, echinoderms, brachiopods, mollusks, arthropods, protochordates, fish, birds, and several other groups. As might be expected, filtering devices are diverse.

In the oyster, constantly lashing cilia drive a water current—up to 34 litres (about 36 quarts) per hour—through the openings of perforated gill plates. Particles only two microns (0.002 millimetres) in size are wrapped in mucus and transported by other cilia to special food grooves, along which they pass to the mouth by the action of yet further cilia; particles that are too large, too heavy, or capable of producing irritation are sorted out and rejected by various mechanical means.

The polychaete worm *Chaetopterus* uses a bag of mucus, secreted by special body appendages, to strain the water it pumps through its burrow. The mesh openings of the bag, about 40 angstroms (4×10^{-7} millimetres) wide, can even trap single molecules of large proteins. Every 20 minutes the food-laden bag is taken to the mouth, consumed, and replaced by a new one.

The sessile marine snail *Vermetus gigas* secretes mucus strings up to 30 centimetres (12 inches) long that extend away from the shell and entangle fine plankton. At intervals the strings are drawn back toward the mouth and swallowed.

Selective feeders, found among major animal groups, including coelenterates, annelids, echinoderms, mollusks, arthropods, and vertebrates, show even greater diversity of feeding patterns than do filter feeders. One striking point is that different groups deal in different ways with the same food in accordance with special capacities. Animals feeding on bivalve mollusks provide one example. The starfish *Asterias* forces the valves apart by the relentless pull of its sucker tube feet and then everts its stomach through its mouth to digest the soft tissues inside the shell. The snail *Sycotypus* attacks an oyster by stealth: waiting until the valves open, it thrusts its shell between the valves and pushes its tubular feeding organ, or proboscis, into the soft parts. Another snail, *Natica,* supports the scraping action of a filelike structure called a radula with chemical dissolution by sulfuric acid, which is secreted by a gland on the proboscis, and drills a neat hole in a clam. Another snail, *Fulgur,* cracks a clam shell against its own shell by contracting its columellar muscle. Among birds, the oyster catcher (*Haematopus ostralegus*) adroitly cuts the closing muscles of a cockle with its chisel-shaped bill; herring gulls (*Larus argentatus*) break a shell by dropping it onto a rock. A sea otter (*Enhydra lutris*) cracks a clam on its chest, while floating on its back, by pounding it on a stone held between the forepaws.

Various methods of feeding on bivalve mollusks

A few additional examples further illustrate the wealth of adaptations in selective feeding. The sluggish praying mantis (orthopteran insects of the family Mantidae) stalk insect prey until within reach, then carefully orient themselves and accurately and rapidly extend forelegs adapted for grasping. For detecting prey in murky habitats, bats use an ultrasonic echolocation system; some fish use electric pulses in a somewhat comparable manner. Anglerfish dangle a baitlike appendage of the first dorsal spine (luminous in deep-sea species) to lure the fish on which they feed toward their enormous mouths. Certain labroid fishes, which eat parasites off the bodies of other fish, induce their hosts to submit to treatment by a dancing approach; certain blennies treacherously mimic this behaviour and then rapidly bite the fin of the unsuspecting victim. Shrikes perform special wiping movements to remove the sting from certain prey, even without previous experience of stinging insects. The peregrine falcon (*Falco peregrinus*) dives at birds at speeds above 160 kilometres (100 miles) per hour; and the cheetah, or hunting leopard (*Acinonyx jubatus*), pursues antelope at more than 95 kilometres (60 miles) per hour. More than a thousand cormorants may join in a single fish drive. Instead of hunting, some species rob food collected by members of other species; among these robbers are the marauding skuas and jaegers (*Stercorariidae*) and man-of-war birds (*Fregata*), which force weaker cousins to disgorge already swallowed prey, and various tropical flies that take up positions along the line of march of army ants and rob the passing workers.

The driving force for the evolution of each of these adaptations is the survival value to the species of selecting the food sources for which it can successfully compete. For the same reason, closely related species living in the same area may tend to exploit separate parts of the environment; *e.g.*, woodland titmice (*Parus*) forage in different parts of the trees, and the larvae of different species of the moth genus *Eupithecia* prefer different food plants. The result of such evolution may be that a species becomes specialized to one kind of food, as have many internal parasites and phytophagous (plant-eating) insects. Such food may be exotic, as is that of the larva of a moth (*Galleria*) that feeds on beeswax. In other species, such as the herring gull, on the other hand, each individual exploits a broad range of foods, thereby lessening the risk of starvation, as

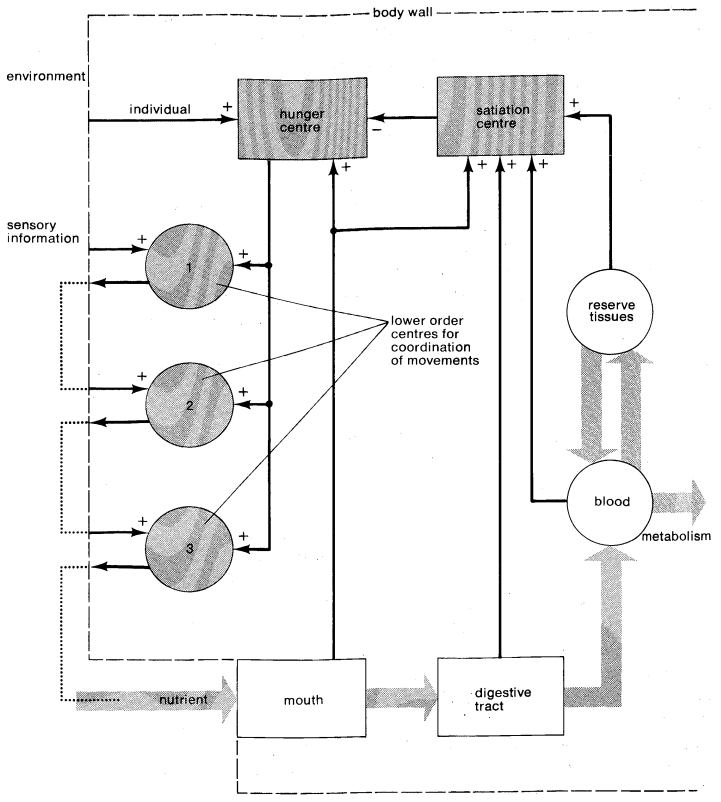


Figure 12: Diagrammatic representation of feeding control in a generalized mammal. Thin black arrows represent information flow; shaded arrows the flow of nutrients. The hunger centre is stimulated by general sensory information (*e.g.*, environment suitable for feeding behaviour). It, in turn, activates lower order brain mechanisms for food searching (1), food getting (2), and eating (3); performance of each of which prepares the animal for the next. Ingested food passes through the various compartments of the body, each of which sends messages to the satiation, or satiety, centre, which inhibits the hunger centre correspondingly. In addition, taste messages from the mouth stimulate the hunger centre.

it is unlikely that all types of food will become exhausted at the same time.

REGULATION OF FOOD INTAKE

Metabolic expenditure cannot exceed food intake for very long if an animal is to survive. One way to equalize the two processes is to decrease metabolism to a level sustainable by maximum intake, which may be limited by the ability to extract food from a meagre habitat. Data for filter feeders suggest that, in certain cases, continuous filtration at maximal rates may be barely sufficient to support normal growth and maintenance. Selective feeders have been found to undergo a more or less drastic reduction of metabolism during temporary starvation. Secondly, the capacity of the digestive system may set a limit on nutrient supply to the body. There is evidence that this is so in the minute filter-feeding crustacean *Daphnia magna*. Such limitations are known to play a role in human feeding behaviour.

In man and many other selective feeders, nevertheless, the capacities of food-gathering and digestive systems exceed all but the most extreme demands of metabolism. To maintain nutritional balance, feeding must then be geared to metabolic rate. Information on the mechanisms and even on the existence of such regulation of intake is scanty, except for mammals and some insects.

Vertebrates. Most information on the control of feeding behaviour in vertebrates has come from studies of mammals, but the general patterns found in mammals appear to be present in fish, amphibians, reptiles, and birds. Food intake requires a well-ordered sequence of searching, food getting, and ingestive activities. Sometimes the behaviour is elaborate. The following elements are distinguished in the various cats: stalking, spying, pouncing, thrusting down with the head, biting the neck, carrying into cover, plucking, and devouring. In grazing animals, the pattern is much simpler. In any case, the movement a feeding animal performs at a given moment depends largely on external stimuli; search and pursuit, for example, are unnecessary when prey is within reach. In this sense, any feeding act is a response to the environment, but it is not a simple "reflex." On repeated presentation of the same food situation, the individual sometimes shows the appropriate response but at other times will fail to do so. These fluctuations in responsiveness are roughly parallel in all elements of feeding behaviour. Responsiveness tends to be higher with increasing lack of food in the body. It appears that responsiveness of the brain mechanisms for feeding is governed by messages reporting the nutritional state of the body. The contents of these messages, in other words, are primary determinants of the level of feeding motivation (for other influences see below *Relation of feeding to other functions*). High and low levels of feeding motivation are the objective counterparts of the everyday concepts of hunger and satiety. Regulation of food intake, then, must hinge on the physiological mechanisms of the feeding motivation.

Elements of feline feeding behaviour

Specific hungers. Lack of any nutrient with a specific anabolic function, such as vitamins or minerals, must be redressed by increased uptake of the particular substance. Little is known thus far of the specific hunger mechanisms that ensure increased uptake, but good evidence exists that a nutrient deficiency causes a specific rise in responsiveness to food containing the substance needed. In the case of thiamine (vitamin B_1), a learning process is involved. The deficient animal tries various kinds of food and concentrates on those that remove the deficiency. Specific appetite for salt in a sodium deficient subject, on the other hand, appears to rest on a genetically determined increase in reaction to the taste of sodium chloride and does not require any learning.

Caloric regulation. Lack of fuel in the body can be corrected by intake of any of a variety of possible substances that provide energy. Most natural food contains a mixture of such substances. Energy deficiencies can be alleviated by increased responsiveness to food in general. Ingested food (*i.e.*, calories) passes from (1) the mouth to (2) the digestive tract to (3) the bloodstream; if not needed at once for catabolic processes, the digested food passes to (4) storage sites, of which the fat tissues are the most important. These four regions are continuously monitored. A considerable amount is known about the monitoring roles of the organs for taste, smell, and touch in the mouth region; in addition, distension receptors in the digestive tract monitor the volume there, and chemoreceptors monitor the nature of the contents. Information concerning the availability of glucose (the most commonly utilized sugar) and possibly other fuels in the blood is recorded by cells located probably both in the brain itself and elsewhere (*e.g.*, in the liver). Finally, circumstantial evidence suggests that the contents of fat tissues are also monitored. All food that passes through the body contributes to each of these four messages in succession, until it is eventually catabolized.

The signals converge on the brain mechanisms for the feeding motivation over nervous and, possibly, humoural (chemical) pathways. Here they have effects of two kinds: (1) if signals from the four regions report increased fuel contents, the feeding motivation is lowered (satiety is raised), and (2) if taste (and perhaps other, *e.g.*, visual) receptors are stimulated by palatable food the feeding motivation is increased. Intake stops when accumulation of signals of the first kind, overriding those of the second kind, causes hunger to drop below a critical level. Feeding is resumed when hunger surpasses this level as a result of fuel depletion by catabolism and emptying of the digestive tract by digestion and absorption. Once started, intake is enhanced by the positive effects of the food stimulus. The net result of this interplay of positive and negative feedbacks from food responses is that caloric intake, observed over a sufficiently long period (at least several days), is equal to energy output over that period, so that body fuel content (body weight in fully grown individuals) remains constant.

The brain mechanisms involved in vertebrate feeding motivation consist of a complex network, not yet well understood, encompassing, among other areas of the brain, the limbic system (the marginal zone of the forebrain) and the hypothalamus. The lateral hypothalamus ("hunger centre") facilitates feeding responses. Electrical or chemical stimulation of this area elicits voracious feeding in satiated subjects, and its destruction causes more or less prolonged noneating (aphagia). If the subject is kept alive by artificial feeding, however, other brain areas may take over and reinstate more or less normal feeding. In contrast, the ventromedial (lower central) nucleus of the hypothalamus appears to be a clearinghouse for satiety signals. Subjects with lesions in this area stop feeding only at an abnormally high level of energy content (obesity) and grossly overeat (hyperphagia) until this level is reached.

Control of feeding behaviour in the blowfly

Invertebrates. One of the few invertebrates in which the physiology of feeding behaviour has been extensively studied is the blowfly *Phormia regina.* Sucking is elicited by food stimuli on taste organs of the tarsi (the terminal sections of the legs) and proboscis. The meal continues until adaptation of these receptors causes their signals to decrease below the threshold of the sucking-response mechanism. This threshold is modulated, in the following manner, by food present in the digestive tract and in body fluids. As long as food is present in the foregut, the threshold is raised by signals from distension receptors in that area. The foregut is kept filled after a meal by release of food from the crop, where food taken up at the meal in excess of the capacity of the gut is temporarily stored. The threshold will remain high, therefore, until the crop is completely voided. The rate of crop emptying is directly related to the nutrient concentration of body fluids. The latter depends on the balance between absorption from the gut and uptake by the metabolizing tissues. The harder the fly works, therefore, the sooner sucking will be resumed, with the result that food intake is kept equal to caloric expenditure through appropriate spacing of meals.

SELECTION OF FOOD ITEMS

Most natural habitats offer a diversity of food objects, and most selective feeders are more or less euryphagic—*i.e.,* they ingest a variety of different foods; strict monophagy is less common. On the other hand, no euryphagic species includes in its diet all potential food objects present in the habitat, nor are those that it does eat taken in proportion to the amounts in which they are available. On what grounds, then, are diets selected?

Vertebrates. A plant species constituting only a fraction of 1 percent of a pasture may make up the greater proportion of the diet of a sheep. Insectivorous birds also take a highly biassed selection from the insect menu offered by the habitat. Although the relative abundance of different kinds of food is reflected in diets to some extent, this does not usually go so far that a single kind of food, however attractive and abundant, will become the sole constituent. Most vertebrates appear to take a varied diet whenever possible.

Responses to encountered food. Diet selection in adult vertebrates proves to be largely the result of individual learning processes that guide the genetically determined response potentialities of the newborn individual into certain definite channels.

Instinctive feeding in young animals

Innate responsiveness appears to be broad in species that forage for themselves from birth and thus must deal with many different food situations. The pecking of newly hatched chicks of domestic fowl at all kinds of small objects, edible or not, is an example. Yet these chicks have certain innate preferences for colour and other features. Such preferences may foreshadow the composition of adult diets. In newly hatched snakes, for instance, feeding responses are more easily elicited with extracts of the natural food of adults of the same species than with preparations of food of closely related species. In contrast, colour preferences of ducklings of different species are similar, although the adult diets differ.

Innate responsiveness may be narrow, however, in young vertebrates for whom the parent is the only source of food. Herring-gull chicks beg for food in response to a few "sign stimuli" provided only by the parent's head among all objects in the natural habitat. Sucking behaviour of newborn mammals is a somewhat comparable example. In such cases, responsiveness must be profoundly reorganized when the individual forages on its own.

Responsiveness is channelled into the adult pattern through experience of taste, nutritional value, and possible noxious properties of various objects. In this way the individual is able to attach a definite palatability rating to each type of food regularly encountered and to associate this with visual or other characters by which it recognizes objects from a distance. As demonstrated in experiments, insectivorous birds may discriminate precisely among as many as 40 different prey species in this manner.

In addition to palatability, detectability of food objects is a factor in diet selection. This has been studied in detail in visually foraging vertebrates. Detectability of an object depends on its degree of contrast with the background as to colour, shape, and movement. The individual predator can learn to detect prey that it finds only with difficulty at first; such "searching image formation" occurs only if the prey is palatable and encountered often.

Finally, responses to encountered prey also depend on (1) the hunger level of the individual and (2) its experience regarding the general food situation. Hungrier predators have lower palatability requirements and may take greater risks to secure prey. At one and the same hunger level, a prey of slight palatability may be rejected if the predator "knows" that further search will probably bring better food but accepted if it "knows" that nothing tastier is available. As a result of these two influences, animals concentrate during scarcity on food scorned in times of plenty.

Determination of search area by a predator

Food searching and diet. The general type of food taken is often determined by the innate search method of the animal and the section of the whole habitat being exploited. A fish-eating bird, such as the osprey (*Pandion haliaetus*), which secures prey by diving into water (but not swimming), is limited in its diet to fish species that are active near the surface. The much discussed question of whether food searching is random is relevant here, for certain kinds of nonrandomness can influence diets. No simple answer can be given. Search must be random in the sense that oriented reactions to food objects can be made only after detection; at the same time, however, the search may be systematic in that (1) places not recently traversed are favoured over those just unsuccessfully explored, and (2) the locality where a prey has just been caught or seen may be searched with special attention. Further, (3) it is most common for individuals to restrict their foraging to parts of the home range where ample food has been previously found, although exploration of other parts is interspersed and may change the destination of further trips if successful. In all, food searching appears to have sufficient nonrandomness to influence diets provided that different kinds of food concentrate in different parts of the home range, as is often the case.

To sum up, vertebrate diet selection is largely molded by learning processes. Insofar as their course depends on chance experiences of individuals, differences in diet may develop even among members of one population of a species. On the whole, however, patterns of food selection are typical of the species, as all its members have similar genetic makeup and live in broadly similar ecological situations.

Invertebrates. Learning processes appear to play a relatively small role in food selection by invertebrates. Diets are largely, though not entirely, determined by genetically fixed preferences. Intensive studies have been made of host-plant selection by phytophagous insects. Here, as in host selection by animal parasites, the question is one of the choice of a place to live rather than of food alone, and the selection criteria may be largely a matter of compromise between nutritional requirements and other ecological functions. Leaving aside these complications, the factors leading to selection of a particular plant as food are predominantly chemical, although other properties, such as structure, also play a role. The chemicals involved in part are the nutrients themselves, but often the feeding responses are largely elicited by token substances that are

not nutritionally essential but are characteristic of the species or family of plants that provide the natural hosts for the insect concerned.

SPECIALIZED ASPECTS OF FEEDING BEHAVIOUR

Relation of feeding to other functions. In principle, feeding must proceed throughout life at a pace equal to that of metabolism, but in many cases intake does not closely follow expenditure. It is permissible for intake to lag when there are reserves in the body. In some cases it is clear that large reserves are present in anticipation of increased metabolic demands or predictable food shortage—*e.g.*, hibernating mammals store large amounts of tissue fat before the onset of dormancy; migrating birds do the same before departure. Insect larvae store nutrients to last them through the pupal stage. Adults of many insects, such as mayflies, do not eat at all and have reduced mouthparts. Other species, such as the hamster, solve similar problems by laying in extracorporeal hoards of food.

Reserves and stores of food

Discrepancies between intake and expenditure, whether large or small, amount to distortion of the basic pattern of caloric regulation for the benefit of other functions.

Like most biological processes, feeding has a diurnal periodicity; *i.e.*, depending on the species, the active period may fall in daylight or during the night. Only in filter feeders is the activity often continuous.

Priority claims of other functions may lead to suppression of feeding even in hungry animals. Thirsty mammals or birds eat much less than normal because food intake would aggravate water shortage in the body in various ways. The same is true of mammals in a hot environment—*i.e.*, food intake increases heat production in the body and would thus intensify the heat stress—and of female mammals during estrus (periods of fertility). In all these cases, more or less marked loss of body weight results.

Social facilitation is a further cause of discrepancies as here considered. Individuals often start feeding when they observe other members of the same (or other) species doing so. Both timing of feeding and choice of food are affected in this way. Unfamiliar food is accepted more readily by individuals observing others eating it. Such phenomena have been noted in mammals, birds, and fish.

Food-directed activities in social situations. A further complication is that food-directed activities may be performed for the benefit of other individuals (see also below *Behaviour of animals in groups*). This may serve their nutrition or some other function. Marked weight loss may occur in songbirds as they feed most of the prey to their nestling young. Courtship feeding in many birds (and insects), in which the male gives food to the female, strengthens the pair bond rather than having a role in nutrition.

Remarkably intricate is the behaviour by which individuals of social-insect species—honeybees, for example—ensure nutrition of the colony. Tropical honey ants store nectar collected by the workers of the colony in the crops (stomachs) of certain workers that remain inside the nest and become so gorged that they are hardly more than storage bins. They disgorge droplets upon solicitation by other ants in the nest. "Dairying" ants keep aphids as suppliers of honeydew, a sugar- and protein-rich secretion. They milk the aphids by gently stroking them and, in return, protect them against enemies. The aphids may even be carried to the nest at the approach of winter and returned to a plant the following spring.

Fungi cultivation

A number of ants and termites cultivate fungi for food. Workers of tropical leaf-cutting ants carry pieces cut off the green leaves of trees to the nest, where other workers use them for making a bed on which the fungi grow. When a queen sets out to start a new nest, she carries a pellet of mycelium (the "root" system of the fungus) in a special pocket on her head during her nuptial flight and subsequent burrowing. After depositing it in the new nest, she manures it with a special secretion until the first workers start bringing in leaf fragments.

The motivational background of behaviour as discussed above has not yet been sufficiently analyzed. Much remains to be done in the more intricate—and even in the more straightforward—cases before satisfactory insight into the functions and causes of the behaviour of animals toward their food is achieved. (L. de R.)

Locomotion

To locomote, all animals require both propulsive and control mechanisms. The diverse propulsive mechanisms of animals involve a contractile structure—muscle in most cases—to generate a propulsive force. The quantity, quality, and position of contractions are initiated and coordinated by the nervous system: through this coordination, rhythmic movements of the appendages or body produce locomotion.

Animals successfully occupy a majority of the vast number of different physical environments (ecological niches) on Earth; in a discussion of locomotion, however, these environments can be divided into four types: aerial (including arboreal), aquatic, fossorial (underground), and terrestrial. The physical restraints to movement—gravity and drag—are the same in each environment: they differ only in degree. Gravity is here considered as the weight and inertia (resistance to motion) of a body, drag as any force reducing movement. Although these are not the definitions of a physicist, they are adequate for a general understanding of the forces that impede animal locomotion.

To counteract the force of gravity, which is particularly important in aerial, fossorial, and terrestrial locomotion, all animals that live in these three environments have evolved skeletal systems to support their body and to prevent the body from collapsing upon itself. The skeletal system may be internal or external, and it may act either as a rigid framework or as a flexible hydraulic (fluid) support.

To initiate movement, a sufficient amount of muscular work must be performed by aerial, fossorial, and terrestrial animals to overcome inertia. Aquatic animals must also overcome inertia; the buoyancy of water, however, reduces the influence of gravity on movement. Actually, because many aquatic animals are weightless—*i.e.*, they possess neutral buoyancy by displacing a volume of water that is equal in weight to their dry weight—little muscular work is needed to overcome inertia. But not all aquatic animals are weightless. Those with negative buoyancy sink as a result of their weight; hence, the greater their weight, the more muscular energy they must expend to remain at a given level. Conversely, an animal with positive buoyancy floats to and rests on the surface and must expend muscular energy to remain submerged.

Buoyancy

In water, the primary force that retards or resists forward movement is drag, the amount of which depends upon the animal's shape and how that shape cleaves the water. Drag results mainly from the friction of the water as it flows over the surface of the animal and the adherence of the water to the animal's surface (*i.e.*, the viscosity of the water). Because of the water's viscosity, its flow tends to be lamellar; *i.e.*, different layers of the water flow at different speeds, with the slowest layer of flow being the one adjacent to the body surface. As the flow speed increases, the lamellar pattern is lost, and turbulence develops, thereby increasing the drag.

Another component of drag is the retardation of forward movement by the backward pull of the eddies of water behind the tail of the animal. As they flow off an animal, the layers of water from each side meet and blend. If the animal is streamlined (*e.g.*, has a fusiform shape), the turbulence is low; if, however, the water layers from the sides meet abruptly and with different speeds, the turbulence is high, causing a strong backward pull, or drag, on the animal.

Aerial locomotion also encounters resistance from drag, but, because the viscosity and density of air are much less than those of water, drag is also less. The lamellar flow of air across the wing surfaces is, however, extremely important. The upward force of flight, or lift, results from air flowing faster across the upper surface than across the lower surface of the wing. Because this differential in flow produces a lower air pressure on the upper surface, the animal rises. Lift is also produced by the flow of water across surfaces, but aquatic animals use the lift as a steering aid rather than as a source of propulsion.

Lift and drag

Drag is generally considered a negligible influence in terrestrial locomotion; and, in fossorial locomotion, the friction and compactness (friability) of soils are the two major restraints. If the soil is extremely friable, as is sand, some animals can "swim" through it. Such fossorial locomotion, however, is quite rare; most fossorial animals must laboriously tunnel through the soil and thereafter depend upon the tunnels for active locomotion.

Movement in animals is achieved by two types of locomotion, axial and appendicular. In axial locomotion, which includes the hydraulic ramjet method of ejecting water (*e.g.,* squid), production of a body wave (eel), or the contract–anchor–extend method (leech), the body shape is modified, and the interaction of the entire body with the surrounding environment provides the propulsive force. In appendicular locomotion, on the other hand, special body appendages interact with the environment to produce the propulsive force.

There are also many species of animals that depend upon their environment for transportation, a type of mobility that is called passive locomotion. Some jellyfish, for example, develop structures called floats that extend above the water's surface and act as sails. A few spiders have developed an elaborate means of kiting; when a strand of their web silk reaches a certain length after being extended into the air, the wind resistance of the strand is sufficient to lift and carry it away with the attached spider. In one fish, the remora, the dorsal (top) fin has moved to the top of the head and become modified into a sucker; by attaching itself to a larger fish, the remora is able to ride to its next meal.

AQUATIC LOCOMOTION

Microorganisms. Most motile protozoans, which are strictly aquatic animals, move by locomotion involving one of three types of appendages: flagella, cilia, or pseudopodia. Cilia and flagella are indistinguishable in that both are flexible filamentous structures containing two central fibrils (very small fibres) surrounded by a ring of nine double fibrils. The peripheral fibrils seem to be the contractile units and the central ones, neuromotor (nervelike) units. Generally, cilia are short and flagella long, although the size ranges of each overlap.

Flagellar locomotion. Most flagellate protozoans possess either one or two flagella extending from the anterior (front) end of the body. Some protozoans, however, have several flagella that may be scattered over the entire body; in such cases, the flagella usually are fused into distinctly separate clusters. Flagellar movement, or locomotion, occurs as either planar waves, oarlike beating, or three-dimensional waves. All three of these forms of flagellar locomotion consist of contraction waves that pass either from the base to the tip of the flagellum or in the reverse direction to produce forward or backward movement. The planar waves, which occur along a single plane and are similar to a sinusoid (S-shaped) wave form, tend to be asymmetrical; there is a gradual increase in amplitude (peak of the wave) as the wave passes to the tip of the flagellum. In planar locomotion the motion of the flagella is equivalent to that of the body of an eel as it swims. Although symmetrical planar waves have been observed, they apparently are abnormal, because the locomotion they produce is erratic. Planar waves cause the protozoan to rotate on its longitudinal axis, the path of movement tends to be helical (a spiral), and the direction of movement is opposite the propagation direction of the wave.

Types of flagellar movement

In oarlike flagellar movements, which are also planar, the waves tend to be highly asymmetrical, of greater side to side swing, and the protozoan usually rotates and moves with the flagellum at the forward end. In the three-dimensional wave form of flagellar movement, the motion of the flagella is similar to that of an airplane propeller; *i.e.,* the flagella lash from side to side. The flagellum rotates in a conical configuration, the apex (tip) of which centres on the point at which the flagellum is attached to the body. Simultaneous with the conical rotation, asymmetrical sinusoidal waves pass from the base to the end of the flagellum. As a result of the flagellar rotation and its changing angle of contact, water is forced backward over the protozoan, which also tends to rotate, and the organism moves forward in the direction of the flagellum.

Ciliary locomotion. Cilia operate like flexible oars; they have a unilateral (one-sided) beat lying in a single plane. As a cilium moves backward, it is relatively rigid; upon recovery, however, the cilium becomes flexible, and its tip appears to be dragged forward along the body. Because the cilia either completely cover, as in ciliate protozoans, or are arranged in bands or clumps, the movement of each cilium must be closely coordinated with the movements of all other cilia. This coordination is achieved by metachronal rhythm, in which a wave of simultaneously beating groups of cilia moves from the anterior to the posterior end of the organism. In addition to avoiding interference between adjacent cilia, the metachronal wave also produces continuous forward locomotion because there are always groups of cilia beating backward. Moreover, because the plane of the ciliary beat is diagonal to the longitudinal axis of the body, ciliate organisms rotate during locomotion.

Metachronal rhythm

Pseudopodial locomotion. Although ciliar and flagellar locomotion are clearly forms of appendicular locomotion, pseudopodial locomotion (Figure 13) can be classed as either axial or appendicular, depending upon the definition of the pseudopodium. Outwardly, pseudopodial locomotion appears to be the extension of a part of the body that anchors itself and then pulls the remainder of the body forward. Internally, however, the movement is quite different. The amoeba, a protozoan, may be taken as an example. Its cytoplasm (the living substance surrounding the nucleus) is divided into two parts: a peripheral layer, or ectoplasm, of gel (a semisolid, jellylike substance) enclosing an inner mass, or endoplasm, of sol (a fluid containing suspended particles; *i.e.,* a colloid). As a pseudopodium, part of the ectoplasmic gel is converted to sol, whereupon endoplasm begins flowing toward this area, the cell wall expands, and the pseudopodium is extended forward. When the endoplasm, which continues to flow into the pseudopodium, reaches the tip, it extends laterally and is transformed to a gel. Basically, the movement is one of extending an appendage and then emptying the body into the appendage, thereby converting the latter into the former. Although the flow of the cytoplasm is produced by the same proteins involved in the mechanism of muscle contraction, the actual molecular basis of the mechanism is not yet known. Even the mechanics of pseudopodial formation are not completely understood.

From T.L. Jahn, *How to Know the Protozoa*

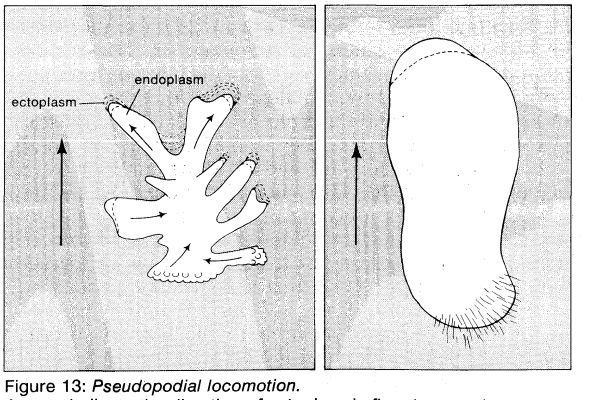


Figure 13: *Pseudopodial locomotion.*
Arrows indicate the direction of cytoplasmic flow (see text).

Undulating and gliding locomotion. Two other types of locomotion are observed occasionally in protozoans. Some protozoans, usually flagellates, have along their bodies a longitudinal membrane that undulates, thereby producing a slow forward locomotion. A gliding locomotion is commonly seen in some sporozoans (parasitic protozoans), in which the organism glides forward with no change in form and no apparent contractions of the body. Initially, the movement was thought to be produced by ejecting mucus, a slimy secretion; small contractile fibrils have been found that produce minute contraction waves that move the animal forward.

Invertebrates. As in the protozoans, aquatic locomo-

tion in invertebrates (animals without backbones) consists of both swimming and bottom movements. In swimming, the propulsive force is derived entirely from the interaction between the organism and the water; in bottom movements, the bottom surface provides the interacting surface. Whereas some bottom movements are identical with terrestrial locomotor patterns, others can occur effectively only in the water, where buoyancy is necessary to reduce body weight.

Bottom locomotion. Small flatworms (Platyhelminthes) and some of the smaller molluskan species move along the bottom by ciliary activity. On their ventral (bottom) surface, a dense coat of cilia extends from head to tail. The direction of the ciliary beat is tailward, causing the animal to glide slowly forward. Generally, all animals that move by this type of ciliary activity secrete a copious stream of mucus over which the animal glides. The mucus not only attaches the animal to the surface but also raises its body so that the cilia can beat. Because ciliary forces are too weak for the movement of large flatworms, they must use muscular contraction for their propulsive force.

Aquatic invertebrates possess several other types of bottom locomotion. In species with well-developed legs, such as crabs and lobsters, bottom walking is common. Whereas the gaits in such cases are identical to those used on land, they tend to be slightly faster in water, because the buoyancy increases the animal's stability. (Walking gaits are described below; see *Terrestrial locomotion.*)

Bottom creeping

Another form of bottom locomotion is bottom creeping, which employs the contract–anchor–extend method of movement. Bottom creeping is best developed in leeches, which have two suckers, one at the anterior end and one at the posterior end. After the posterior sucker anchors the animal, it stretches its body forward and attaches the anterior one. It then releases the posterior sucker and contracts its body toward the anterior end. For effective contract–anchor–extend locomotion, the body musculature must consist of both circular and longitudinal muscles: the contraction of the circular muscles extends or elongates the body; the contraction of the longitudinal muscles flexes and shortens the body. Moreover, the skeleton should be hydrostatic; that is, a fluid skeleton that changes shape but not volume, thereby providing a firm but flexible base.

In pedal locomotion, which is a slow, continuous gliding that is superficially indistinguishable from ciliary locomotion, propulsion along the bottom is generated by the passage of contraction waves through the ventral musculature, which is in contact with the bottom surface. The pedal contraction waves are either direct (in the same direction as the movement) or retrograde (in the direction opposite to the movement). The direct waves produce locomotion in a manner analogous to that in which a caterpillar walks. When a direct wave reaches a muscle, the muscle contracts and lifts a small part of the body; the body is carried forward and set down anterior to its original position as the wave passes. With direct waves, the surfaces of the body touching the bottom surface are not the ones that contract; with retrograde waves, however, these are the surfaces that do contract. As the retrograde wave approaches, the body area immediately adjacent to it is extended upward. The body surface within the contraction area then anchors itself to the bottom surface, after which the body is pulled forward.

Large flatworms use pedal locomotion instead of or in alternation with ciliary activity. In the gastropod and amphineuran molluscans (*e.g.,* snails and chitons, respectively), pedal locomotion is the primary locomotor mode and has become highly complex. The foot of these creeping animals is extremely muscular, penetrated by nerves, and capable of generating one, two, or four laterally adjacent contraction waves. If the foot generates a pair of waves, the lateral halves of the foot may alternate, thereby producing a shuffling movement, or they may be opposite. Generally, a foot can contain no more than one whole and two partial waves moving along a single axis.

Peristaltic locomotion is a common locomotor pattern in elongated, soft-bodied invertebrates, particularly in segmented worms, such as earthworms. It involves the alternation of circular- and longitudinal-muscle-contraction waves. Forward movement is produced by contraction of the circular muscles, which extends or elongates the body; contraction of the longitudinal muscles shortens and anchors the body (see below *Fossorial locomotion*).

Although peristaltic locomotion is frequently used by sea cucumbers, they and other echinoderms, such as sea urchins and starfishes, possess rows of tube feet that provide the main locomotor force. In starfishes, each arm bears hundreds of tube feet. Only one arm, however, becomes dominant in locomotion; while the tube feet on that arm move toward the tip of the arm, the tube feet of the other arms move in the same plane as those of the lead arm. Because there is no apparent metachronal wave of contraction within an arm, the movement of the tube feet is poorly coordinated, but small areas of the tube feet do move in synchrony. Each tube foot is a hollow elastic cylinder capped by a hollow muscular ampulla (a small, bladder-like enlargement). When the ampulla contracts, it forces fluid into the tube foot and extends it. Preferential contraction of muscles in the wall of the tube foot controls the direction of and the retraction of the tube foot. When the tube foot is fully contracted, fluid is withdrawn from it by relaxation of the ampulla, after which the muscles of the tube foot swing it forward in preparation for another step.

Tube feet

Swimming. Invertebrates have developed two distinct propulsive mechanisms for swimming: some use hydraulic propulsion; all others utilize undulations of all or parts of their bodies. The medusa (umbrella-shaped) body of coelenterates and ctenophores (*e.g.,* jellyfish and comb jelly, respectively) is a flexible hemisphere with tentacles and sense organs suspended from the edge; a manubrium (handle-shaped structure) bearing the digestive system hangs from the internal tip of the hemisphere. Enclosed in the outer margin of the medusa is a wide muscular band; when this band contracts, the opening of the medusa narrows. Simultaneously, water is ejected from the medusa through the narrow opening, and the animal is propelled upward. Because the contractions tend to be regular but slow, locomotion is somewhat jerky.

Scallops are the best swimmers among bivalve molluscans that can swim. Locomotion is produced by rapid clapping movements of the two shells, creating a water jet that propels the scallop. The muscular mantle (a membranous fold beneath the shell) acts as a valve and controls the direction of flow of the ejected water, thereby controlling the direction of movement. Nomally, the flow is directed downward on each side of the hinge that joins the two shells, and the resulting water jet lifts the scallop and moves it in the direction of the shell's opening. If necessary, however, escape movement may occur in the opposite direction. The scallop is adapted to swim even though it is two or three times as dense as seawater. The hinge is elastic and opens the shell rapidly; this action, coupled with rapid and repeated contractions of the adductor muscle, which closes the shell, produces a powerful and nearly continuous water jet. Moreover, the body form of a closed scallop is an airfoil (like a wing, the curvature of its upper surface is greater than that of its lower surface); this shape, combined with the downward ejection of water, produces lift.

Cephalopods (*e.g.,* squids, octopuses) are another group of mollusks that use hydraulic propulsion. Unlike the scallops, they have lost most of their heavy shell and have developed fusiform bodies. The mantle of cephalopods encloses a cavity in which are contained the gills and other internal organs. It also includes, on its ventral surface, a narrow, funnel-shaped opening (siphon) through which water can be forcibly ejected when all the circular muscles surrounding the mantle cavity contract rapidly and simultaneously. This water jet shoots the cephalopod in a direction opposite to that in which the siphon is pointed.

Many invertebrates, particularly elongated ones such as open-sea-dwelling annelids and mollusks, swim by undulatory movements produced by contraction waves that alternate on each side of the body. Although the arrangement of the musculature differs between invertebrates and vertebrates, the mechanics of undulatory swimming are the same in both and are described in the following section.

Undulatory swimming

Fish and fishlike vertebrates. Undulatory swimming is roughly analogous to using one oar at the stern of a boat. The side-to-side movements of the oar force the water backward and the boat forward. The undulatory movement of a fish acts similarly, although the motions involved are much more complex.

Anguilliform locomotion. When an elongated fish such as an eel swims, its entire body, which is flexible throughout its complete length, moves in a series of sinuous waves passing from head to tail. In this type of movement, which is called anguilliform (eel-like) locomotion, the waves cause each segment of the body to oscillate laterally across the axis of movement. Unlike the simple side-to-side movement of the oar, however, each oscillating segment describes a figure-eight loop, the centre of which is along the axis of locomotion. It is these oscillations and the associated orientation of each body segment that produce the propulsive thrust.

The undulatory body waves are created by metachronal contraction waves alternating between the right and left axial musculature. During steady swimming, several contraction waves simultaneously pass down the body axis from head to tail; the resultant undulatory waves move backward along the body faster than the body moves forward. As the undulatory wave passes backward, its amplitude and speed increase, thereby producing the greatest propulsive thrust in the tail (caudal) region. Propulsion, however, is not limited to the caudal region, for all undulating segments contribute to the thrust. Because the speed, amplitude, and inclination of each body segment differ, the thrust of each differs. In all segments, the greatest thrust is obtained as the segment crosses the locomotor axis, for here it is travelling at its greatest speed and inclination.

Carangiform and ostraciiform locomotion. All undulatory swimming movements generate forward thrust in the manner described above. Not all swimming animals, however, possess the elongated shape of an eel; only those with a similar body form, in which the surface area of the head end is the same as that of the tail end, have anguilliform locomotion. Fish with fusiform bodies exhibit carangiform locomotion, in which only the posterior half of the body flexes with the passage of contraction waves. This arrangement of body form and locomotion apparently is the most efficient one, for it occurs in the most active and fastest of fish. The advantage of carangiform locomotion appears to be related to the effectiveness of the posterior half of the body as a propulsive unit and the fact that the shape of the body and its small lateral displacement create little water turbulence. In contrast to ostraciiform locomotion, in which only the caudal fin oscillates from side to side in a manner similar to moving a boat with one oar, the length of the propulsive unit of carangiform fish enables the unit to obtain maximum oscillatory speed and inclination.

Locomotion of aquatic mammals

Whales also use undulatory body waves, but unlike any of the fishes, the waves pass dorsoventrally (from top to bottom) and not from side to side. In fact, many mammals that swim mainly by limb movements tend to flex their body in a dorsoventral plane. Whereas the body musculature of fish and tail musculature of amphibians and reptiles is highly segmental—that is, a muscle segment alternates with each vertebra—an arrangement that permits the smooth passage of undulatory waves along the body, mammals are unable to produce lateral undulations because they do not have this arrangement. Nor does the muscle arrangement of mammals permit true dorsoventral undulations; however, with an elongated caudal region, as in whales, they can attain a form of carangiform locomotion as effective as that of any fish.

Stabilization and steering. To stabilize and steer, most aquuatic vertebrates have, in addition to the caudal fin, a large dorsal fin and a pair of large anterolateral fins. Although they may possess other fins, these are of less importance. The balance of a swimming animal may be maintained in several ways. Rolling, or rotation, along the longitudinal axis of the body is reduced or controlled by any fins that extend at right angles to the body. Pitching, or dorsoventral seesawing, movements are counteracted by the anterolateral fins, which are also the major steering organs of fish, whales, and seals. Yawing, or lateral seesawing, is prevented by the dorsal fin and, if present, a ventral fin; for these fins to be effective, however, most of their exposed surface area should be behind the fish's centre of gravity. Because fins of the above type are not common in most invertebrates that swim by undulation, their locomotion is less stable.

Tetrapodal vertebrates. Many of the various types of undulatory locomotion described above are also widely used by aquatic tetrapods (those with walking appendages). Larval frogs, crocodilians, aquatic salamanders, and lizards, for example, have long muscular tails that propel them by undulatory motion. Most aquatic tetrapods, however, move by appendicular locomotion, for which the major propulsive units are the hindlegs. The exceptions are sea turtles, auks, penguins, and fur seals; in these, the hindfeet are webbed and are used as rudders. For propulsion, these animals use their forelegs, which have become bladelike flippers in which the forearm and hand region are dorsoventrally compressed to form a single, inflexible unit. The movements of such flippers are analogous to the aerial flight of birds; by moving synchronously, they provide lift and thrust in the water. Unlike aerial flight, however, the upper arms do not produce lift or thrust; instead, they serve only as a pivotal or leverage point for driving the flippers.

Swimming movements in sea turtles, penguins, and auks are accomplished by the rotation of the flippers or wings through a figure-eight configuration. In the birds, however, the stroke is relatively faster than in sea turtles, because the entire cycle appears to be proportionately smaller in amplitude. Moreover, because the birds' bodies are more streamlined, they can attain greater speeds than the turtles. Penguins may attain speeds of 40 kilometres (25 miles) per hour in water and have sufficient speed and thrust to enable them to leap two metres (six feet) or more above the water. The wings of penguins are so highly modified, however, that they have lost the ability to fly. The auks, on the other hand, are able to use their wings for both aerial and aquatic locomotion.

Aquatic birds

Some of the other aquatic birds, such as ducks and water ouzels, are said to propel themselves underwater through wing movements, but the evidence for such propulsion is incomplete and still open to question. The wing movements of ducks may be for steering and hydroplaning (skimming through the water) rather than for actual propulsion. The wings of the water ouzels, or dippers, were once thought to function as hydroplanes, but investigations have revealed that, although the wings are flapped underwater, the ability of dippers to bottom walk or fly underwater depends upon the velocity of the water flowing past the wings rather than the movement of the wings themselves.

Most aquatic birds are propelled by their webbed hindfeet, which tend to move alternately in surface swimming and in unison when the bird is submerged. Of all the swimming birds that use their hindfeet, the loons show the most extreme adaptations: the body, head, and neck are elongated and streamlined; the hindlegs are at the very posterior end of the body; the lower legs are compressed and bladelike; and the feet are strongly webbed. The webbing increases the surface area exposed to the water during limb retraction and also permits the folding of the foot, thereby reducing water resistance during protraction.

In frogs and freshwater turtles, the hindlegs are elongated and the feet enlarged and strongly webbed. But, whereas the hindlegs of frogs move synchronously, except occasionally in slow swimming, when they alternate, the limb movements always alternate in freshwater turtles. Some aquatic turtles, however, such as snapping, mud, and musk turtles, are very poor swimmers and will swim only under extreme conditions. These turtles are bottom walkers, and their limb movements in water are identical to those on land except that they can move faster in water than on land.

The swimming movements of many mammals are also identical with their terrestrial limb movements. Hippopotamuses spend much of their time in the water, yet

they bottom walk rather than swim. Most of the aquatic mammals—*e.g.*, otters, hair seals, aquatic marsupials, insectivores, and rodents—use their hindlegs and frequently their tails for swimming. The feet are webbed and usually move alternately; the tail tends to be flattened. Fur seals, polar bears, and platypuses swim mainly with forelimbs; only in the seals, however, are the movements of the forelimbs similar to those of sea turtles and penguins.

FOSSORIAL LOCOMOTION

The speed, manner, and ease with which animals move depends directly upon the compactness of the material and its cohesiveness. Many aquatic animals can swim through semisolid mud or muck suspensions, which lack compactness. Some lizards and snakes that live in an arid environment can swim through friable sand, which is compact but lacks strong cohesiveness. Although these swimming movements can be considered a form of fossorial locomotion, the following discussion considers only locomotor patterns in which most of the activity of the animals involved is confined to tunnels that they leave behind.

Burrowing

Fossorial invertebrates. Burrowing or boring invertebrates have evolved a number of different locomotor patterns to penetrate soil, wood, and stone, of which soil or mud is the easiest to penetrate. The soft-bodied invertebrates, such as worms and sea cucumbers, burrow either by peristaltic locomotion or by the contract–anchor–extend method. Their hydrostatic, or fluid, skeleton, combined with their circular and longitudinal musculature, permits controlled deformation of their shape, which allows them to squeeze into narrow spaces and then enlarge the spaces, thus creating a burrow or tunnel. Worms with a protrusible proboscis (a tubular extension of the oral region) generally burrow by the contract–anchor–extend method. Contraction of the circular muscles in the posterior half of the body drives the body fluids forward, causing the proboscis to evert (turn outward) and forcing it into the soil. When the proboscis is fully everted, the part of the body (collar) directly behind it dilates and anchors the proboscis in the soil. The entire body is then pulled forward by the longitudinal muscles and reanchored. This pattern produces the very jerky and slow forward progression typical of most fossorial locomotion.

Peristaltic locomotion (see Figure 14), which is generated by the alternation of longitudinal- and circular-muscle-contraction waves flowing from the head to the tail, is similar to the above pattern. Forward progression is more continuous, however, because of the contraction waves. The sites of longitudinal contraction are the anchor points; body extension is by circular contraction. The pattern of movement is initiated by anchoring the anterior end. As the longitudinal contraction wave moves posteriorly, it is slowly replaced by the circular contraction wave. The anterior end slowly and forcefully elongates, driving the tip farther into the surface as the circular contraction wave moves down the body. The tip then begins to dilate and anchor the anterior end as another longitudinal contraction wave develops. This sequence is repeated continuously, and the worm moves slowly forward. Reversing the direction of the contraction waves enables the worm to back up.

From J. Gray, *Animal Locomotion*, Weidenfeld & Nicolson Ltd., London

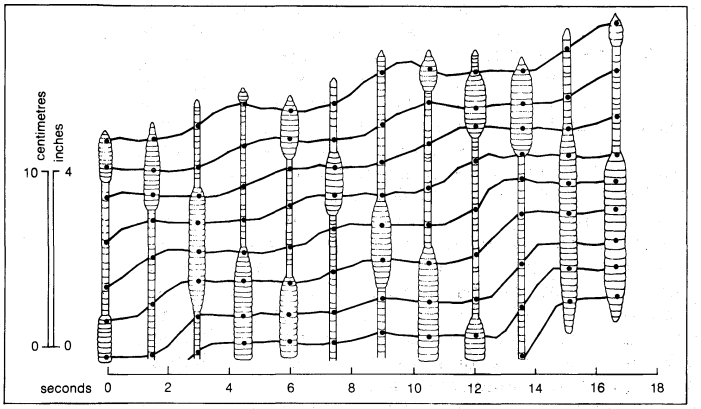


Figure 14: *Peristaltic locomotion in worms.*
Segments move forward except when they are longitudinally contracted (shown as wide areas with large dots). Individual points on the worm's body and their movements relative to each other are shown by the oblique lines.

Burrowing bivalve mollusks, such as clams, use the contract–anchor–extend locomotor mode. Such bivalves have a large muscular foot that contains longitudinal and transverse muscles as well as a hemocoel (blood cavity). The digging cycle begins with the extension of the foot by contraction of the transverse muscles. The siphons (tubular-shaped organs that carry water to and from the gills) are closed, and the adductor muscle of the shell contracts, thereby forcing blood into the tip of the foot and causing it to dilate. With the tip acting as an anchor, the longitudinal muscles then contract, pulling the body down to the anchored foot. Frequently, the longitudinal muscles contract in short steps and alternate between the left and right sides; this causes the shell to wobble and penetrate deeper as it is pulled down.

Rock borers

Some invertebrates are able to bore through rock. Most of the rock borers are mollusks; they bore either mechanically by scraping or chemically by the secretion of acid. The piddock, or angel's wing, bivalves, for example, attach themselves to a rock with a sucker-like foot. The two valves, held against the rock, grind back and forth by the alternate contraction of two adductor muscles; the grinding slowly produces a tunnel.

Fossorial vertebrates. The fossorial vertebrates are found in three classes: amphibians, reptiles, and mammals. Although some fishes and birds dig or bore shallow burrows, they can hardly be considered truly fossorial, as are moles or earthworms. Locomotion of fossorial amphibians and reptiles tends to be axial; it is appendicular only in mammals. Fossorial mammals have strong forelegs with a tendency toward flattening; their hands and particularly the claws are enlarged. Forelegs show the greatest modification in such species as moles and gophers, whose entire lives are spent in burrows. These animals tend to dig with a breast stroke, either synchronously or alternately, by extending the foreleg straight forward in front of the snout and then retracting it in a lateral arc. The loosened soil is compacted against the side walls of the burrow. In those fossorial species that dig burrows as nests but forage above the ground—many rodents, such as prairie-dogs, ground squirrels, and groundhogs—the digging movements tend to be dorsoventral with alternating limb movement. The forelegs are extended forward and then retracted downward and backward; the loosened soil passes beneath the body and is frequently pushed to the surface.

Fossorial reptiles and amphibians are usually legless, or the legs are so reduced that they serve no locomotor function; in most species, the head is flattened dorsoventrally, and the snout extends beyond and somewhat over the mouth. Burrowing is accomplished by one of three patterns analogous to the contract–anchor–extend locomotion of invertebrates. In the most common of these, the snout is driven straight forward along the bottom of the tunnel, the head is then raised, and the soil is compacted to the roof. The head tends to be laterally compressed in animals that use the other two patterns. In one of these patterns, the snout is shoved forward and then swung from side to side; in the other, the snout is rotated as it swings from side to side and seems to shave the walls of the tunnel.

TERRESTRIAL LOCOMOTION

Legs in arthropods and vertebrates

Walking and running. Only arthropods (*e.g.*, insects, spiders, and crustaceans) and vertebrates have developed a means of rapid surface locomotion. In both groups, the body is raised above the ground and moved forward by means of a series of jointed appendages, the legs. Because the legs provide support as well as propulsion, the sequences of their movements must be adjusted to maintain the body's centre of gravity within a zone of support; if the centre of gravity is outside this zone, the animal loses its balance and falls. It is the necessity to maintain stability that determines the functional sequences of limb movements, which are similar in vertebrates and arthro-

pods. The apparent differences in the walking and slow running gaits of these two groups are caused by differences in the tetrapodal (four-legged) sequences of vertebrates and in the hexapodal (six-legged) or more sequences of arthropods. Although many legs increase stability during locomotion, they also appear to reduce the maximum speed of locomotion. Whereas the fastest vertebrate gaits are asymmetrical, arthropods cannot have asymmetrical gaits, because the movements of the legs would interfere with each other.

Cycle of limb movements. The cycle of limb movements is the same in both arthropods and vertebrates. During the propulsive, or retractive, stage, which begins with footfall and ends with foot liftoff, the foot and leg remain stationary as the body pivots forward over the leg. During the recovery, or protractive, stage, which begins with foot liftoff and ends with footfall, the body remains stationary as the leg moves forward. The advance of one leg is a step; a stride is composed of as many steps as there are legs. During a stride, each leg passes through one complete cycle of retraction and protraction, and the distance the body travels is equal to the longest step in the stride. The speed of locomotion is the product of stride length and duration of stride. Stride duration is directly related to retraction: the longer the propulsive stage, the more time required to complete a stride and the slower the gait. A gait is the sequence of leg movements for a single stride. For walking and slow running, gaits are usually symmetrical—*i.e.*, the footfalls are regularly spaced in time. The gaits of fast-running vertebrates, however, tend to be asymmetrical—*i.e.*, the footfalls are irregularly spaced in time.

The different gaits of insects are based on the synchrony of leg movements on the left (*L*) and right (*R*) sides of the animal. The wave of limb movement for each side passes anteriorly; the posterior leg protracts first, then the middle leg, and finally the anterior leg, producing the sequence $R_3 R_2 R_1$ or $L_3 L_2 L_1$. There is no limb interference, because the legs of one side do not have footfalls along the same longitudinal axis. The slowest walking gait of insects is the sequence $R_3 R_2 R_1$ followed by the sequence $L_3 L_2 L_1$. As the rate of protraction increases, the protractive waves of the right and left sides begin to overlap. Eventually, the top speed is reached when the posterior and anterior legs of one side move synchronously. This gait occurs because the protraction times for all legs are constant, the intervals between posterior and middle legs and between middle and anterior legs are constant, and the interval between posterior and anterior legs decreases with faster movements. Other gaits are possible in addition to those indicated above by altering the synchrony between left and right sides.

Limb movements in millipedes and centipedes

The limb movements of centipedes and millipedes follow the same general rules as those of insects. The protraction waves usually pass from posterior to anterior. Because each leg is slightly ahead of its anteriorly adjacent leg during the locomotory cycle, one leg touches down or lifts off slightly before its anteriorly adjacent one. This coordination of limb movement produces metachronal waves, the frequency of which equals the duration of the complete protractive and retractive cycle. The length of the wave is directly proportional to the phase lag between adjacent legs.

Whereas the millipedes must synchronize leg movements to eliminate interference, the tetrapodal vertebrates must synchronize leg movements to obtain maximum stability. Four legs are the minimum requirement for symmetrical terrestrial gaits. Although bipedal (two-legged) gaits require extensive structural modifications of the body and legs, they still retain the leg-movement sequence of tetrapodal gaits (see Figure 15). The basic walking pattern of all tetrapodal vertebrates is left hindleg (*LH*), left foreleg (*LF*), right hindleg (*RH*), right foreleg (*RF*), and then a cyclic repetition of this sequence, which is equivalent to the slow walking gait of insects but with the middle legs removed. Unlike the insects, however, vertebrates can begin to walk with any of the four legs and not just the posterior pair. The faster symmetrical gaits of vertebrates are obtained by overlapping the leg-movement sequences of the left and right sides in the same manner as insects; for example, an animal can convert a walk to a trot by moving diagonally contralateral legs (those on opposite sides) simultaneously, or to a pace by moving the ipselateral legs (those on the same side) simultaneously. Many other symmetrical gaits occur between the walk and the pace and the trot, which are extreme modifications of the walk.

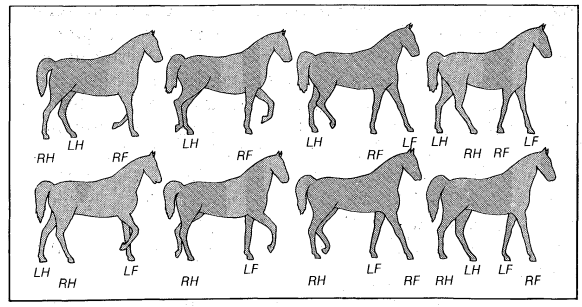


Figure 15: *Walking sequence of a five-gaited horse.* LH and RH refer to left and right hindlegs; LF and RF refer to left and right forelegs.

From A. Howell, *Speed in Animals*, © 1944; University of Chicago Press

Running

Cursorial vertebrates. Cursorial (running) vertebrates are characterized by short, muscular upper legs and thin, elongated lower legs. This adaptation decreases the duration of the retractive–protractive cycle, thereby increasing the animal's speed. Because the leg's cycle is analogous to the swing of a pendulum, reduction of weight at the end of the leg increases its speed of oscillation. Cursorial mammals commonly use either the pace or the trot for steady, slow running. The highest running speeds, such as the gallop, are obtained with asymmetrical gaits. When galloping, the animal is never supported by more than two legs and occasionally is supported by none. The fastest runners, such as cheetahs or greyhounds, have an additional no-contact phase following hindfoot contact.

In cursorial birds and lizards, both of which are bipedal, the feet are enlarged to increase support and the body axis is held perpendicular to the ground, so that the centre of gravity falls between the feet or within the foot-support zone. The running gait is, of course, a simple alternation of left and right legs. In lizards, however, bipedal running must begin with quadrupedal (four-footed) locomotion. As the lizard runs on all four legs, it gradually builds up sufficient speed so that its head end tilts up and back, after which it then runs on only its two hindlegs.

Saltation. The locomotor pattern of saltation (hopping) is confined mainly to kangaroos, anurans (tailless amphibians), rabbits, and some groups of rodents in the vertebrates and to a number of insect families in the arthropods. All saltatory animals have hindlegs that are approximately twice as long as the anteriormost legs. Although all segments of the hindleg are elongated, two of them—the tibial (between upper segment and ankle) and tarsal (ankle) segments—are the most elongated.

Types of saltatory patterns

There are at least four different saltatory patterns, but all are similar in that the simultaneous retraction or extension of the hindlegs is followed by an aerial phase of movement. The aerial phase in all patterns is governed by the physical principles of ballistics (the flight characteristics of an object): the height and the length of the jumps are functions of the takeoff velocity and angle. The longest jumps are attained when the takeoff angle is 45°.

Before jumping, the femur (upper segment of the hindleg) of the flea is held perpendicular to the ground, the tibia extends obliquely posterior, and the remainder of the hindleg extends posteriorly along the ground. Just prior to the jump, the middle legs flex and tilt the body upward; then the femur of the hindlegs swings sharply backward simultaneously with the extension of the tibia. This retraction forces the animal upward and forward at an angle of 50°. As the flea approaches touchdown, the front legs are swung forward and downward, the middle legs are held perpendicular to the body axis, and the hindlegs project obliquely posterior. The anterior two pairs of legs thus act to absorb the landing shock.

The frog jump (Figure 16) is initiated with three simul-

taneous movements: the forelegs flex, and the back arches to tilt the entire body upward; the tarsus of the hindleg swings to a vertical position and locks; and the femur, extending anteriorly along the body, swings in a horizontal plane. When the femur is perpendicular to the body, the knee joint snaps open, and the frog jumps forward at a 30° to 45° angle. As the frog begins to land, the forelegs are protracted and held downward in front of the chest. The forefeet touch down first, the forelegs acting as shock absorbers. Simultaneously, the hindlegs are protracted so that they can be in jumping posture by the completion of landing.

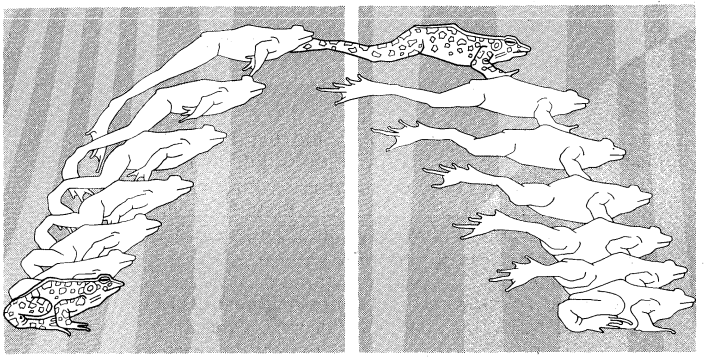


Figure 16: (Left) Jumping and (right) landing sequences of frogs.

The positions and movements of the hindlegs in rabbits and kangaroos are similar to those of the frog. The major difference is that rabbits, kangaroos, and all other mammals move their legs in a vertical plane instead of a horizontal plane, as do the frogs; because the femur and tibia move vertically, the tarsus need not be elevated to prevent the hindleg from hitting the ground. The saltatorial gait of rabbits is quadrupedal, whereas that of kangaroos is bipedal. A jumping rabbit stretches forward and lands on its forefeet; generally, both forefeet do not touch ground simultaneously, however. As the forefeet touch, the back flexes, and the hind end rotates forward and downward. The hindfeet touch down lateral to the forefeet, and, as the back extends, a new jump begins. In contrast, the kangaroo lands on its hindfeet, and the back is held fairly straight through all phases of the jump, although the body inclines forward at takeoff and posteriorly when landing.

Crawling patterns

Crawling. Invertebrates crawl either by peristaltic locomotion or by contract–anchor–extend locomotion, both of which have been described previously (see above *Fossorial locomotion*). Limbless vertebrates, however, crawl in one of four patterns: serpentine, rectilinear, concertina, and sidewinding. The most common pattern, serpentine locomotion, is used by snakes, legless lizards, amphisbaenids (worm lizards), and caecilians (wormlike amphibians). Rectilinear locomotion is used by the giant snakes and almost exclusively by fossorial vertebrates when burrowing. Concertina and sidewinding locomotion are largely confined to snakes.

Serpentine locomotion. In serpentine locomotion, in which the body is thrown into a series of sinuous curves, the movements appear identical to those of anguilliform swimming, but the similarity is more apparent than real. Unlike anguilliform swimming, when a snake starts to move, the entire body moves, and all parts follow the same path as the head. When the snake stops moving, the entire body stops simultaneously. Propulsion is not by contraction waves undulating the body but by a simultaneous lateral thrust in all segments of the body in contact with solid projections (raised surfaces). The muscular thrust against the projection is perpendicular to the axis of the pushing segment. To go forward, therefore, it is necessary for the strongest thrust to act against the side of the projection facing in the direction of movement. Because of this, thrust tends to occur at the anterior end of the concave (inward-curving) side of the loop of the snake's body.

Concertina locomotion. Concertina locomotion is used when there is not enough frictional resistance along the locomotor surface for serpentine locomotion. After the body is thrown into a series of tight, sinuous loops, forming a frictional anchor, the head slowly extends forward until the body is nearly straight or begins to slide. The anterior end forms a small series of loops and, with this anchor, pulls the posterior regions forward, after which the sequence of movements is repeated. This crawling pattern is analogous to the contract–anchor–extend locomotion of invertebrates, but, because snakes lack the body flexibility provided by a hydrostatic skeleton, they must depend upon the body loops.

Sidewinding. Sidewinding, which is also used when the locomotor surface fails to provide a rigid frictional base, is a specific adaptation for crawling over friable sandy soils. Like serpentine locomotion but unlike concertina locomotion, the entire body moves forward continuously in sidewinding locomotion (see Figure 17). Although the body moves through a series of sinuous curves, the track made by the snake is a set of parallel lines that are roughly perpendicular to the axis of movement. This is because only two parts of the body touch the ground at any instant; the remainder of the body is held off the ground. To begin sidewinding, the snake arches the anterior part of the body forward and forms an elevated loop with only the head and the middle of the body in contact with the ground. Because each part of the body touches the ground only briefly before it begins to arch forward again, the snake seems to roll forward like a short, coiled spring. In a continuously repeating cycle, as a segment arches forward, the posteriorly adjacent segment touches down.

Tracks of sidewinders

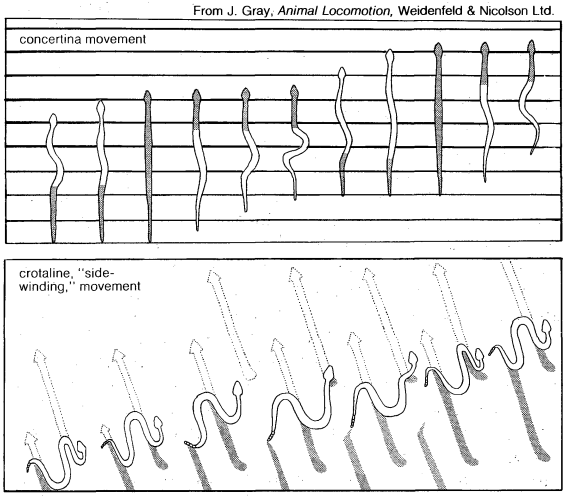


Figure 17: Two modes of locomotion in the snake.

Rectilinear locomotion. Unlike the three preceding patterns of movement, in which the body is thrown into a series of curves, in rectilinear locomotion in snakes the body is held relatively straight and glides forward in a manner analogous to the pedal locomotion of snails. The ventral (belly) surface of snakes is covered by scales elongated crosswise that overlap like roof shingles, with the opening of the overlap facing toward the posterior. Each ventral scale is moved by two pairs of muscles, both of which are attached to ribs but not to ribs of the same segment as the scale. One pair of muscles is inclined posterior at an angle (obliquely); the other is inclined anterior at an angle. As contraction waves move rearward from the head simultaneously on both sides, the anterior oblique muscles of a scale contract first and lift the scale upward and forward. When the posterior oblique muscles contract, the scale is pulled rearward, but its edge anchors it, and the body is pulled forward. This sequence is repeated by all segments as the contraction wave passes posteriorly, and, as a series of contraction waves follow one another, the body slowly inches forward.

ARBOREAL AND AERIAL LOCOMOTION

Climbing. The adaptation for climbing is unique for each group of arboreal animals. All climbers must have

strong grasping abilities, and they must keep their centre of gravity as close as possible to the object being climbed. Because arthropods are generally small and, thus, not greatly affected by the pull of gravity, they show little specific structural adaptation for climbing. In contrast, the larger and heavier bodied vertebrates have many climbing specializations. In both arthropods and vertebrates, however, no leg is moved until the others are firmly anchored.

Arboreal amphibians and reptiles. Arboreal frogs are slender-bodied anurans with tapering legs and feet. The tips of the toes (digits) are expanded into large, circular disks that may function as suction cups, although such an action has not yet been definitely demonstrated. The disks, however, do increase the contact area, thereby improving grasping ability. The leg-movement sequence during climbing is that of a walking gait.

Surface adherence

Arboreal lizards have the same type of climbing gait as arboreal frogs, and their climbing specializations are also similar to those of anurans. They have a different type of climbing foot, however, because of the presence of claws and scales on the digits. Moreover, the entire digits, rather than just their tips, may be expanded. On the bottom of each of these spatula-shaped expansions are one or two rows of transversely elongated scales. Although not visible to the naked eye, the surface of these scales is covered with fine projections that increase their ability to adhere to a surface. Because of this strong adherence, the toes roll off and on the surface on which the animal is walking. Unlike other arboreal lizards, chameleons possess a prehensile (grasping) tail and zygodactylous feet—*i.e.,* the toes are fused into two opposable units. Although these adaptations are inferior for vertical climbing, they are superior for locomotion on vertical or inclined, slender branches. Arboreal snakes tend to have either prehensile tails or extremely elongated bodies.

Climbing birds and mammals. Although the strong, clawed feet of birds permit many of them to climb occasionally, most truly scansorial (climbing) birds cling with their strong feet and brace themselves with stiffened tail feathers. Birds such as woodpeckers and tree creepers usually climb vertically upward, usually with both feet moving simultaneously in short, vertical hops. This mode of locomotion, however, prevents vertical descent. Only the nuthatch can descend as easily as it can ascend; it climbs obliquely, using the upper foot for clinging and the lower foot as a brace. Parrots have developed zygodactylous feet as an aid to climbing; in addition, they frequently use their bills when climbing vertically.

Use of prehensile fingers or claws

Several locomotor patterns for climbing are used by arboreal mammals, the grasping ability of which has been enhanced by the presence of either strong claws or prehensile fingers. Many monkeys use a climbing gait similar to the leg sequence of walking. Occasionally, however, they use a leg sequence equivalent to that of a trot. Small-bodied climbers with sharp claws, such as squirrels, climb by the alternate use of forelegs and hindlegs; essentially, they hop up a tree. Prehensile-fingered climbers descend backward and generally with a walking type of leg sequence. Sharp-clawed species descend with a similar gait sequence but with the head downward.

Leaping. The mechanics of arboreal leaping do not differ from those of terrestrial saltation; the upward thrust in both is produced by the rapid, simultaneous extension of the hindlegs. Because of the narrowness of the arboreal landing site, however, landing behaviour does differ. Arboreal leaping also tends to be a discontinuous locomotor behaviour that is used only to cross wide gaps in the locomotor surface. Leaping from limb to limb, although occasionally employed by most climbers, appears to occur most frequently in animals with opposable or at least prehensile forefeet, particularly tree frogs and primates. Such forefeet enable the animal to grasp and hold onto the landing site.

Brachiation. True brachiation (using the arms to swing from one place to another) is confined to a few species of primates, such as gibbons and spider monkeys. Because the body is suspended from a branch by the arms, brachiation is strictly foreleg locomotion. When the animal moves, it relaxes the grip of one hand, and the body pivots on the shoulder of the opposite arm and swings forward; then the free arm reaches forward at the end of the body's swing and grabs a branch. The sequence is then repeated for the other arm. This locomotor pattern produces a relatively rapid and continuous forward movement but is restricted to areas with thick canopies of trees. Brachiators have arms that may be as long or longer than the body and a very motile shoulder joint.

Gliding. There are two functionally distinct forms of gliding, gravitational gliding and soaring: the former is used by gliding amphibians, reptiles, and mammals; the latter is restricted to birds. All gliders are able to increase the relative width of their bodies, thereby increasing the surface area exposed to wind resistance. The few gliding frogs flatten their bodies dorsoventrally and spread their limbs outward. Gliding snakes not only flatten their bodies but also draw in the ventral scales, thereby creating a trough. The best adapted gliding lizards have elongated ribs that open laterally like a fan.

Gliding mammals, such as the African flying squirrel and the colugo, usually have, on each side of the body, a fold of skin (the patagium) that extends from their wrist or forearm backward along the body to the shank of the hindleg or the ankle. When gliding, they assume a spread-eagle posture, and the patagia unfold.

Gravitational gliding. Gravitational gliding is equivalent to parachuting. Because the expanded lateral surface of the body increases the wind resistance against the body, the speed of falling is reduced. The directions of gliding can be controlled by adjusting the surface area—to curve to the right, the right patagium is relaxed. Gliders can land on vertical surfaces by suddenly turning the anterior end of the body up as it reaches the surface. Mechanically, this stalls the flight—*i.e.,* the horizontal component of flight is eliminated.

Soaring. Gravitational gliding is one of the basic mechanisms of soaring, which is restricted to birds, although birds must obtain their initial elevation by means of flapping flight. The second basic mechanism of soaring involves wind or air currents. Soaring requires that air currents meet one of two conditions: either the air must have a vertical velocity exceeding the rate of descent in gravitational gliding, or it must have a horizontal velocity that is nonuniform in time and space. Whereas static soaring depends upon vertical air currents, dynamic soaring depends upon horizontal air currents. Both types of soaring are described below.

Static and dynamic soaring

Vertical air currents for static soaring are produced when wind strikes an obstruction and is deflected upward. The sites of deflection are very local and discontinuous and seldom extend more than 30 metres (100 feet) above the obstruction. The height of deflection and the vertical velocity of the air are a function of the angle of deflection and the velocity of the wind. If the vertical velocity of the air equals the descent speed of the bird, the bird remains stationary in height relative to the ground. If, however, the vertical velocity is greater, the bird rises, and, if less, the bird falls at a speed equal to the gravitational descent speed minus the air's vertical ascent speed. The horizontal velocity of the air determines the bird's movements relative to the ground in the same manner as that of the vertical velocity.

The soaring flights of vultures and hawks depend upon vertical hot-air currents called thermals. Such currents are not continuous updrafts or downdrafts originating from a specific spot; instead, as a local region of the ground is heated, a vertical, hot-air updraft is created. At the top of the column, a thermal bubble is formed by the hot air curving outward, downward, and then around the bubble. It is then pinched off by cool air flowing into the column and floats upward. The free-floating thermal bubble is doughnut shaped, with the air rising in the centre and cycling outward and downward. Soaring birds spiral downward in the updraft; however, because the bubble rises faster than birds descend, soaring birds are carried upward, but at a speed less than that of the bubble. When a bird reaches the bottom of the bubble, it begins a straight gravitational glide until it reaches the next thermal bubble. Thus, static soaring in a thermal bubble can be

recognized by its alternating flight pattern of circling and straight gliding.

Unlike static soaring, which is done at relatively high altitudes over land, dynamic soaring is done at low levels and is usually restricted to oceanic areas. Dynamic soaring depends upon a steady horizontal sea wind, which is laminated into layers of different velocities because of the frictional interaction between the water and the air; the lower layers have the lowest velocity. The flight path of a bird performing dynamic soaring tends to be a series of inclined loops that are perpendicular to the direction of the wind. A soaring albatross, for example, will begin its gravitational glide approximately 15 metres (50 feet) above the sea. Because it glides downwind, its velocity is increased both by descent and by the wind at its tail. As the bird nears the sea, it makes a turn into the wind, and the forward flight velocity derived from the downwind glide and the tailwind combine to lift the albatross slowly back to its initial gliding height, but with a loss of horizontal velocity. The bird therefore turns downwind again and begins to repeat the soaring cycle.

Because it depends upon the presence of a horizontal air current, the flight of flying fish is more akin to soaring than to true flying. As a flying fish approaches the water surface, its pectoral and pelvic fins, which are analogous to the forelimbs and hindlimbs of quadrupeds, are pressed along the side of the body. The greatly enlarged, winglike pectoral fins then spread out as the fish leaves the water. The wind against the fins provides lift to raise the body above the water, and the tail continues to undulate to provide additional thrust. When the entire body is out of the water, the enlarged pelvic fins extend, and the fish glides for a short distance until its forward velocity is lost. Occasionally, as a fish drops back into the water, it will undulate its tail to initiate another short flight.

True flight. Three animal groups have developed true flight: insects, birds, and mammals. All generate forward thrust by flapping lateral appendages, and all are free of any dependence on gravitational descent or air currents. It should be noted at the outset, however, that, although the aerodynamics of flight are identical in all three, the following cycles of wing movements described for the different animal groups are generalizations; each species in a group has a distinctive flight pattern and, therefore, a distinctive pattern of wing movement.

Wing movements

Flight is produced by the simultaneous rotation of the left and right wings in a circle or in a figure eight. This rotation produces the upward thrust, or lift, necessary to overcome gravity and the forward thrust required to overcome drag. As the downward and backward phase of rotation forces the air backward and the body forward, lift is produced by the unequal velocities of the air across the upper and lower wing surfaces.

Wings of insects. In flies with one pair of wings, the rotation of the tip inscribes a posterior inclined oval. At the top of the wing cycle, the tip lies above the junction of the thorax and abdomen. The wing then beats downward and forward so that the tip ends anterior and below the head. To insure maximum thrust, the broad surface of the wing lies parallel to the horizontal body plane during the downstroke. During the path of the upstroke, which is the reverse of the downstroke, the wing is feathered (turned) by inclining it perpendicular to the body plane. Although the rotational cycle of those insects with two pairs of wings follows a similar path, the upward and downward strokes of the anterior and posterior wings are not simultaneous; the anterior pair usually lags behind the posterior pair.

The wings of insects are rotated by pulsation of the thorax, not by a set of muscles. Basically, the thorax is a rigid box to which the wings are attached by a pair of longitudinal lateral hinges that enable the thorax to move dorsoventrally. Four sets of muscles control the major movements. Contraction of a perpendicular set, which extends from the centre of the floor of the thorax to its roof, depresses the thorax and, because of a reverse linkage between wing and thorax, raises the wing. Contraction of a diagonal set, which extends from the anterior roof of the thorax to its posterior floor, elevates the thorax and lowers the wing. Two diagonal sets of muscles extend laterally from the floor to the wall of the thorax and are responsible for maintaining a relatively constant width in the thorax.

Wings of birds and bats. Unlike insect wings, the wings of birds and bats are linked structures, the lateral extent and regional inclination of which are altered intrinsically by muscular and bony segments. The up-and-down strokes of a bird's wing are produced by large chest (pectoral) muscles that extend from the sternum (breastbone) to the lower surface of the humerus (a bone in the upper arm). When these muscles contract, the wing is lowered; it is raised by the contraction of a small anterior pectoral muscle that is attached to the upper surface of the humerus by a long tendon.

Hovering and propulsive flight

Birds exhibit two major flight patterns, hovering flight and propulsive flight. Hovering flight is of fairly restricted use and is observed most frequently in the hummingbirds. The path of the wings inscribes a horizontal figure eight whose centre is perpendicular to the shoulder joint. The downward stroke of the wings is actually a slightly inclined anterior stroke, and, because the longitudinal body axis is nearly perpendicular to the ground, the upward stroke is a horizontal posterior stroke. Both strokes are power strokes that produce lift: on the downstroke the dorsal wing surface is the top of the airfoil surface; on the upstroke the ventral surface is the top of the airfoil surface.

Most birds and bats, however, utilize propulsive flight. Because the body is not stationary, as it is in hovering flight, the wing always moves forward relative to the air, and its tip generally inscribes an oval or figure-eight path. In a pigeon, for example, the downstroke begins with the wing fully extended and perpendicular to the back (Figure 18). As the wing moves downward and anterior, it draws

From J. Gray, *Animal Locomotion*; Weidenfeld & Nicolson, London

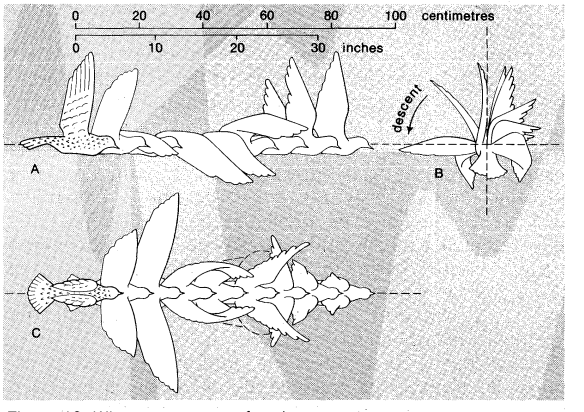


Figure 18: *Wing movements of a pigeon.*
(A) View from right side; (B) View from behind; (C) View from above.

level with the body, at which point the upper arm section stops while the distal part completes the downward path. At the bottom of the downstroke, the distal part turns outward and is elevated rapidly by the combined protraction of the humerus and the extension of the distal section.

DIRECTIONAL CONTROL

Although an animal's locomotor pattern may be controlled by its nervous system, directional control is impossible without sensory input. Two factors are involved in directional control: orientation, the ability of an animal to determine and to alter its position in the environment; and steering, the mechanical alteration of the locomotor pattern through which the animal adjusts its position.

Orientation. Orientation of locomotor behaviour is usually categorized as either kinesis or taxis. In kinesis, as previously explained, an animal's body is not oriented in relation to a sensory stimulus; rather, the stimulus causes an alteration in speed or direction of movement. In wood lice, for example, the kinetic response alters only the rate of movement. Because wood lice tend to aggregate in moist areas, their ambulatory activity increases or decreases as the relative humidity decreases or increases, respectively. In the planarian (an aquatic, ciliated flatworm), on the other hand, the kinetic response affects only the rate at

which the planarian changes its direction. Because planaria tend to stay in or return to darker areas, an increase in light intensity causes an increase in their turning responses. Generally, however, animals tend to alter both direction and speed as a single kinetic response.

Taxis

In taxis, an animal orients itself in a specific spatial relationship to a stimulus. The orientation may be simply an alteration of body position or it may be an alteration of locomotor direction so that the animal moves toward, away from, or at a fixed angle to the source of the stimulus. Sources that elicit a taxis response, which may cause a modification of speed, direction, or both, seem to encompass the entire range of environmental stimuli, such as gravity (geotaxis), temperature (thermotaxis), light (phototaxis), or chemicals (chemotaxis). If the response is negative, the animal moves away from the source; if it is positive, the animal moves toward the source.

The control of the response to a taxis is of two types. In open-system control, the initial response to a stimulus has no effect on subsequent responses to the same stimulus. A male firefly, for example, locates a female by the latter's brief flashes of light. When a male sees a female's flash, the male turns in the direction of the female, even though the source is no longer visible. If another female flashes, however, the male responds to the second flash in exactly the same manner as it did to the first. In close-system control, on the other hand, the response is progressively altered by feedback so that all subsequent responses are adjusted to the initial response. A bat chasing a flying insect will alter its flight path to intercept that of the insect. The bat's initial change in direction is only a general alteration of its course, but, as it approaches the insect, the bat constantly modifies its course to obtain an accurate interception.

Steering. Animals obtain accurate directional response (steering) by changing their propulsive response. Because steering relies heavily on continuous feedback (the communication cycle in which the motor output, or behaviour, is constantly being modified by the sensory input, or stimulus), it requires a precise integration of the central and peripheral nervous systems. (The central nervous system—in vertebrates, the brain and spinal cord—is that part of the nervous system that receives sensory impulses and sends out motor impulses; the peripheral nervous system consists of all the nerves that carry impulses between the central nervous system and other parts of the body.) Exteroceptive stimuli (those that originate outside the body) received by the peripheral nervous system establish the animal's spatial position in the environment; proprioceptive stimuli (those that originate inside the body), also received by the peripheral nervous system, establish the relative position of the body units to each other (see further SENSORY RECEPTION). Through integration of these two sets of stimuli, the central nervous system continuously adjusts the contraction of the motor units (*e.g.*, muscles) in order to obtain the desired orientation.

During locomotion, steering is a continual process. The direction of movements must be constantly adjusted to counteract environmentally produced deviations of direction. The apparently simple act of a bird flying from a tree to the ground illustrates the complexity of directional control. As the bird flies to the ground, it must be constantly aware of its height above the ground, the orientation of its body axis relative to the ground, deviations in flight direction resulting from air currents, and its speed of fall. All these parameters are determined primarily by exteroceptive stimuli received through the eyes and inner ears. The downward flight is constantly adjusted in response to these exteroceptive stimuli, and the fine control necessary for these adjustments is obtained by proprioceptive feedback.

(G.R.Z./Ed.)

Avoidance behaviour

In one of its major meanings "avoidance" is any behaviour induced by adverse stimuli. The underlying implication that a single neural mechanism is involved (such as a specific part of the brain, which, under electrical stimulation, seems to inflict punishment) remains only a hypothesis. Clearly, the same kinds of avoidance behaviour might result from different underlying physiological mechanisms. Thus, although the various dichotomies, or polarities, of behaviour such as positive and negative, psychoanalytic life and death instincts, and approach and withdrawal concepts may be logical or philosophical conveniences, they seem, nevertheless, to lack clear meaning physiologically.

Alternative usage defines avoidance behaviour by describing a number of patterns: active avoidance (fleeing), passive avoidance (freezing stock-still or hiding), and a pattern of protective reflexes, as seen in the startle response. There is good reason to suppose that, in cats, for example, each of these patterns is coordinated separately by the brain. One kind of fleeing, in which the cat moves continuously and shows much upward climbing, is produced by electrical stimulation of specific parts of the brain (hypothalamic sites). Stimulation of other sites (in the thalamus) generates other types of fleeing movements, causing the animal to crouch, look around, move, slink close to the floor, and hide, if possible. In general, among birds and mammals, brain sites for fleeing of the first type occur in hypothalamic and mesencephalic zones.

Brain function in coordination of patterns

Protective reflexes in mammals include ear retraction to a position of safety—pressed against and somewhat behind the skull—as when a horse is seen to lay its ears back. Among the monkey-like bush babies (*Galagos*) the outer ear folds up laterally and longitudinally at the same time, under threat. The eyes are closed, and the muscles around the eye are contracted, adding to the protection. During this so-called startle reflex, breathing is checked, and the mouth corners are pulled back to expose the teeth; this prepares both for biting in defense and also for movements of the tongue and for head shaking to free the mouth of any dangerous or distasteful substance that may have been taken in. In most mammals, the limbs flex as if ready for a leap; in the human startle reflex, the arms are thrust outward as if ready to grasp at a support.

It is helpful to consider avoidance behaviour in terms of factors that elicit it (*e.g.*, specific stimuli) and regulate it (*e.g.*, hormones).

FACTORS IN AVOIDANCE BEHAVIOUR

Specific stimuli. Warning calls and visual signals that are unique to different species of birds and mammals effectively and specifically evoke avoidance patterns. In some cases, learning clearly emerges as a factor; thus, members of a colony of birds seem to learn to respond to the alarm calls of all species present in the colony. Among ducklings, a visual model to evoke fleeing and hiding can be fashioned as a cardboard cutout. When moved overhead in one direction, the model resembles a short-necked, long-tailed hawk, and the ducklings flee from it; when moved in the other direction, the model looks like a harmless, long-necked goose, and the ducklings tend to stay calm. The model is effective, however, in eliciting the two kinds of behaviour only when the ducklings are accustomed to geese flying over but not hawks.

Innate factors also contribute to such responses (see above *Instinctive behaviour*). Domestic chicks, for example, show crouching and freezing in response to the long alarm call of their species. Many of the perching birds (passerines) will gather to mob when stimulated by the sight of an owl. The eyes in the characteristic owl face have been found to be especially important; even birds reared in isolation respond to man-made models with appropriate eyespots painted on. It has been suggested that many human beings are specifically (and perhaps instinctively) disturbed by the sight of snakes—the notion of a legless object perhaps being a key stimulus. Human responses to spiders and centipedes with conspicuous legs also may be intense. In the reaction to snakes at least, notwithstanding Freudian explanations that they symbolize male sex organs, the behaviour of people may be compared with owl mobbing among passerine birds.

Specific chemical signals can induce avoidance behaviour; some are released by minnows and tadpoles when their skin is damaged (usually indicating to fellows that there is danger). These chemicals appear to be specific for each species of fish and are highly effective in producing fleeing (see SENSORY RECEPTION: *Chemoreception*). Many

ants produce volatile alarm substances (terpenes) that are attractants to other ants at low concentrations and, in high concentrations near their source, produce rapid locomotion, defense postures, and, sometimes, fleeing. Some invertebrate avoidance responses are reflexes evoked by very specific stimuli; rapid swimming by cockles clapping their shells, for example, is elicited by starfish extract. Shell jerking is produced in a freshwater snail (*Physa*) by contact with a leech, another specific response to a major predator.

Pain, startle, and novelty. Painful stimuli are preeminent among those that produce avoidance. Among mammals (including man) many such responses are patently inborn, as is the reflex withdrawal of one's finger from a hot griddle.

To classify a stimulus as startling or novel requires some comparison with previous stimulation. Human responses (orientation reflex) to startling or interesting stimuli may be studied by presenting a series of repeated tones; the orientation reflex tends to appear at the moment at which some change in the usual sequence (such as a longer or shorter tone) occurs. There is some evidence that the hippocampus (a brain structure) is involved in the human experience of novelty. Surgical removal of the hippocampus in many animals makes avoidance responses to strange objects far more persistent; a comparable operation in small parrots (lovebirds) greatly increases the persistence of calls that gather others for mobbing. Probably the hippocampus takes part in establishing memory of any new stimulus, and once this has occurred, the stimulus is no longer novel. Removal of other brain structures (the amygdala) reduces avoidance of strange objects (*e.g.*, in lovebirds) and also makes fleeing and defensive attack less likely.

Probable role of the hippocampus

Passive and active avoidance. Passive avoidance is achieved by the inhibition of a previously exhibited response. Thus, after a laboratory animal has learned to approach a food dish, it may then be punished by an electric shock whenever a selected visual or auditory stimulus is present. In passive avoidance, the animal may freeze as soon as the stimulus is given; in active avoidance, the animal is given the opportunity of fleeing.

Freezing proper entails general motor inhibition, which, if sustained, may pass into signs of reduced arousal. States of considerable loss of muscle tonus, of eye closure, and many signs of deep sleep have been variously termed feigning death or animal hypnosis. In very young fowl, such signs can be induced simply by holding the animal firmly if the experience is novel (and thus presumably frightening). Such states tend to occur as an alternative to fleeing when the apparently frightening stimulus is difficult to locate or to escape. Among social mammals (*e.g.*, cats or dogs) the status and confidence of an animal may be inferred from its degree of leg extension, arched back (vertebral tonus), and cocky tail elevation. Threat from which there seems no escape may induce a progressive approach to immobility and to general motor inhibition.

Punishment. Inhibitory interconnections have been postulated between the punishment and reward systems within the brain. One line of evidence suggesting a single punishment system rather than a number of them includes behavioral and neurological resemblances in the responses of animals to fear-inducing and to frustrating circumstances. If either fear or frustration is induced during conditioning, both produce resistance to extinction. Both are specifically opposed by barbiturate drugs.

Whatever its physiological basis, negative reinforcement (punishment) can induce in an animal both the inhibition of the response that produced the punishment and the avoidance of the location at which it occurred. Sometimes the tendency to show avoidance behaviour develops further with time, even without additional training. Thus, when being conditioned to discriminate between stimuli (*e.g.*, two tones), some breeds of dog (*e.g.*, Alsatians), if made to wait for food reward or given an impossible discrimination to perform, will howl and show great excitement. On later days, they may first resist mounting the conditioning stand and finally resist approaching the room to which earlier they ran eagerly, presumably for the rewards of food. Stimuli associated with the training room are sometimes said to act as secondary negative reinforcers (secondary punishments) in such a case.

Even a piece of cockroach nervous system (metathoracic ganglion) and the leg it controls have been shown to be capable of avoidance conditioning. If each contact of the leg with a water surface is paired with an electric shock, the leg comes to be retracted on contact with the water; no such change occurs in a control leg receiving the same number of shocks at random. The conditioning is accompanied by a very marked decrease in a chemical (acetylcholinesterase) found in the nervous system; since it greatly facilitates transmission of some nerve impulses, such a chemical may well be basic to this primitive kind of learning.

Avoidance conditioning of the leg of a cockroach

Hormonal effects. Male hormones (androgens) cause the performance of new mobbing calls in the breeding season by many male passerine birds (*e.g.*, chaffinch) and also some other birds (*e.g.*, farmyard cock); it is not certain whether the effect is specific to the vocalization or whether the hormone produces a general change in responsiveness to frightening stimuli. Female hamsters are initially faster than males to emerge from a box and also move about more in a strange place; perhaps females innately tend to be less nervous. Females behave more like male hamsters if given a small injection of male hormone (testosterone) in the second day of life; the adult difference survives castration, so it probably rests on sexual differentiation of the nervous system rather than on adult hormone levels.

The adrenocorticotropic hormone (ACTH) from the pituitary glands of many animals may facilitate avoidance behaviour. ACTH has other direct effects on the nervous system (*e.g.*, facilitating male sexual behaviour).

FUNCTIONS OF AVOIDANCE BEHAVIOUR

Fleeing and escape. Most animals capable of locomotion show a rapid locomotor reflex to painful or startling stimuli. Such a reflex is very ancient in an evolutionary sense; it is present even in such primitive marine animals as the slender, tiny, translucent amphioxus. The rapid propulsion of an octopus or squid by its own jet of water or of a crayfish by a blow of its tail, the sudden leap and flight of a grasshopper, and the retraction of a worm into its hole—all are examples of such avoidance behaviour.

Reflex escape movements

Many invertebrates commonly compete in speed against their vertebrate predators, which tend to have faster conducting individual nerve cells; in order to compete successfully, the invertebrates seem to have evolved giant nerves (bundle of individual cell fibres), for the broader a nerve is, the faster it conducts. Among such lower animals, perhaps one-third or more of the nerve cord running the length of the body is made up of fibres responsible for initiating the escape response of the species. The fibres of a cockroach, for example, activate a mechanism that produces rapid running when the rear end (anal cerci) is disturbed by air movements. Bony fishes also have such structures, the Mauthner cells, that initiate escape swimming when stimulated.

Escape may be facilitated not only by speed of response but also by its explosive onset (*e.g.*, after a period of shamming death), making it difficult for a predator to predict the behaviour of a prospective meal. Escape movements may stop as suddenly as they start. Many animals may even be especially conspicuous in escape (*e.g.*, showing coloured hind wings, as do some grasshoppers and moths), so that their disappearance appears even more sudden. Presumably, the predator, engaged in pursuing and tracking a moving prey, finds its difficult to shift quickly enough to a different kind of search and so is unable to localize the exact point of disappearance.

In many instances, rapid locomotion is enough to frustrate a predator; in others, direction is crucial (*e.g.*, a fish moving upward to the water surface or downward to the bottom or, among birds, a more elaborate celestial orientation). Under threat, insects such as pond skaters (*Vellia*) flee toward the nearest shore; beach fleas (amphipods) flee to the sea; and particular populations of ducks have a preferred compass direction for escape (so-called nonsense orientation).

Freezing. Immobility usually makes detection less like-

ly. For stick insects and other animals resembling twigs or leaves, when immobility itself becomes conspicuous against moving foliage, the animals' compensatory swaying increases the camouflage effect. There seems to be an evolutionary conflict between camouflage and the need for conspicuous signals in communication. Social groups commonly keep in touch by calls or by movements such as tail flicks, which are inhibited during freezing or even under incipient immobility. Movements may be made conspicuous by patches of white or colour on a bird's outer tail feathers, or under a mammal's tail. The well-known white rear patch of hair among antelopes, for example, is hidden when the tail is folded or lowered under conditions of safety.

Protection reflexes, armour, and spines. Facial protective reflexes are usually well developed in flat-faced mammalian predators like cats and tarsiers, whose eyes and ears are especially exposed to injury by prey. The reflexes also are exaggerated in social species for use in communication; thus ear flattening in horse and dog displays has a counterpart in scalp retraction among Old World monkeys. The scalp movements and raised brows are effectively used in communication, despite the greatly reduced mobility of ears among monkeys. Limbs and other appendages (*e.g.*, antennae) are withdrawn or used to protect sensitive areas by both vertebrates and invertebrates. Among mollusks and such groups as sea squirts and barnacles, the whole soft body can be retracted into a protective shell, or carapace; a kind of door (operculum) may be used to stop the entrance (*e.g.*, among snails and barnacles), and trap-door spiders pull the stopper in place behind them. Bone may have evolved in fossil vertebrates as protective armour in jawless ancient fishes (ostracoderms), probably as a result of natural selection in the face of dominant arthropod predators (eurypterids). With the evolution of jaws, the vertebrates themselves gave rise to nearly all later large predators. Evolutionary advantage then apparently came with complex sense organs and behaviour; in most vertebrate lines there is evidence of a progressive reduction in body armour (dermal bone). Thus, although such cartilaginous fishes as sharks and rays do not exhibit such bony skins, they may well have evolved from heavily armoured ancient fishes (placoderms).

The completely armoured turtles and tortoises

Armour nevertheless has evolved repeatedly, particularly among animals incapable of fast locomotion; trunkfish (boxfish), for example, have a body entirely boxed by bony plates; and tortoises and turtles are perhaps the most completely armoured of four-legged animals. The turtles seem to have evolved early from the basal stock of the reptiles; thanks to the shell (carapace) within which they can withdraw head, limbs, and tail, they represent one of the few reptilian orders that have remained consistently successful. The turtle's dorsal carapace appears to consist of newly evolved plates of dermal bones, but the belly plates (ventral plastron) may well be retained in part from fish ancestors. Reptilian land vegetarians usually tended to evolve armour, as in the fossil dinosaurs such as stegosaurs and ankylosaurs.

South American toothless animals (edentates) such as anteaters are probably survivors of a comparable early development in mammals. The armour of armadillos and the presence of bony plates in the skin of the extinct sloths suggest that the whole group may derive from an armoured ancestor. The appearance of hair in the mammal line, partly inferred from the presence of fleas in early evolutionary periods, seems to have led to the evolution of a light, spiny type of armour. Such modern mammals as hedgehogs, echidnas, insect-eating tenrecs, and some rodents and their relatives (lagomorphs) all possess defensive spines that are commonly erectile and are often able to roll into a ball like an armadillo.

Chemical means of defense may be widely distributed in the body, making the animal distasteful to predators. Some of these chemical compounds may be derived from plants eaten or synthesized by the animal itself (*e.g.*, bufotoxin in toads). Although the animal attacked may be killed and thus not benefit, his fellows do since they are likely to be avoided by the predator. Poison or distasteful substances may also be ejected from a bodily reservoir and squirted at the enemy (*e.g.*, the skunk, some ants) or inserted into a puncture made by a spine (*e.g.*, triggerfish) or teeth (*e.g.*, certain snakes). Many poisons act to paralyze muscles by blocking nerve transmission at the neuromuscular junction (*e.g.*, cobra venom).

Warning behaviour. Mobbing behaviour apparently advertises the presence of a predator that is potentially but not immediately dangerous; thus mammalian nest predators can be safely mobbed by flying birds, as can owls in the daytime. From the safety of trees, such mammals as monkeys and squirrels mob predators on the ground. Mobbing calls are typically easy to locate, the calls being short and staccato, and they provide excellent cues of distance and direction. Conspicuous movements, such as tail flicks among small birds and squirrels, accompany the calls.

More urgently, intense warning behaviour is given in response to sources of immediate danger (*e.g.*, hawks actively hunting). Under these circumstances, warning calls are usually long whistles that make location difficult because of their gradual onset and termination and their narrow ranges of pitch. The evolution of warning behaviour that puts the displaying animal in danger (such as these intense warning calls) seems likely to come about only if the benefit to offspring and other members of the species is great. Indeed, it has been calculated that if an individual loses his life as the result of his warning behaviour, increased transmission of his family's genes will result only if the reproductive rate of relatives (as close as sisters) is doubled as a result of his sacrifice. (R.J.A.)

Aggressive behaviour

The term aggressive behaviour is used in so many different ways that no single definition can possibly cover all of its meanings. Behaviour that serves to injure an opponent or a prey animal, or to cause an opponent to retreat, is usually considered aggressive. When considering human aggression, some psychiatrists consider any act that has destructive consequences (including suicide) to be aggressive. (For a discussion of aggressive behaviour in man, see EMOTION AND MOTIVATION.) The role of aggression in behaviour has been—and continues to be—debated by psychologists and ethologists, as does the meaning of the term itself. Frequently, aggressive behaviour encompasses both attack and defense. Other investigators exclude food-gathering behaviour, though it may involve attack on another animal. In order to avoid these ambiguities of definition, a distinction must be made between causation, function, and description of observed behaviour. It is frequently assumed that a single motivational system (aggression) causes all recognizably aggressive behaviour in higher animals. This assumption is certainly invalid for invertebrates and for most higher vertebrates, in which a variety of motivational bases appear to exist. A motivational definition of aggression is thus difficult. The only possible rigorous approach is to list patterns of behaviour, usually held on both functional and causative grounds to be aggressive. Because aggressive behaviour has been most studied in mammals, mammalian behaviour will be examined here first as a basis for comparison with other animals.

BASIC AGGRESSIVE PATTERNS

Most modern theories of aggressive behaviour are based on laboratory experimentation, especially with mice and cats, and on field observations. Mammals show two basic aggressive patterns: attack and defensive threat. Direct evidence from brain-stimulation experiments in marsupials and eutherian (placental) mammals shows that each pattern is a closely coordinated series of events, obtainable in relatively constant form from a wide range of brain sites. Attack behaviour results from stimulating the lateral and anterior regions of the hypothalamus, and defensive threat from some areas near the centre of the hypothalamus and from the central gray region of the mesencephalon (midbrain). In the opossum, attack involves visual fixation on the prey, seizing it with the jaws, and shaking it from side to side through a figure eight. In defensive threat, the jaws

Mammalian attack behaviour

are opened and the animal backs into a corner, hissing. Newborn mice exhibit escape behaviour. Helpless at birth, a mouse will respond defensively to an outside stimulus (*i.e.,* pinching the tail) by squeaking, moving its legs, and moving away. When the mouse is older, its behaviour becomes more offensive in that it will attempt to bite. Still later, at about 12 days of age, it will stand on its hind legs in a posture of defense. Not until it is a month old, however, will it attack. The pattern of fighting in male mice includes kicking, biting, and rolling over until one of the combatants is hurt. The injured mouse might attempt escape, with the attacker in pursuit. If escape is not attempted or is not possible, the injured mouse may show defensive behaviour or feign death, submitting to the attacks of the aggressor. The aggressive patterns in the cat have been studied extensively. There is disagreement about the relation in this species between movements of prey catching and those of attack on other cats, and indeed whether prey catching should be termed aggressive at all. Since the coordination of stalking, biting, seizing with the forepaws, and scratching with the hind legs can be the same in both situations—and the eliciting stimuli also are often similar—it is reasonable to classify them together. The dispute stems from the possibility of the same pattern having two or more rather different causations.

In the cat, it is clear that defensive threat involves a number of component patterns that can occur independently of each other. The patterns include reflexes that protect vulnerable areas, such as narrowing the eyes or flattening the ears and cardiovascular preparations for exertion (*e.g.,* a rise in arterial blood pressure and active dilation of the blood vessels that supply muscles). Hisses and growls, together with erection of back and tail hair, are also typically part of defensive threat, serving to make the animal conspicuous.

Similar patterns can be recognized in birds and can be induced by brain stimulation in roughly the same areas as in the cat. In the collared dove (*Streptopelia decaocto*) attack involves pecks and wing blows, whereas an immobile threat display includes bill opening and feather erection.

Human attack behaviour, defined in this way, might consist of a direct stare, exaggerated by a slight frown, as preparatory behaviour, and then, at full intensity, striking with hand or fist, seizing, and biting. Staccato, loud, or breathy speech, given at such times, may be compared to the violent expiration (hissing, growling) of other primates in threat.

Acquired skilled movements (*e.g.,* boxing, wrestling) could reasonably be included, just as fighting in adults of other species probably also involves skilled movements (*e.g.,* blows with leg spurs in cocks).

CAUSATION

Evocative stimuli. Few stimuli that are highly specific for attack have been described. A scent produced by adult male mice is probably one example. Another is colouring. It has been shown that chaffinches (*Fringilla coelebs*) have a personal territory about them that they protect from intrusion by other chaffinches. This territory may be related to a source of food or water or to space for perching. Invasion of the personal space of one chaffinch by another usually results in either attack or withdrawal. Males may penetrate up to 20 centimetres (eight inches) before either alternative occurs. Females, however, are permitted to penetrate up to 10 centimetres. Experiments have been performed in which the feathers of the female chaffinch's breast are dyed to resemble those of the male. When such a disguised female penetrates the personal territory of a male, attack or withdrawal occurs within 20 centimetres. Experiments with inexperienced birds indicate that the effect is innate.

Members of an animal's own species (conspecifics) that are unfamiliar tend to evoke attack, perhaps partly because they lack the cues that suppress attack in familiar fellows (*e.g.,* group smell in mice; see below *Aggression as a drive*). Another very widely effective stimulus is that of a conspecific approaching within a certain distance, which, in the case of species the members of which ordinarily do not tolerate close approach, may be very uniform; for example, the chaffinches' inviolate personal territory, which has already been mentioned above. This distance may be maintained around objects other than the animal itself. A typical songbird territory includes points, such as song perches or the nest site, that are defended in this way. Approach to the mate often evokes attack; this is particularly clear in species in which paired birds remain in flocks (*e.g.,* some grackles, *Quiscalus*) and approach to the mate is not the same as entry of the territory. Both birds and mammals commonly attack in response to an approach that threatens their offspring. In many ways (*e.g.,* grooming) offspring are treated as an extension of the maternal body, thus, a threat to the offspring is a threat to the parent.

Stimulus of a fleeing animal

A conspecific moving rapidly away tends to evoke chasing and attack in many birds and mammals. In such predators as dogs attacks like this may be partly inhibited, taking the form of aggressive play, but their resemblance to prey catching is so close that any distinction would be artificial. The stimulus is equally effective, however, in such species as baboons, which rarely or never take large prey. For young domestic fowl, inexperienced in attack, repeated rapid retreats and approaches by another individual are the optimal stimuli for sparring and pecking.

Submissive behaviour (*e.g.,* exposing throat or belly) has been described as providing stimuli that specifically inhibit attack. It is more likely that such behaviour is effective either by removing stimuli that tend to evoke attack, or by evoking systems of responses incompatible with attack. Examples of the first category are immobility, as in the resting attitudes assumed by many mammals and birds when persecuted, and the avoidance of behaviour (such as the direct stare) suggestive of intent to attack. The assumption of infantile behaviour (*e.g.,* food begging in many birds) makes attack less likely. Many social mammals, indeed, have evolved special features in their young (*e.g.,* juvenile fur colour of baboons and guenon monkeys) that allow their possessor to behave freely in ways that would certainly produce attack if shown by an adult. The assumption of sexual postures is another example of the second category of submissive behaviour. In baboons and macaque monkeys "presenting" (the type of solicitation used by a sexually receptive female) has become a common response of any subordinate, male or female, when approached by a superior. The Austrian ethologist Konrad Lorenz has argued that predators have unusually effective submissive behaviour because they are so well armed and, extrapolating from animals to man, that it is therefore unfortunate that man is not more genuinely a carnivore. In fact, as with most displays, the clearest correlation is that the more social the species, the more elaborate and easily elicited the communicating displays. Man is even better equipped than other primates with infantile displays (weeping, dependent behaviour) and friendly greetings (smiles) that avert attack.

Dominance hierarchies

Dominance hierarchies, first described in terms of pecking right in the domestic fowl, may be regarded as a means of reducing attack within a social group. Once the hierarchy is established, disputes rarely need arise because the inferior immediately gives way when confronted by a superior.

A dominant animal is best defined as one whose actions are not constrained by possible responses of its fellows. Its position need not derive from earlier successful aggressive behaviour although it usually does, but may instead reflect superior persistence or strength. Although best known in birds and mammals, dominance hierarchies have also been described in crabs and bumblebees. In bumblebees the dominant animal opens its mandibles and butts the inferior one.

Facilitating stimuli. Startling and painful stimuli facilitate attack, which may be directed to objects not associated with the source of pain or startle. The effect has been shown in both rats and squirrel monkeys, following electric shock administered through a floor grid. A clear distinction has usually not been made between attack, defensive threat, and a mixture of the two; but both patterns are clearly facilitated. The human irritability that often accompanies persistent discomfort or pain provides a more familiar example.

The thesis that frustration causes aggression has been very influential since it was first proposed. In a laboratory experiment, a pigeon immediately attacked another whenever a lever that normally delivered food when pressed failed to do so. There is no reason, however, to elaborate such evidence into the proposition that all of the attack is caused by frustration, as has been suggested by some investigators.

Reinforcing patterns. It has been suggested that a number of important patterns of behaviour—most or perhaps all of which have major representation in lateral hypothalamic brain centres—include component responses that produce self-reinforcing stimuli. An animal in which the pattern is activated, either naturally or by hypothalamic stimulation, will learn to run a maze in order to be able to perform the response. This observation has been made of the rat, not only for such obvious patterns as feeding and drinking but also for gnawing on wood and other objects. Following appropriate electrical stimulation, a cat will run a maze in order to be able to attack a rat, even though most cats normally do not respond to rats. This behaviour involves prey catching, rather than attack, if a distinction is to be drawn. There appears to be no direct evidence of this type for attack in animals not artificially stimulated.

Play as aggressive behaviour

It is clear, observationally, that once attack has become likely an animal will search for the opponent and perform a variety of responses in order to reach him. The study of play behaviour suggests that the movements of prey catching (*e.g.*, kittens pouncing on a ball) or attack (*e.g.*, the play fighting so common in primates and other mammals) in themselves can be reinforcing. In such play, hurting or inducing signs of hurt are avoided, but it is possible that these sometimes are sought and serve as reinforcers, as is certainly suggested by human behaviour.

It is almost certain, however, that attack and defensive threat are also performed under the control of quite other reinforcers, such as removing a social fellow who is an inconvenience or an obstacle. It has been shown that a pigeon can be conditioned to attack a cage mate as an operant response to obtain a food reward from a hopper quite remote from the animal attacked. Full intensity attack was given, but other experimenters have found that attack caused in this way shows the simpification and reduction in intensity typical of operant responses. This reduction is obvious in the light peck used by dominant hens to make inferiors retreat from food or to control them in other ways. There is no reason why attack should be used for such a purpose rather than some other behaviour, if some other type of behaviour was more convenient.

Aggression as a drive. Lorenz is the most recent author to treat aggression as a drive that energizes attack (and some other behaviour), is exhausted by performance of the behaviour, and then accumulates afresh, making performance more and more likely. The resemblance to the death wish of Freudian theory is obvious, particularly since both suggest that quite unlikely behaviour may be motivated by the drive. This approach to the causation of behaviour can be traced back through the writing of Charles Darwin to an influential essay by Herbert Spencer on laughter and its role in discharging accumulated nervous energy. No convincing proof for such theories has ever been presented.

There is now direct evidence that threshold (the amount of stimulus required to elicit the response) for attack or threat may be raised or lowered over periods as long as 20 minutes. In the ring dove, following the evocation of either response by a natural stimulus or by brain stimulation, some brain sites that normally give the response show a changed electrical threshold for such a period. In many species, once a stimulus capable of evoking attack has been presented, the individual may remain likely to attack for some time, especially when the attack is evoked by an animal itself too dominant to be attacked. Other conspecifics or, in man at least, inanimate objects may then be attacked instead, a phenomenon called redirection. What evidence there is, however, suggests that such states of readiness, which in man would be termed irritability or anger, will dissipate just as well or better in the absence of actual performance of aggressive behaviour.

There is no reliable evidence that aggressive behaviour becomes progressively more likely over long periods in which it is not evoked, as has been suggested by some workers. Such an increased likelihood is clearly found for feeding behaviour, but the consequences of not eating—such as reduced availability of glucose and free fatty acids to tissues and a reduction of fat stores—provide internal stimuli that make feeding more likely. These effects, of necessity, become more intense the longer a fast is prolonged. In the case of performance of copulation, the act is followed by a reduced likelihood of response to further stimuli. In the male of certain species (domestic cattle, for example) this reduction is almost entirely restricted to copulation with the female just mounted, the bull remaining receptive to stimulation by other cows. In other species there is some contribution from a general rise in the threshold for copulation. There is no evidence for either type of effect in the case of aggressive behaviour, except that isolation promotes fighting. This is well established for male laboratory mice. Attacks on intruders, however, depend largely on odour cues (*e.g.*, male cues increase, female cues decrease, attack). Fighting within a social group of males is much reduced by group cues, and isolation may act to abolish the effectiveness of such cues.

Fighting in laboratory mice

The increased incidence of aggressive behaviour in small groups of isolated men, such as polar explorers or prisoners, has been explained by Lorenz as resulting from a build-up of undischarged aggression. But the comparison is not strictly relevant because the stress and frustration of such situations might well facilitate aggressive behaviour directly.

Physiological effects on aggressive behaviour. It has long been known that the administration of male sex hormones (androgens) makes aggressive behaviour more likely in many male mammals and birds. It has no effect on aggressive behaviour in female rats and mice, apparently because of the general insensitivity of the female central nervous system to androgens. Male individuals castrated at birth are less aggressive than those castrated at the time of weaning, which suggests that masculine aggressiveness is partly determined by early differentiation. Androgens also affect male aggressive behaviour indirectly. In the red deer, testosterone implants cause migration to the rutting (mating) area, which will be defended, and also cause the loss of velvet from the antlers, so that these become effective weapons. The increase in aggressive behaviour caused by the possession of a territory is particularly obvious in passerine birds, the males of which may continue to flock freely with other males on feeding grounds at a time when, if on their territories, they would attack other males fiercely. Hormonal effects on aggressive behaviour also occur in females. The increased aggressivity shown by female hamsters at pregnancy is the result of rising progesterone levels.

Role of hormones in aggression

Lesions of the amygdala, a structure near the base of the forebrain of great importance in motivation, have been long known to produce taming in a wide variety of mammals and in ducks, and the effective area has been localized to the rostral lateral amygdala in cats. Defensive threat, attack, and fleeing are reduced as a result. The amygdala is not exclusively concerned with aggressiveness; it appears to be especially important in cases in which the novelty, intensity, or past significance of a stimulus determines whether the source is approached and eaten or avoided. The amygdala may facilitate aggressive behaviour as part of a determination that a stimulus is intense or unusual.

Lesions of the septum, a part of the base of the forebrain containing important tracts connecting to the rest of the brain, have the opposite effect: in the laboratory rat, such lesions cause defensive attack to be given to such stimuli as handling or a light (but presumably startling) touch, which were previously ineffective. The septum is also implicated in the inhibition of certain responses (*e.g.*, those that have ceased to be rewarded), and this may be related to its role in suppressing attack.

Threat displays. The main modern work on displays has explained their causation in terms of conflict between such major drives as aggression, fear, and sex. There is considerable evidence, especially from experimental brain

stimulation in the cat, that defensive threat is a unitary pattern caused by the same stimuli that evoke attack or fleeing but independent of the latter. Confident threat displays are given by animals unconstrained by the possibility of retaliation, which do not attack because, for example, they do not wish to move. These displays include a wide variety of components. A direct gaze, which may be exaggerated by a frown in mammals with mobile faces, indicates a sustained interest. Protective reflexes indicate that the animal showing them anticipates close and possibly dangerous contact. They are assumed in rough, friendly play; but when they appear during approach—when still at a little distance—they suggest the expectation of attack (*e.g.*, ear flattening in the horse); the greater the distance at which they appear, the less confident the animal showing them. Finally, the animal may perform actual intention movements of attack, such as mouth opening, withdrawing the lips, or striking at the ground.

Many displays, termed threat displays because they are given to strangers at the territorial boundary or during early phases of pair formation when attack is still likely (*e.g.*, long call or choking display of gulls), would be best given a more neutral name.

EVOLUTION OF AGGRESSIVE BEHAVIOUR

It is instructive to attempt to identify attack in the familiar frogs (*Rana*) and toads (*Bufo*) of the amphibian order Anura. Prey catching in most frogs is highly specialized, involving a lunge of the head or a movement of the tongue only and directed only at small moving prey. In this respect, it is quite unlike attack shown by most land vertebrates. The mounting and clasp of females are so obviously related to copulation that they, too, are not readily confused with attack. Males clasped in error use wiping movements of hind legs that also are evoked by dirt on the body surface, and they struggle as they would if caught in some obstruction. There is thus no obvious attack response. Some of the functions of attack, however, are achieved by displays that are not derived from aggressive behaviour: females that have spawned, and all males, when clasped, give a special call that tends to produce unclasping. In higher tetrapods (land vertebrates), such unwanted copulation attempts would certainly evoke and be terminated by attack. In some anuran species, the calls given by males to attract females elicit avoidance by other males, ensuring optimum spacing of breeding males.

Relation of aggression to mating

Iguanid lizards resemble such frogs in possessing a female rejection display (legs extended, back arched, and tail turned toward the male), which appears to be specific in effect. Males give what appears to be an aggressive display to other males from lookout sites, but there is no evidence that it has evolved from movements of attack; it consists of vertical head bobbing that may well be an exaggeration of movements used in depth perception. The movements precede copulation as well as attack, which in lizards involves bites and sometimes tail blows, and the two may have evolved quite separately. The bites presumably derive originally from prey-catching behaviour, which is clearly as ancient in the vertebrates as the evolution of the jaw. As soon as dangerous bites became possible for feeding in vertebrate evolution, there was strong selection for their use in defense, and the facilitation of attack by pain or startle may be as ancient as this selection. The use of prey-catching behaviour to drive off conspecifics from areas offering important resources (*e.g.*, food, mate, nest) would also be strongly selected, for, once the pattern was available (see below), it would require initially only a slight change in evoking stimuli (*e.g.*, response to rather larger moving objects) to allow this result. The crossopterygian ancestors of the tetrapods and their immediate amphibian descendants were dominant predators, so that prey catching and biting in defense can be assumed to have been present. Comparison with present-day inhabitants of muddy shores, both air-breathing fish (*e.g.*, mud skipper) and crabs, suggests that conspicuous displays given to the approach of a conspecific had probably also already evolved. Causally then, as now, some or all of such displays might have been independent of attack. No doubt they already could be followed either by copulation attempts or by attack. There would be strong selection for appropriate hormones to confine the displays to breeding periods; this selection has happened with frog breeding calls and with bird songs, as well as with bird territorial displays, all of which are now facilitated by testosterone. The fact that this is true also of attack and male copulation has been taken mistakenly in the past to indicate causation of the former categories of bird behaviour by aggression and sex.

Prey-catching movements in displays

Aggressive behaviour of invertebrates. An account of invertebrate aggressive behaviour must be based on an analogy with the functions of that of vertebrates, since there can be no question of homology (common origin). Obviously, even in vertebrates, many of these functions are also served by other behaviour as well.

Aggressive behaviour may allow a vertebrate to maintain, exclusively if possible, its access to some important resource such as food, territory, or mate. In some invertebrates, there is the same use of attack derived from prey-catching responses, as in higher vertebrates. The larger dragonflies (Odonata) maintain a feeding territory by hawking at conspecifics in the same way as at prey. The clearest example is that of nest defense in social wasps and bees. Here a sting and a secretion, originally evolved to paralyze insect prey, have come to be used both against conspecifics (*e.g.*, when fighting with "robber" bees attempting to enter the hive) and against vertebrates, which raid bee and wasp nests. Bees are not predatory, so their sting has no food-capturing function. The scent of bee venom strongly facilitates the stinging of large, hairy objects, and a whole hive may pass into a state entirely analogous with that of heightened aggressivity in a vertebrate but with the persistence of lowered threshold (*i.e.*, greater readiness to attack) dependent on the persistence of a chemical signal (pheromone). Both termites and ants have evolved facilitation of colony defense by pheromones released by disturbed individuals, like that described for bees.

Nest defense by social insects

In their attack behaviour, ants and bees show an interesting use of responses evolved to keep the nest clean of dead conspecifics and other debris. Drones (males) are seized and thrown out of the hive entrance. Strangers are treated similarly if they remain passive. If the stranger is identified at a distance (*e.g.*, by the side-to-side flight characteristic of robber bees), the guard bees may open their mandibles in a manner reminiscent of vertebrate threat, although there is no clear evidence that it is effective as a signal.

Freely ranging spiders, such as wolf spiders (Lycosidae), respond to conspecifics as they do to prey; the animal attacked commonly gives a low-intensity version of a display shown at full intensity in courtship, which tends to terminate the attack. Perhaps the least advanced group in which behaviour analogous with attack has been described is that of polychaete worms. Predatory ragworms defend their burrows against conspecifics with biting attacks. In other instances, biting is the analogue of vertebrate defensive threat, which has evolved from prey-catching behaviour; most threatened crabs extend and open their chelipeds (large claws), which evolved for feeding.

The function of defending resources against conspecifics is also served by a variety of adaptations that should not be termed aggressive (*e.g.*, Orthoptera stridulation).

Defensive displays. In addition to territorial defense, aggressive behaviour may function to terminate some behaviour of a fellow animal. The crying of human infants effectively manipulates mothers to remove sources of discomfort or fear; it is misleading to term it aggressive, however. In the chimpanzee, screaming tantrums are known to allow infants to win disagreements with the mother and access to food. Tantrums are also used by adults to some extent. In many primates (*e.g.*, baboons), such behaviour shades into the screeches of defensive threat, but it is probably best separated from aggressive behaviour proper. That the distinction is needed is clear for an analogous series of behaviour patterns, the distress calls of young gallinaceous birds (pheasants, quail, and domestic fowl), which are given when there is marked discrepancy between the stimulation expected or sought after and that obtained. Aggressive behaviour can thus be regarded as a source of only one of a range of reinforcers used by higher vertebrates to punish and manipulate their fellows.

Tantrums

Defense against predation can depend on many adaptations other than attack and defensive threat. Those that involve the infliction of damage, pain, or startle by a response aimed at the predator may be considered here as analogous to aggressive behaviour. A number of mustelids, especially skunks, have evolved a new pattern of defensive threat in which the readiness of the scent glands to discharge is conspicuously advertised, in one species by standing on forepaws to present glands. Various petrels (seabirds of the families Hydrobatidae and Procellariidae) discharge the oily and repellent contents of their crops over substantial distances at nest predators; the accuracy of aim may depend on the fact that they face the predator in order to examine him binocularly.

The loud call given by most tetrapods, including frogs and toads, when seized, apparently serves to startle the predator so that a final opportunity for escape may be given. Such a call may well have been involved in the evolution of the screeches given by many mammals in defensive threat. Many invertebrates have specialized startling displays that are comparable: the sudden display of eyespots by many moths when disturbed by a sudden movement or tap, the forewings typically moving forward to expose markings on the hind wing. (R.J.A.)

Migratory behaviour

Migration can be contrasted with emigration, which involves a change in location not necessarily followed by a return journey; invasion or interruption, both of which involve the appearance and subsequent disappearance of great numbers of animals at irregular times and locations; and range expansion, which tends to enlarge the distribution of a species, particularly its breeding area.

The migration cycle is often annual and thus closely linked with the cyclic pattern of the seasons. The migration of most birds and mammals and many of the fishes are on a yearly cycle. In many cases (*e.g.*, salmon and eels) animals with a relatively long life-span return to the place of birth in order to reproduce and eventually die. In other cases, as in certain invertebrates, where the animal has a relatively brief life-span and reproduces rapidly, migrations may not occur in every generation. The daily movements of certain fishes and invertebrates have also been called migrations because of their regular occurrence. This type of movement, however, is not to be confused with migration in the strict sense.

Most migrations involve horizontal travel. The distance traversed may be a few miles or several thousands of miles.

Some migrations take a vertical direction and involve no appreciable horizontal movement. Certain aquatic animals, for example, move from deep water to the surface according to the season. Certain birds, mammals, and insects migrate altitudinally in mountainous areas, going from the upper zones, where they breed, to the foothills or plains during seasons when the weather is severe and unfavourable. Such vertical travels involve essentially the same type of environmental change as horizontal, or latitudinal, migrations over long distances.

SURVEY OF MIGRATORY BEHAVIOUR IN ANIMALS

Planktonic migration

Lower invertebrates. Many marine invertebrates travel considerable distances during certain seasons. A large proportion of them, however—particularly planktonic organisms, plant and animal aquatic drifters—do not travel deliberately but are carried by ocean currents. Planktonic organisms also travel vertically in a daily rhythm. Very small or microscopic animals remain at great depths during the day and rise at dusk, concentrating in the upper layers of water during the night. Their predators, particularly fishes, follow them in their cycle. The daily activity of pelagic birds (*i.e.*, those living on the open sea rather than along the shore), such as petrels and shearwaters, which feed on planktonic crustaceans and squids, follows this same rhythm.

A seasonal change of habitat, analogous to migration, is made by some Polychaeta (sandworms). Along the coast of Europe, clam worms (*Nereis*) live during the colder months in rock crevices and among algae. During the summer, however, they become planktonic and swim out some distance from the coast, where reproduction occurs. In the South Pacific, near Samoa and Fiji, the palolo worm (*Palola siciliensis*) lives among coral reefs, where it develops posterior segments filled with genital (reproductive) cells. These are cast off, and the worm rises to the surface. The phenomenon occurs regularly on the first day of the last quarter of the October–November moon.

Some of the best known migrations among the invertebrates occur in crustaceans during the reproductive period, when some of them travel as far as 240 kilometres (150 miles). Generally in the crabs, females move into shallow coastal waters to mate and to lay their eggs. After the eggs have been laid, the females return to deep water.

Some freshwater crabs, such as the Chinese crab (*Eriocheir sinensis*), after remaining for three to five years in freshwater, migrate to brackish water, where mating occurs. Females with eggs externally attached then travel to the sea and remain a few miles offshore for several months during winter. The following spring they enter shallower water near the shore. Here the eggs hatch. Young crabs spend a year in brackish water and migrate upstream the following spring, settling in freshwater and growing to maturity.

Some crabs, such as robber crabs (*Birgus*) and land crabs of tropical regions (*Geocarcinus*), have adapted to life on land. They migrate to the sea to reproduce and then return inland and are followed at a later time by the young.

Insects. Migration among the insects is best known in locusts and butterflies; a great number of other insects, however, including some of the smallest, are migrants. Broadly speaking, insect migration is of three types—some insects emigrate on one-way journeys to breed; others migrate from a breeding area to a feeding area; still others migrate from breeding areas to hibernation sites.

Types of insect migration

In the first type, adults with a life-span limited to a single season emigrate from their breeding site, deposit their eggs, and die. Such migratory flights can be very short or very long but, because they are always one-way journeys, cannot be regarded as migration in its strictest sense. The best known examples of such flights are those of locusts, particularly the desert locust (*Schistocerca gregaria*), a species found in tropical and subtropical countries of the Old World. The migratory, gregarious form arises from the solitary form as a result of various conditions—*e.g.*, lack of food, crowding.

The desert locusts breed only when and where seasonal rains permit; as a consequence of climatic conditions, therefore, the insects migrate from one breeding area to another. If the available food decreases and the numbers of insects increase too drastically in a particular area, migratory locusts develop. They differ from nonmigratory forms in colour, structure, behaviour, and physiology. Swarms numbering up to 10,000,000,000 individuals periodically invade territories in Africa, southwestern Asia, and southern Europe, covering areas as large as 1,000 square kilometres (400 square miles).

Other migrant insects travel beyond the limits of their breeding range and either die or return to the breeding range. The painted lady butterfly (*Vanessa cardui*) "migrates" in the spring, when its population becomes too large for local conditions, from the peninsula of Lower California in Mexico to the Mojave Desert in Southern California. Eggs are laid in the desert region, but the species does not become permanently established there and makes no return flight. Such movements are known in about 250 species of butterflies.

In the second type of migration—migration in the strict sense—insects migrate from a breeding area to a feeding area. In the feeding area the females develop mature ovaries and then return to lay their eggs in the place from which they came or a similar region.

Cockchafers (*Melolontha melolontha*), a species of beetle, leave the site where they emerge as adults and move to a feeding area, generally in a forested region, where maturation of the eggs takes place. They then return to the area where they developed from eggs and lay their own eggs. This process may be repeated several times during the life of the insect. Although the distances covered by

the cockchafer may not be great, the regularity of the phenomenon is characteristic of true migration.

In the third type of migration, insects travel from their breeding areas to places where they hibernate or estivate—*i.e.*, pass the summer in a dormant state. The place of hibernation or estivation may be outside the zone where climate permits breeding. The following season, they return to the breeding place and lay their eggs. This type of migration, which can involve great distances, is made by insects with unusually long life-spans. The lives of these insects include a diapause, or period of dormancy during which development is suspended.

In warm countries the coccinellids, a family of beetles, and certain moths leave the hot lowlands and migrate to the mountains, where they estivate and later hibernate. In spring they return to the breeding areas.

Migration of the convergent ladybug

One coccinellid, the convergent ladybug (*Hippodamia convergens*), lives in valley regions of California, where the eggs hatch in March or April and develop into adults one month later. In early summer they migrate to the mountains, particularly to the Sierra Nevada, where they may lay eggs if food is abundant and the weather warm. Generally, however, the adults gather in clusters and remain inactive until October, when rains initiate a period of activity, after which they travel to lower altitudes and hide in forest litter, passing the winter in a state of dormancy. As many as 30,000,000 ladybugs may congregate on a quarter acre. In spring they mate, fly back to the valleys, lay their eggs, and die.

The flight before diapause of some insect groups may cover thousands of miles. In North America, the monarch butterfly (*Danaus plexippus*) is a well-known example of a wide-range migrant with an extensive breeding range. The number of generations varies with the latitude; as many as five generations may occur each year in the south and only one in the north. In summer the insects travel northward to Hudson Bay. Individuals of the last generation of the year migrate southward in autumn to Florida, Texas, and California, where they hibernate after travelling nearly 3,200 kilometres (2,000 miles). They gather in sheltered sites, particularly on trees where they cluster on trunks and big branches. In spring part of the populations migrate back to the northern breeding areas. Some of the returning butterflies are members of the first generation that develops from the overwintered insects; others represent successive generations that develop as the insects progress toward more northern latitudes. The recapture of marked butterflies has revealed that they travel as much as 130 kilometres (80 miles) in one day. The longest distance recorded thus far for the complete flight of a migrant monarch butterfly is 3,010 kilometres (1,870 miles).

From M. Ricard, *The Mystery of Animal Migration*

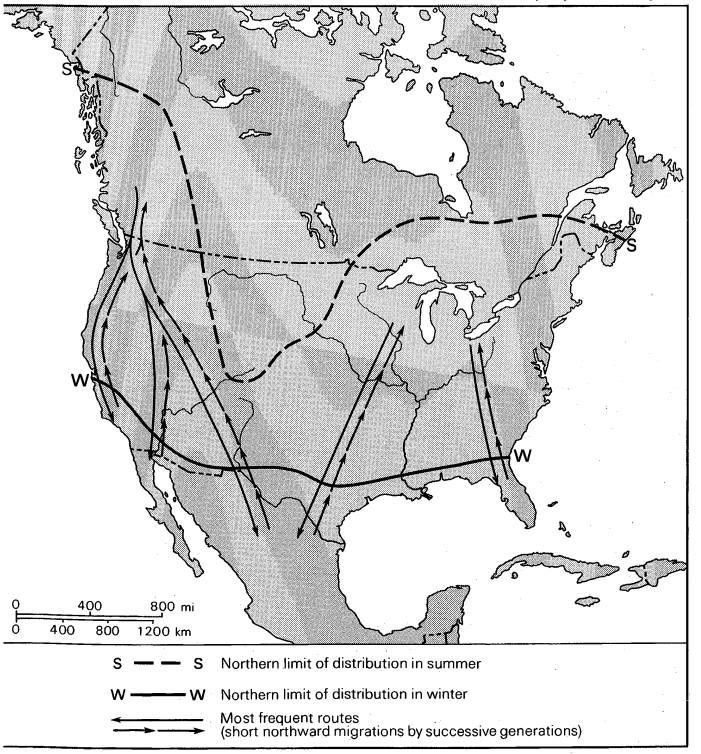


Figure 19: Migrations of the North American monarch butterfly (*Danaus plexippus*).

Fish. Many species of fish wander annually through a particular area of the ocean. Some are true migrants, travelling regularly over great distances. Young fish usually leave the spawning grounds for areas where they develop into juveniles, before joining the adult stock at the feeding grounds. Adults move to the spawning grounds, then return to the feeding grounds. Migratory patterns of fish are related to oceanographic factors and to currents. Eggs, larvae, and young drift passively with the current, although migration of adult fish toward breeding grounds is usually against the current. Adult movements thus are directional rather than passive, and the fish respond to environmental conditions—*e.g.*, climate.

Three categories of migratory fishes can be distinguished: oceanodromous, anadromous, and catadromous.

Oceanodromous fish. Oceanodromous fish, which occur widely throughout the world's oceans, live and migrate wholly in the sea. They differ mainly from one another by the method and extent of their migration.

Herring migratory patterns

Herring (*Clupea harengus*), extensively studied because of their economic importance, are the best known of the oceanodromous type and can be classified into several populations, or local races, which do not mix freely. In addition, each has a particular migratory behaviour. In the North Sea, distinct groups spawn in different seasons and on different grounds: Buchan herring spawn in August and September off the coast of Scotland and migrate to the coast of southwestern Norway; Dogger Bank herring spawn in September and October in the central part of the North Sea and along the English coast and then migrate to the Skagerrak, an arm of the North Sea between Denmark and Norway; Downs herring spawn from November to January off the French coast, mainly between Dunkirk and Fécamp, then feed in summer in the middle and northern parts of the North Sea, sharing the feeding grounds with other populations. The diversity of migration and of reproductive seasons is closely connected with the annual cycle of oceanographic conditions in the North Sea.

Cod (*Gadus morhua*) have migration patterns similar to those of herring. The migrations of other fish cover even greater distances; in the Atlantic, for example, white tuna (*Germo alalunga*) are found in winter around the Azores and the Canary Islands, where they spawn in spring. They then migrate northward to the Gulf of Gascogne and afterward to the waters around Iceland, arriving there in July. Populations of red tuna (*Thunnus thynnus*) occur throughout the Mediterranean Sea and the eastern Atlantic. In May and June they spawn in the western Mediterranean. During summer they spread northward to feed, finally reaching the Arctic Ocean. Similar migrations occur along the North American coast in the Atlantic and throughout the Pacific.

Anadromous fish. Anadromous fish live in the sea and migrate to freshwater to breed. Their adaptations to conditions of different habitats are precise, particularly with regard to salinity of the water.

Salmon (*Salmo, Oncorhynchus*) spawn in the cold, clear waters of lakes or upper streams. Eggs are laid in gravel beds. The young of the Atlantic salmon remain in freshwater for two to three years, sometimes as long as six; Pacific salmon sometimes migrate to the sea in their first year. Adult fish usually remain in the sea for two or three winters, but sometimes only one. Then, as grilse (adolescents) or as adults, they return to freshwater and spawn, after changes occur in colour and other external features. Some Atlantic salmon die in freshwater after a single spawning; others return to the sea.

The tagging of salmon has shown that European types may cross from Norway to Scotland, as well as the reverse. Pacific salmon are probably distributed over the Pacific Ocean and Bering Sea, between latitudes 45° N and 65° N with surface waters of 2° to 11° C (36° to 52° F).

Experiments in Canada and the United States, in which young salmon migrating to the Pacific have been tagged, have shown that a high proportion of the fish return to the river where they hatched. Tagging of Atlantic salmon has shown that a few survivors have migrated two or even three times to a particular river in successive years. Adults reared from experimentally transplanted eggs return to the stream where they were hatched or grew, not to the stream where the eggs were laid. Aside from other means of orientation, such as reference to celestial features, topographical features are believed to play an important part in recognition of the original habitat. The sense of smell, or olfaction, however, has the most important role. Experiments have shown that migrating salmon are attracted to the waters of the stream in which they are going to spawn. Experiential imprinting at an early stage of development enables a grown fish to respond to waters that contain substances with a particular odour or that have a characteristic temperature.

Catadromous fish. Catadromous fish spend most of their lives in fresh water, then migrate to the sea to breed. This type is exemplified by eels of the genus *Anguilla*, numbering 16 species, the best known of which are the North American eel (*A. rostrata*) and the European eel (*A. anguilla*).

Eel migration

European eels and North American eels spawn in warm saline waters of the Atlantic, at depths of 400 to 700 metres (about 1,300 to 2,300 feet), in an area centred near latitude 26° N longitude 55° W called the Sargasso Sea. The pelagic eggs develop into leptocephali—transparent, leaflike forms with relatively small heads—that are carried by the Gulf Stream to the shallow waters of the continental shelves. When they are about two and one-half years old and about eight centimetres long (a little more than three inches), a metamorphosis occurs. The leptocephali are transformed into so-called elvers, which are bottom-dwelling, pigmented, and cylindrical in form. They arrive in coastal waters as glass eels and begin to swim upstream in freshwater streams in spring. Their migration upstream is spectacular, as the young fish gather by millions, forming a dense mass several miles long. In freshwater the eels grow to full size, becoming yellow eels. They live as such for 10 to 15 years before changing into silver eels, with enlarged eyes; they swim downstream to the sea, return to the spawning grounds (Sargasso Sea), and die.

The migration of these eels is not entirely understood, particularly their return to the Sargasso Sea. It may be that European eels and North American eels belong to the same species.

Reptiles and amphibians. The range of seasonal movements of most reptiles and amphibians is probably very limited. Generally incapable of travelling any great distance, they respond to unfavourable conditions by lapsing into a state of lethargy. This type of response makes it possible for them to remain in a particular area for the entire length of the year.

The only migration-like movements of reptiles and amphibians are made during the reproductive period. Frogs and toads then concentrate in particular areas such as ponds and lakes; thousands travel toward these sites from year to year. After reproduction, the animals disperse and again settle over their usual range.

Turtle migration

The South American river turtle, or arrau (*Podocnemis expansa*), migrates along rivers in large masses that may impede the passage of boats. The turtles gather on sandbars of large rivers to lay their eggs. In the Galápagos Islands, giant land tortoises (*Testudo elephantopus*) stay chiefly in the upper humid zone, where food is abundant, but go down to the dry zone to lay their eggs. Despite their great body weight and slow pace, they travel some 50 kilometres (30 miles) across rough country.

Sea turtles, on the other hand, migrate over long distances, lay their eggs on special beaches, and then disperse over a wide area. Green turtles (*Chelonia mydas*), which deposit their eggs on the coast of Costa Rica in Central America, disperse through the Gulf of Mexico and the West Indies. Green turtles that have been tagged on Ascension Island, halfway between Africa and South America, have been recovered on the coast of Brazil, 2,300 kilometres (1,400 miles) away.

Birds. Migration is most evident among birds. Most species, because of their high metabolic rate, require a rich, abundant supply of food at frequent intervals. Such a situation does not always prevail throughout the year in any given region. Birds have thus evolved a highly efficient means for travelling swiftly over long distances with great economy of energy.

The characteristics of migratory birds do not differ greatly from those of nonmigratory forms; many intermediate types exist between the two groups. All transitional forms, in fact, may be manifested in a single species or in a single local population, which is then said to undergo partial migration.

In addition to regular migration, nomadic flights may also occur. This phenomenon takes place, for example, among birds of the arid zones of Australia, where ducks, parrakeets, and seedeaters appear in a locality following infrequent and unpredictable rains, breed, and then move to other areas. Nomadism is a response to irregular ecological conditions.

In Europe. The populations of many northern and eastern European species of birds have pronounced migratory tendencies; the populations of western Europe, on the other hand, are more sedentary.

Some birds are nomadic in winter, others spend the colder months in the southwestern part of the continent or in the Mediterranean region. Many migrant populations migrate to Africa south of the Sahara. Geographical conditions determine several main routes. The Alps are an important barrier to migratory birds. About 150 species travel westward and southwestward; others travel southeastward.

Tits (*Parus*), goldfinches (*Carduelis carduelis*), and blackbirds (*Turdus merula*) are usually sedentary in western Europe; they are usually migratory, however, in northern Europe, where their flights resemble a short migration. Starlings (*Sturnus vulgaris*) are sedentary in western Europe, where large numbers gather from eastern Europe. Large flocks also pass the winter in North Africa.

Insectivorous (insect-eating) species, such as warblers, flycatchers, and wagtails, are highly migratory and spend the winter in the tropics, chiefly in Africa. They migrate to Sierra Leone on the west coast, Tanzania on the east coast, and all the way southward to the tip of the continent. Most of these migrants use different routes to cross the Mediterranean, chiefly in the western portion, although some migrate only southeastward. Golden orioles (*Oriolus oriolus*) and red-backed shrikes (*Lanius collurio*) go to East Africa by way of Greece and Egypt. Swallows—particularly barn swallows (*Hirundo rustica*) and house martins (*Delichon urbica*)—and swifts (*Apus apus*) pass the winter in Africa south of 20° N latitude, particularly in South

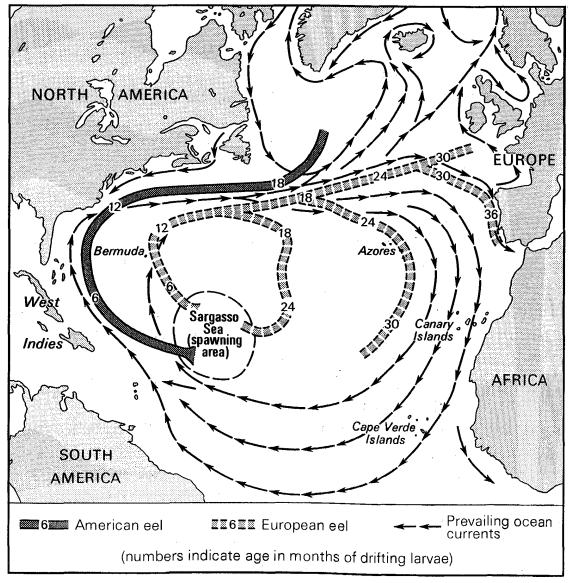


Figure 20: Migration of eels in the Atlantic Ocean.

Africa, in the Congo River region, and in some coastal areas of West Africa.

Migration routes of storks, ducks, geese, and swans

Among nonpasserines—*i.e.*, nonperching birds—one of the best known migrants is the stork (*Ciconia ciconia*), which migrates to tropical Africa along two well-defined flyways. The stork population nesting west of a line that follows the Weser River in West Germany flies southwestward through France and Spain, past the Strait of Gibraltar, and reaches Africa by way of West Africa; the eastern population, by far more numerous, takes a route over the straits of the Bosporus, through Turkey and Israel, to east Africa. These well-separated routes are probably a result of the stork's aversion to long flights over water.

Ducks, geese, and swans also are migrants. These birds winter partly in western Europe and partly in tropical

From J. Dorst, *The Migrations of Birds*; published by Houghton Mifflin Company

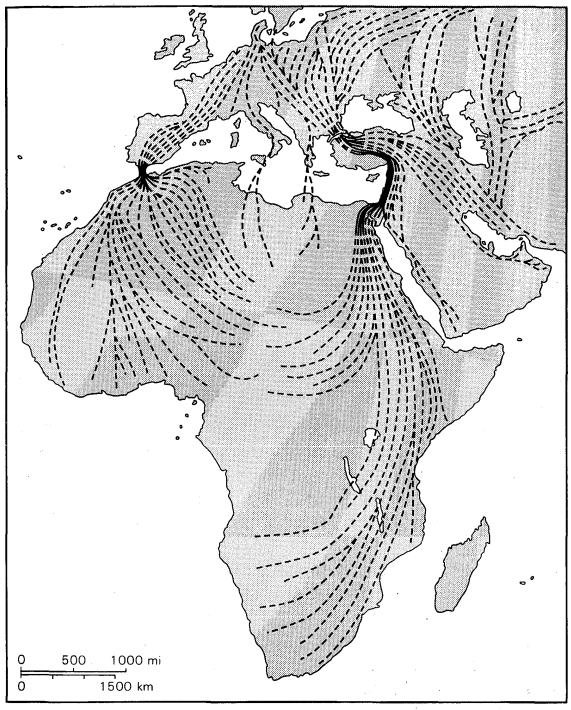


Figure 21: Principal routes taken by the European white stork (*Ciconia ciconia*) between nesting grounds in Europe and wintering grounds in Africa.

Africa. In Africa they are likely to spend the winter in lake and river regions from Senegal in western Africa to the Sudan in eastern Africa, where thousands of garganeys (*Anas querguedula*) and pintails (*A. acuta*) congregate annually. Some ducks leave their breeding grounds to molt (a process by which old feathers are replaced) in areas where they are most secure from predators during the time they are unable to fly; this is known as a molt migration. After molting, the ducks fly to their final winter quarters. Wading birds (Limicolae) also are typical migrants, most of them nesting in tundra of the Arctic region and wintering along the seacoasts from western Europe to South Africa.

In North America. North American birds must endure the same hazards of winter as European species. The geographical arrangement of the continent determines the main routes of migration, which run from north to south and include the Atlantic oceanic route, the Atlantic Coast route, the Mississippi flyway, the central flyway, the Pacific flyway, and the Pacific oceanic route. A great many birds pass the winter in the Gulf States, where the climate is favourable and food is abundant, but the principal wintering area extends through Mexico and Central America to Panama, which has the greatest density of winter bird residents in the world.

The ruby-throated hummingbird (*Archilochus colubris*) nests in southern Canada and winters in Central America as far south as Panama. Some of these birds fly nonstop across the Gulf of Mexico. Because of their food requirements, many American flycatchers (Tyrannidae), which are mainly insectivorous, have the same migratory behaviour as the hummingbirds. Others, like the phoebe (*Sayornis phoebe*), spend the winter in the Gulf States. Birds such as the American robin (*Turdus migratorius*) and several species of grackles assemble in the Gulf States in enormous flocks. The seasonal flights of the American wood warblers (Parulidae) are among the most spectacular on the North American continent. Some spend the winter in the Gulf States and in the West Indies; others, such as the blackpoll warbler (*Dendroica striata*), travel to Guiana, Brazil, and Peru by way of the West Indies. The spring migration routes of the Canada goose span the Continent of North America in an east–west direction from Hudson Bay as far south as Chesapeake Bay.

South America is winter quarters for several tanagers, such as the scarlet tanager (*Piranga olivacea*) and the bobolink (*Dolichonyx oryzivorus*); these birds migrate through the eastern United States and past Cuba to the swampy regions of Bolivia, southern Brazil, and northern Argentina. This area of South America is also winter quarters for the American golden plover (*Pluvialis dominica dominica*), which travels in an enormous loop over much of the New World. After nesting in the tundras of Alaska and Canada, the plover assemble in Labrador in easternmost Canada, then fly to Brazil over an oceanic route (the shortest possible route) about 3,900 kilometres (2,400 miles) long. Their return flight traverses South America, Central America, and the Gulf of Mexico, then follows the Mississippi Valley.

The role of wet and dry seasons

In intertropical regions. Birds of tropical regions migrate according to the rhythmic succession of wet and dry seasons—a profoundly influential factor on the annual cycle of animals and plants alike.

The migratory behaviour of birds has a unique regularity in Africa, where life zones are arranged symmetrically by latitudes away from the equator. Some migrants never cross the equator. The standard-wing nightjar (*Macrodipteryx longipennis*), which nests in a belt extending from Senegal in the west to Kenya in the east along the equatorial forest, migrates northward to avoid the wet season. The plain nightjar (*Caprimulgus inornatus*), on the other hand, nests in a dry belt from Mali in the west to the Red Sea and Kenya in the east during the rains and then migrates southward to Cameroon and the northern Congo region during the dry season.

Other birds migrate across the Equator to their alternate seasonal grounds. Abdim's stork (*Sphenorhynchus abdimii*) nests in a belt extending from Senegal to the Red Sea; after the wet season, it winters from Tanzania through most of southern Africa. The pennant-wing nightjar (*Cosmetornis vexillarius*), in contrast, nests in the Southern Hemisphere south of the Congo forests during the austral, or Southern Hemisphere, summer, then starts north with the onset of the rainy season. It spends its winters in savannas from Nigeria to Uganda.

In coastal and pelagic regions. Among the migrating seabirds, a distinction must be made between the coastal and the pelagic, or open-sea, species. Birds such as guillemots, auks, cormorants, gannets, and gulls—all common to the seashore—stay in the zone of the continental shelf. Except during the breeding season, they are dispersed over a vast area, often preferring specific directions of travel. Gannets (*Sula bassana*) nesting around the British Isles spread in winter along the Atlantic coast of Europe and Africa to Senegal, the young travelling farther than the adults. Pelagic birds, most of which belong to the order Procellariiformes (petrels and albatrosses), cover much greater distances and, from a few small nesting areas, roam over a large part of the oceans.

Wilson's petrels (*Oceanites oceanicus*), which nest in the western sector of the Antarctic (South Georgia Island, Shetland Islands, and South Orkney Islands), spread rapidly northward in April along the coasts of North and South America and stay in the North Atlantic during the summer. In September they leave the western Atlantic, travelling east, then southeast, along the coasts of Europe and Africa toward South America and their Antarctic breeding grounds, arriving there in November. These petrels thus travel in a great loop through the whole Atlantic

The significance of prevailing winds

Ocean, in a flight pattern correlated with the direction of prevailing winds. The same pattern is used by other seabirds normally carried by the winds. Albatrosses, such as the wandering albatross (*Diomedea exulans*) that nests on small Antarctic islands, circle the globe during their migrations. One such bird, banded as a chick at Kerguelen Island in the southern Indian Ocean and recovered at Patache, Chile, travelled in less than 10 months at least 13,000 kilometres (8,100 miles)—perhaps as much as 18,000 kilometres (11,200 miles)—by drifting with the prevailing winds.

In the Pacific, short-tailed shearwaters (*Puffinus tenuirostris*) nest in enormous colonies along the coasts of southern Australia and in Tasmania, then migrate across the western Pacific to Japan, remaining in the North Pacific and the Arctic Ocean from June to August. On the return migration they go east and southeast along the Pacific coast of North America, then fly diagonally across the Pacific to Australia.

Arctic terns (*Sterna paradisaea*), whose breeding range includes the northernmost coast of Europe, Asia, and North America, spend the winter in the extreme southern Pacific and Atlantic, chiefly along Antarctic pack ice 17,600 kilometres (11,000 miles) from their breeding range. American populations of the Arctic tern first cross the Atlantic from west to east, then follow the coast of western Europe. Arctic terns thus travel further than any other bird species.

Modes of migration. The migration flights of birds follow specific routes, sometimes quite well defined over long distances. The majority of bird migrants, however, travel along broad airways. A single population of migrants may be scattered over a vast territory so as to form a broad front hundreds of miles in width. Such routes are determined not only by geographical factors—*e.g.,* river systems, valleys, coasts—and ecological conditions but are also dependent upon meteorological conditions; *i.e.,* birds change their direction of flight in accordance with the direction and force of the wind. Some routes cross oceans. Small passerine (perching) birds migrate across 1,000 kilometres (620 miles) or more of sea in areas such as the Gulf of Mexico, the Mediterranean Sea, and the North Sea. American golden plover, wintering in the Pacific, fly directly from the Aleutian Islands (southwest of Alaska) to Hawaii, the 3,300-kilometre (2,050-mile) flight requiring 35 hours and more than 250,000 wing beats.

The speed of migratory flights depends largely on the species and the type of terrain covered. Birds in migration go faster than otherwise. Rooks (*Corvus frugilegus*) have been observed migrating at speeds of 51 to 72 kilometres (32 to 45 miles) per hour; starlings (*Sturnus vulgaris*) at 69 to 78 kilometres (43 to 49 miles) per hour; skylarks (*Alauda arvensis*) at 35 to 45 kilometres (22 to 28 miles) per hour; and pintails (*Anas acuta*) at 50 to 82 kilometres (31 to 51 miles) per hour. Although the speeds would permit steadily flying migrants to reach their wintering grounds in a relatively short time, the journeys are interrupted by long stops, during which the birds rest and hunt for food. The redbacked shrike (*Lanius collurio*) covers an average of 1,000 kilometres (620 miles) in five days as follows: two nights for migration, three nights for rest, five days for feeding.

Most migrations occur at relatively low altitudes. Small passerine birds often fly at less than 60 metres (200 feet). Some birds, however, fly much higher. Migrating passerines, for example, have been observed at altitudes as great as 4,000 metres (14,000 feet). The highest altitude recorded thus far for migrating birds is 9,000 metres (29,500 feet) for geese near Dehra Dūn in northwest India.

Pelicans, storks, birds of prey, swifts, swallows, and finches are diurnal (daytime) migrants. Waterbirds, cuckoos, flycatchers, thrushes, warblers, orioles, and buntings are mostly nocturnal (nighttime) migrants. Studies of nocturnal migrants using radar on telescopes focussed on the Moon show that most migratory flights occur between 10 PM and 1 AM, diminishing rapidly to a minimum at 4 AM.

Gregariousness during migration

Most birds are gregarious during migration, even those that display a fierce individualism at all other times, such as many birds of prey and insectivorous passerines. Birds with similar habits sometimes travel together, a phenomenon observed among various species of shorebirds. Flocks sometimes show a remarkable cohesion; the most characteristic migratory formation of geese, ducks, pelicans, and cranes is a "V" with the point turned in the direction of flight.

Mammals. Seasonal movements are not widespread among terrestrial species of mammals, because walking speed is relatively slow and energy consumption great. Marine and flying mammals have a much greater tendency to migrate, a tendency that is directly related to their locomotive powers.

Terrestrial mammals. True migration among mammals occurs mostly among large artiodactls (split-hoofed animals) living in habitats with wide fluctuations of climatic and biotic conditions.

In North American Arctic regions, herds of caribou (*Rangifer tarandus*) settle during the summer in the barrens—rather flat wasteland with little vegetation. In July, the animals begin to move irregularly southward and spend the winter in the taiga, or northern forests, through which they wander freely with no general directional trend. Each herd seems to move in accordance with local conditions and without a well-defined pattern. The caribou again move northward as early as late February and return to the barrens. These migrations follow the same routes from year to year.

In former times, American bison (*Bison bison*) migrated regularly through the Great Plains. Herds of as many as 4,000,000 animals moved from north to south in fall and returned when spring rains brought fresh grass to the northern part of their range. Bison travelled over more or less circular routes and spent the winter in areas 320 to 640 kilometres (200 to 400 miles) from the summer range. Other North American mammals, such as elk (*Cervus canadensis*), mule deer (*Odocoileus hemionus*), and dall sheep (*Ovis dalli*) still migrate regularly in areas undisturbed by man.

Large African mammals migrate in accordance with the succession of wet and dry seasons, which can greatly modify the habitat. Some antelope remain in small areas throughout the year, but many species undertake seasonal movements over a large range. In the Serengeti region of Tanzania, plains animals, particularly wildebeests (*Connochaetes taurinus*) and zebras, travel more than 1,600 kilometres (1,000 miles) in their seasonal migrations. Herds spread outward during the rains and concentrate during the dry season around water holes. Elephants (*Loxodonta africana*) wander great distances in search of the best food and water supply.

In southern Africa, hundreds of thousands of springbok (*Antidorcas marsupialis*) once migrated according to the rhythm of rainfall over their vast range. They moved in herds so dense that any animal encountered was either trampled or forced along with the herd. These huge migrations often resulted in enormous losses from starvation, drowning, or disease—natural methods for controlling overpopulation. Such movements, involving lesser numbers, still occur in parts of South West Africa/Namibia and in Botswana.

Flying mammals (bats). A few bats native to Europe and Asia make short flights to winter quarters. In the Soviet Union, the common pipistrelle (*Pipistrellus pipistrellus*) and the frosted bat (*Vespertilio murinus*) withdraw to hibernating places at some distance from their summer range. In Germany, the large mouse-eared bat (*Myotis myotis*) leaves its winter quarters in Brandenburg in March or April and travels as much as 260 kilometres (160 miles) to its summer habitat in northwestern and northeastern Germany. It regularly returns to the same winter locality. Schreiber's long-fingered bat (*Miniopterus schreibersii*) changes its habitat in winter and moves over 160 kilometres (100 miles) or more according to a complex pattern. These local movements represent an adjustment to winter conditions and the search for more habitable caves.

Migratory bats in the United States

Other bats travel even greater distances. In the United States, the red bat (*Lasiurus borealis*), the large hoary bat (*L. cinereus*), and the silver-haired bat (*Lasionycteris noctivagans*)—three species that roost primarily in trees and

shrubs—are true migrants with strong powers of flight. They summer in the northern U.S. and in Canada and winter in Georgia, South Carolina, Florida, and probably also in the southwestern states. The southward movement is made from mid-August to November. Migration flights occur at night, and during the day under favourable conditions. Large numbers follow the coast some distance from land, and all three species are found at sea far from the coast and in Bermuda. Fruit bats and flying foxes (*Pteropus*) native to the tropical regions of the Old World make regular mass migrations, following the seasons for fruit ripening.

Marine mammals. Antarctic whale migrate regularly to the tropics, a fact long known to whalers. By systematically marking whale by shooting into them steel tubes engraved with a serial number, man has obtained evidence of actual movements. A young fin whale (*Balaenoptera physalus*) marked in February in the Antarctic, at latitude 65° S, was captured two years later, in July, off the coast of South Africa, 3,000 kilometres (1,900 miles) north. During the austral (Southern Hemisphere) winter, whale migrate to areas rich in food, particularly the northwestern coast of Africa, the Gulf of Aden, and the Bay of Bengal. Antarctic whale—particularly humpbacks (*Megaptera novaeangliae*), a highly migratory species—can be divided into five distinct populations around Antarctica; each population migrates separately, and individuals usually return to their respective zones, though interchange may occur. The Antarctic population does not, however, migrate entirely into warm waters during the winter, and a segment of the population seems to stay behind at about latitude 50° S.

Northern whale have the same migratory habits as Antarctic whale. Northern blue whale (*Balaenoptera musculus*) migrate northward along the east coast of the United States, then through Davis Strait to Baffin Bay (north of Canada) or Spitsbergen to the waters off northern Scotland or the coast of Norway. They are believed to migrate southward along the same routes. Part of the North Pacific stocks of the northern blue whale winters in the Indian Ocean and in the seas bordering Indonesia.

Smaller cetaceans (porpoises and dolphins) migrate in the same way, as indicated by population fluctuations within a particular area; but little is known about their distribution and migration.

Noteworthy migratory habits occur among the pinnipeds (seals and walrus), some of which disperse over wide areas at times other than the breeding season. The harp seal (*Pagophilus groenlandicus*) lives in summer in northernmost Arctic waters, but reproduces in the White Sea (an arm of the Arctic Ocean extending southward into the landmass of the Soviet Union), in the eastern North Atlantic, and around Newfoundland, where young are born between January and April. The seal then returns to more northern latitudes. Northern fur seals (*Callorhinus ursinus*) reproduce only on the Pribilof Islands, off southwestern Alaska, from May to November, and the colonies then disperse into the open seas. The males stay in the Gulf of Alaska and off the Aleutian Islands; the females go farther south, to southern California, some 4,800 kilometres (3,000 miles) away.

NAVIGATION AND ORIENTATION

Migrants often return to breed in the exact locality where they were hatched or born. This journey homeward, particularly that of birds, may cover thousands of miles.

Homing experiments

Homing experiments have demonstrated the ability of animals to orient themselves geographically. Such experiments involve removing animals from a specific point (usually the nest), transporting them for various distances, and analyzing their speed and degree of success in returning. Starlings have returned to their nests after being transported 800 kilometres (500 miles); swallows have returned a distance of more than 1,800 kilometres (1,100 miles). A Manx shearwater (*Puffinus puffinus*) returned from Massachusetts to Britain, 4,900 kilometres (3,050 miles) across the Atlantic, in 12½ days. Laysan albatrosses (*Diomedea immutabilis*) returned to Midway Island in the Pacific after being released at Whidbey Island, Washington; the journey covered 5,100 kilometres (3,200 miles) and took 10.1 days. Experiments with certain fishes and mammals have demonstrated similar homing ability.

It is apparent that homing animals use familiar landmarks; both random and oriented searches have been observed in birds and fish. Homing experiments with gannets observed from aircraft have demonstrated that, after release, the birds explore the region and hesitate as they apparently look for landmarks. Landmarks vary from topographical (*e.g.*, mountain systems, river systems, coastlines) to ecological (*e.g.*, vegetation zones) to climatic (*e.g.*, air masses differing in temperature and humidity, prevailing winds). Fish may orient themselves by using similar clues in the same way. Passive drifting is an important factor in the movements of larvae and young fishes, such as those of the eel, cod, herring, and plaice, and even in adults that are passive after spawning, such as herring and cod. As a result of drifting with the current, the movements of such fishes are similar from year to year.

Familiar landmarks and exploration do not, however, explain how migrants find their way along routes covering many hundreds or thousands of miles nor do the results of most homing experiments.

Birds. A compass sense has been demonstrated in birds; that is, they are able to fly in a particular constant direction, regardless of the position of the release point with respect to the bird's home area. It has also been shown that birds are capable of relating the release point to their home area and of determining which direction to take, then maintaining that direction in flight. The navigational ability of birds has long been understood in terms of a presumed sensitivity to both the intensity and the direction of the Earth's magnetic field. It has also been suggested that birds are sensitive to forces produced by the rotation of the Earth (Coriolis force); however, no sense organ or physiological process sensitive to such forces has yet been demonstrated to support this hypothesis.

Experiments have shown that the orientation of birds is based on celestial bearings. The Sun is the point of orientation during the day, and birds are able to compensate for the movement of the Sun throughout the day. A so-called internal clock mechanism in birds involves the ability to gauge the angle of the Sun above the horizon. Similar mechanisms are known in many animals and are closely related to the rhythm of daylight, or photoperiodism (see above). When the internal rhythm of birds is disturbed by subjecting them first to several days of irregular light–dark sequences, then to an artificial rhythm that is delayed or advanced in relation to the normal rhythm, corresponding anomalies occur in the homing behaviour.

Two theories have been formulated to explain how birds use the Sun for orientation. Neither theory, however, has thus far been substantiated with proof. One theory holds that birds find the right direction by determining the horizontal angle measured on the horizon from the Sun's projection. They correct for the Sun's movement by compensating for the changing angle and thus are able to maintain the same direction. According to this theory, therefore, the Sun is a compass that enables the birds to find and maintain their direction. This theory does not explain, however, the manner in which a bird, transported and released in an experimental situation, determines the relationship between the point at which it is released and its goal.

The second theory, proposed by British ornithologist G.V.T. Matthews, is based on other aspects of the Sun's position, the most important of which is the arc of the Sun—*i.e.*, the angle made by the plane through which the Sun is moving in relation to the horizontal. Each day in the Northern Hemisphere, the highest point reached by the Sun lies in the south, thus indicating direction; the highest point is reached at noon, thus indicating time. In its native area a bird is familiar with the characteristics of the Sun's movement. Placed in different surroundings, the bird can project the curve of the Sun's movement after watching only a small segment of its course. By measuring maximum altitude (the Sun's angle in relation to the horizontal) and comparing it with circumstances in the usual habitat, the bird obtains a sense of latitude. Details of longitude are provided by the Sun's position in relation

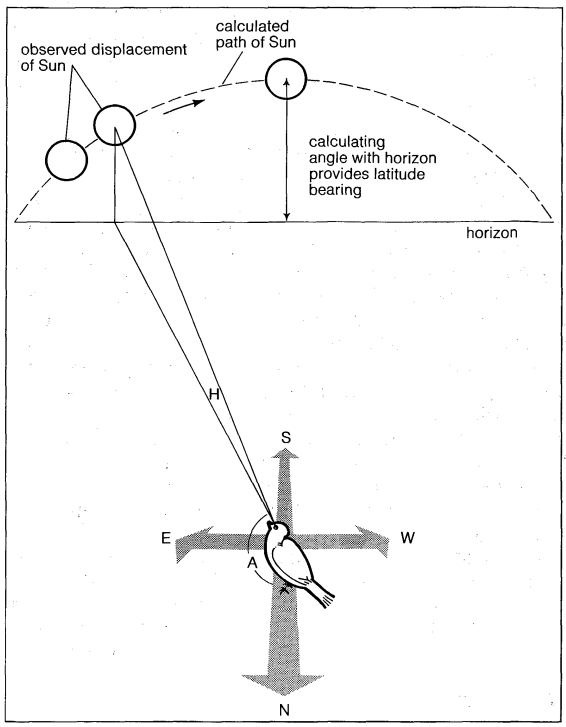


Figure 22: Matthews' hypothesis of solar navigation by birds. Birds can visualize the complete trajectory of the Sun and determine its zenith and altitude at midday (angle H) in order to derive the latitude. Angle A is measured from the north to the point at which a vertical from the Sun cuts the horizon. From this longitude can be derived.
From M. Ricard, *The Mystery of Animal Migration*

to both the highest point and position it will reach—as revealed by a precise internal clock.

Use of celestial navigation by birds

Migrant birds that travel at night are also capable of directional orientation. Studies have shown that these birds use the stars to determine their bearings. In clear weather, captive migrants head immediately in the right direction using only the stars. They are even able to orient themselves correctly to the arrangement of night skies projected on the dome of a planetarium; true celestial navigation is involved because the birds determine their latitude and longitude by the position of the stars. In a planetarium in Germany, blackcaps (*Sylvia atricapilla*) and garden warblers (*S. borin*), under an artificial autumn sky, headed "southwest," their normal direction; lesser whitethroats (*S. curruca*) headed "southeast," their normal direction of migration in that season.

It is known, then, that birds are able to navigate by two types of orientation. One, simple and directional, is compass orientation; the second, complex and directed to a point, is true navigation, or goal orientation. Both types apparently are based on celestial bearings, which provide a navigational "grid."

Other animals. The methods of directional orientation used by birds are similar to those used by other animals. Orientation to the Sun has been demonstrated in various crustaceans, particularly in the sand flea (*Talitrus saltator*). Various insects, particularly bees and certain beetles (families Scarabeidae, Tenebrionidae, and Carabidae) use the Sun to plot their course with remarkable accuracy.

Fishes also are able to use celestial bearings; salmon presumably use the Sun. Experiments with the parrot fish (*Scarus*) have demonstrated a Sun compass reaction that may also occur in other fishes. Localization of the Sun is, however, much more difficult in water than in the air, because of the characteristics of light rays passing through water. Experiments suggest that topographical clues are also used by fishes to recognize their range, particularly their spawning grounds. Visual bearings in this respect have great importance. It is possible that chemical substances also provide clues; the role of olfaction in the salmon has been described (see above, *Anadromous fishes*).

Visible landmarks are used by mammals, at least for orientation within short distances. Scented trails are probably helpful within a limited area, proportionate to the size of the animal; olfaction plays an important role in the life of mammals. Some mammals, however, migrate over enormous distances and are able to return after being taken far away from their home territory; bats, for example, have returned 265 kilometres (165 miles) to their caves. Random exploration plays a part in such movements, but it is possible that some type of true navigation is involved in certain of these movements.

PHYSIOLOGICAL STIMULUS OF MIGRATION

Migration, like reproduction and other phases (as molting in birds), is part of the life cycle and depends on a complex internal rhythm that affects the whole organism, particularly the endocrine glands (glands of internal secretion) and the gonads. Migration must thus be viewed in relation to the entire annual cycle.

Each year birds return to particular areas to breed, and remain there until the members of the brood can care for themselves. There is no relation between the reproductive and migratory stimuli, yet the two phenomena, although independent, are nevertheless stimulated by the same factor.

A physiological study of certain migrants has revealed that metabolic patterns usually change prior to migration, and fats accumulate in the body tissues. The whitethroat (*Sylvia communis*) weighs an average of 12 to 13 grams (about 0.4 ounce) during the breeding season, 16 to 19 grams (about 0.6 ounce) in the autumn, and 20 to 24 grams (about 0.8 ounce) in the winter. Food consumption increases with the autumn molt, reaching a peak at the beginning of the migration season. These fundamental physiological changes, chiefly under the control of the thyroid gland, are correlated with migratory activity. Such fluctuations are not observed in nonmigratory species.

Variations in metabolism and related phenomena are controlled by an endocrine gland, namely the pituitary gland, which is located in the lower part of the brain and acts as a command post, sending out instructions in the form of secretions called hormones. That the pituitary has a cycle independent of environmental factors is demonstrated by the regularity with which phases such as reproduction occur from year to year in the lives of some birds, and by the diverse response of various species and populations to the same environmental factors. That the pituitary is, however, influenced by environmental factors, such as variations in day length and the intensity of the Sun, has been demonstrated experimentally.

The role of the pituitary gland in migration

Gonadal development and the deposition of fat, for example, are influenced by the pituitary, which responds to increasing day length in springtime by accelerating the rate of gonadal development. The pituitary thus governs the development of gonads and, in addition, affects all metabolic processes, including development of the thyroid gland, so as to prepare the animal physiologically for migration. If only the pituitary and variations in day length were involved, migration would be triggered at definite times, because the pituitary cycle is fixed, and photoperiodism is a highly predictable phenomenon; such a lack of flexibility, however, would inevitably cause migrant populations to suffer catastrophes because ecological conditions are irregular—meteorological events, such as the arrival of spring, and biological phenomena, such as flowering, foliation, hatching of insects, and availability of food, are highly variable from year to year. The pituitary thus serves only to prepare the bird for flight; the proper ecological conditions, on the other hand, are necessary to initiate it. The availability of food is an important factor. Temperature and weather conditions also have an influence—a sudden period of cold weather during autumn may induce the immediate departure of many migrants.

Sensitivity to changes in the weather and other environmental conditions varies markedly among species. Some, such as the woodcock, snipe, lapwing, starling, and lark, rely on surrounding conditions to initiate their spring and autumn migrations, and the patterns of their flight depend on temperature and barometric pressure. Others, such as

the swift, cliff swallow, Baltimore oriole, and short-tailed petrel, are less weather dependent, and, since the dates of their arrival and departure are not regulated by the weather, they occur with remarkable regularity each year.

The factors that stimulate migration in animals other than birds are not yet well understood. Ecological conditions play a great part in the migratory activity of mammals, who react to general food shortage by moving to another region. Whale, for example, leave the Antarctic region as winter modifies the oceanographic conditions. Seals disperse when the food supply in the area of their breeding colonies is depleted. Environmental factors are of primary importance in the migration of fishes and marine invertebrates. Annual movements of water masses change physical conditions such as temperature and salinity; biotic conditions are influenced accordingly.

ORIGIN AND EVOLUTION OF MIGRATION

The origins of migration remain in the realm of pure conjecture; neither observation nor experiment has resolved the matter. The explanation, however, must be related to geographical and climatological factors that have prevailed since the Tertiary Period, which ended some 2,500,000 years ago. The great Quaternary ice ages, which came later, were very important in altering the distribution of animals over a large part of the world, but migrations occurred long before.

Migration, as it is now known among modern birds and mammals, probably appeared gradually by stages. Some animals changed their habitat only slightly, never leaving the same general region. The movements of other animals were more erratic, their dispersal being oriented toward the most favourable places. Such movements are the first stages of true migration—a phenomenon characterized by elaborate mechanisms—which gradually acquired stability through natural selection. At first, many populations must have perished rather than attempt to flee from unfavourable conditions. Only a fraction of such populations probably sought more favourable conditions elsewhere, but natural selection favoured the "migrants," and migratory tendencies were retained.

Causes of change of habitat

In some cases, original habitats were in present-day wintering areas, and animals developed a tendency to leave in spring in order to breed in other territories. Seasonal changes of weather and food supply in these newly settled regions forced the animals to migrate in fall, and they thus retreated to their former range. Among birds nesting in the Northern Hemisphere, hummingbirds, tyrant flycatchers, tanagers, orioles, bee-eaters, and swifts have distinct tropical affinities; in recent geological times these birds gradually spread northward as glacial ice receded and the continent became warmer. Other birds, such as plover, ducks, and geese, originally lived in what is now their breeding area. Gradual climatic changes forced them to spend their winters in regions far to the south. Migrations thus appear to be the consequence of invasions or emigrations, during which animals settle in new areas during a segment of the annual cycle.

Migratory birds use the routes by which their ancestors first invaded new regions after the glacial recession. The yellow wagtail (*Motacilla flava*) and the wheatear (*Oenanthe oenanthe*) settled in Alaska; they migrate annually into other parts of the Western Hemisphere but spend their winters in the warm regions of southeastern Asia and even Africa, probably following the migratory route of their ancestors. A typically North American species, the gray-cheeked thrush (*Hylocichla minima*), which has extended its breeding area to northeastern Siberia, returns to spend the winter in the central regions of South America.

ECOLOGICAL SIGNIFICANCE OF MIGRATION

There are many ecological implications of migration. The food resources of some regions would not be adequately exploited without moving populations. The sequence of migratory movement is closely integrated in the annual cycle of ecosystems characterized by productivity fluctuations. Migratory behaviour concerns only species located at specific trophic levels (zones of food availability) where maximal fluctuations occur both in breeding areas and in wintering regions. Migrant birds avoid equatorial forests where productivity is constant throughout the year, and food surpluses do not occur. They do congregate, on the other hand, in savannas where productivity varies with the seasons.

Such a coordinated sequence is particularly apparent in the case of birds migrating from the northern Arctic regions to tropical winter regions; both life zones are characterized by broad fluctuations in productivity. In the Arctic, vegetal and animal production is very high during the summer; ducks and waders nest in great numbers, exploiting these resources. As winter comes, food becomes scarce, and water birds migrate to the tropics, where the rainy season has caused food production to increase to optimal levels. Ducks and wading birds concentrate in the most favourable areas, remaining until spring, when productivity is lowest. By then the condition of breeding areas is again favourable for the birds. The life cycles of these birds are closely attuned with the cycles of their various habitats, and the sizes of bird populations are controlled by the capacity of both areas to sustain them.

Migration, then, has considerable ecological significance. It enables fast-moving animals to exploit fluctuating resources and to settle in areas where life would not be tenable for animals incapable of rapid travel. On the other hand, peaks of food production would be unexploited without the periodic presence of migratory populations.

(J.P.D./Ed.)

Dormancy

There are few environments in which organisms are not subject to some kind of environmental stress. As discussed above, some animals migrate vast distances to avoid unfavourable situations; others reduce environmental stresses by modifying their behaviour and the habitats (immediate surroundings) that they occupy. Arctic lemmings, for example, are able to avoid severe winter weather by confining their life in winter to activities beneath the snow cover. Still another mechanism used by some organisms to avoid stressful environmental conditions is that of dormancy, an inactive state accompanied by a lower than normal rate of metabolism—the chemical processes responsible for the activity, nourishment, and growth of an organism—during which an organism conserves the amount of energy available to it and makes few demands on its environment. Most major groups of animals as well as plants have some representatives that can become dormant; more species hibernate than become dormant, however. Periods of dormancy vary in length and in degree of metabolic reduction, ranging from only slightly lower metabolism during the periodic, short-duration dormancy of deep sleep to more extreme reductions for extended periods of time.

Lowering of metabolic rate

GENERAL OBSERVATIONS

Value of dormancy. In terms of evolution, dormancy seems to have evolved independently among a wide variety of living things, and the mechanisms for dormancy vary with the morphological and physiological makeup of each organism. For many plants and animals, dormancy has become an essential part of the life cycle, allowing an organism to pass through critical environmental stages in its life cycle with a minimal impact on the organism itself. When lakes, ponds, or rivers dry up, for example, aquatic organisms that can enter a period of dormancy survive, while others perish. Moreover, animals that can become dormant during the extreme cold of winter can extend their ranges into regions where animals incapable of dormancy cannot live. Dormancy also ensures that these animals will be free from competition during their periods of activity. Thus, dormancy is an adaptive mechanism that allows an organism to meet environmental stresses and to take advantage of environmental niches that otherwise would be untenable at certain times.

Causes of dormancy. The dormant state that is induced in an organism during periods of environmental stress may be caused by a number of variables. Those of major importance in contributing to the onset of dormancy

include changes in temperature and photoperiod and the availability of food, water, oxygen, and carbon dioxide. In general, because organisms normally exist within a relatively narrow temperature range, temperatures above or below the limits of this range can induce dormancy in certain organisms. Temperature changes also affect such other environmental parameters as the availability of food, water, and oxygen, thus providing further stimuli for dormancy. In Arctic regions, for example, certain animals become dormant during the winter months, when food is less abundant. In desert biomes, on the other hand, the summer months, which may be periods of reduced food availability, intense heat, or extreme aridity, stimulate some desert organisms to become dormant. The lack of water during summer periods of drought or winter periods of freezing, as well as annual changes in the duration and intensity of light, particularly at high latitudes, are other environmental factors that can induce dormant states.

Cyclical environmental changes

Under natural conditions, most of the environmental variables that influence dormancy are interrelated in a cyclical pattern that is either circadian or annual. Fluctuations in the major daily variables—light and temperature—can induce rhythmical changes in the metabolic activity of an organism; annual fluctuations in temperature and photoperiod can influence the availability of food and water. Concentrations of oxygen and carbon dioxide normally do not vary on a cyclical basis but as a result of habitat selection, such as burrowing in the mud, seeking a den, or other similar activities, in which the metabolic responses of the organism can alter the oxygen and carbon dioxide concentrations in its environment.

In an attempt to determine the relative influence of environmental factors upon dormancy, they have been varied experimentally. Investigations indicate that an organism, after it has adapted to a sequence of cyclical rhythms, tends to maintain its adaptive behaviour even though the environmental stimulus that originally elicited such behaviour is no longer present. For example, the Arctic ground squirrel (whose winter period of dormancy is referred to as hibernation), when taken into the laboratory, supplied with adequate amounts of food and water, and exposed to constant temperature and light, exhibits periodic torpor (extreme sluggishness)—an innate behavioral pattern that operates independently of environmental cues. Other animals frequently will continue to respond as if they were exposed to the cyclical changes of their home environments after they are removed from their natural habitats.

DORMANCY IN PROTOZOANS AND INVERTEBRATES

Cysts and cystlike structures. *Protozoans.* Many parasitic and free-living protozoans (one-celled animals) exhibit a dormant stage by secreting a protective cyst. The stimulus for cyst formation in free-living protozoans may be temperature changes, pollution, or lack of food or water. *Euglena,* a protozoan that encysts to avoid environmental extremes, has two kinds of cysts. Apparently one is formed only to avoid stressful conditions; the other is formed for the same reason but also involves asexual reproduction, resulting in a cyst that may contain up to 32 daughter organisms, which emerge under proper environmental conditions.

The role of cysts in parasitic protozoans

Free-living protozoans form cysts around themselves and avoid environmental extremes, but cysts are a part of the life cycle of parasitic protozoans. The causative agent of amebic dysentery, *Entamoeba histolytica,* is found in the intestine of infected individuals, in whom it forms cysts that pass to the outside in feces. When food or water containing cysts enters the digestive tract of another person, the amoebas are released from the cysts and infect the new host. Without encystment, which allows the organism to live in a dormant state in an unfavourable environment (*e.g.,* water), amebic dysentery could be much more easily controlled. Protected by the cyst wall, however, the dormant contents of the cyst can survive for weeks. Although they are not particularly resistant to drying, the cysts of *E. histolytica* can withstand temperatures of up to 68° C (154° F) for five minutes. They are also resistant to certain chemicals.

Invertebrates. Dormant cysts are formed during the life cycles of invertebrate parasites such as the oriental liver fluke (*Clonorchis sinensis*). The cyst stage of this organism develops in fish muscle; if the fish is eaten raw or undercooked, the encysted fluke is transferred to a new host. The encysted stage of the trichina worm (*Trichinella spiralis*), which causes trichinosis, is found in the muscle cells of hogs; it is also an invertebrate parasite in which the dormant stage is an essential part of the life cycle. When undercooked pork is eaten, the cyst wall is dissolved by digestive juices, and the worm is able to make its way into the tissues of a new host.

The cystlike forms found in many other invertebrate groups are all dormant stages that preserve the species during times of environmental stress. All freshwater sponges and some marine species survive cold or drought by forming gemmules within the body of the adult sponge. These structures, which are surrounded by a resistant covering, are released when the sponge dies and disintegrates. When conditions are appropriate, the cell mass escapes from the covering and forms a new sponge.

Rotifers are microscopic aquatic animals that produce winter eggs with thick and resistant coverings similar to protozoan cysts; the eggs may remain dormant for long periods. They can survive drought or freezing and may be dispersed by wind or carried by animals. Thus, the cyst serves not only for survival of the egg under adverse conditions but also for dispersal. Some freshwater bryozoans develop disklike buds, or statoblasts, that are surrounded by a hard, chitinous (horny) shell. These statoblasts are the dormant structures that survive when the bryozoan dies in the fall or during a drought; they form a new bryozoan colony when favourable environmental conditions again prevail.

Among mollusks, land snails remain largely dormant throughout the day, with the soft head and foot withdrawn into the shell. During periods of drought or cold, they retreat into their shells and secrete a membrane (the epiphragm) of mucus and lime that covers the opening of the shell and resists desiccation. Slugs, on the other hand, bore into the ground and secrete a mucus mantle around themselves for protection during periods of unfavourable environmental conditions. Among the arthropods, many freshwater forms develop dormant cystlike stages that resist desiccation and allow the species to survive unfavourable periods.

Diapause in insects. Many insects undergo periods of reduced metabolic activity called diapause. Diapause, which may occur during any stage of the life cycle—egg, nymph, larva, pupa, or adult—is usually characterized by a cessation of growth in the immature stages and a cessation of sexual activity in adults. In some insects, it is a reaction to unfavourable environmental conditions; in others, such as certain moths and butterflies, diapause is a necessary stage of the life cycle. The 17-year larval and pupal periods of the cicada are examples of diapause. This form of dormancy is particularly common among insects that live in arid desert areas, where during the dry and hot summers, the insects usually hide themselves in the soil at suitable depths or under any available protective objects.

Insect survival in winter

Insects may overwinter as egg, larva, nymph, pupa, or adult; because they can stand very low temperatures, few of these forms die if the winter temperatures are within their normal range. Even rather fragile forms, such as mosquitoes and butterflies, survive in sheltered, relatively dry places out of doors. Some butterflies even survive the winter in low shrubbery, where they may be completely covered by snow and ice for three or four months. Other insects prepare for winter by constructing nests or cocoons; still others seek suitable hiding places.

Among some insect species, diapause lasts only until favourable environmental conditions return, after which the insect immediately resumes its normal activities. In other species, favourable environmental conditions alone do not break the diapause; some other stimulus, such as cold or food, is necessary. The eggs of the mosquito *Aedes vexans,* for example, remain in diapause until the damp soil on which the eggs are laid is flooded to form a pool suitable for the larvae. Although the eggs of another mosquito, *Aedes canadensis,* are laid in the same soil as those of

Aedes vexans, they will not hatch—even after flooding—until they have been subjected to cold. Thus, when both species lay their eggs together in early summer, those of *Aedes vexans* hatch in pools formed by late summer rains, but those of *Aedes canadensis* overwinter and hatch in the spring rain pools. Not only are certain conditions required to break diapause but in some species, such as certain cutworms, a specific length of time must elapse before the stimuli are effective.

The onset of diapause depends upon a combination of environmental factors operating on the regulatory mechanisms—*i.e.,* nervous and endocrine systems—of the insect. Photoperiod and temperature influence the endocrine function of the brain, which synthesizes and secretes a substance (hormone) that controls other endocrine organs, specifically the prothoracic glands. Under the stimulation of the brain hormone, the prothoracic glands secrete a hormone called ecdysone. When stimulation by the brain hormone ceases, ecdysone is no longer secreted, and, in its absence, all insect growth and metamorphosis are halted. Thus, provision is made for the overwintering of immature insects in a state of developmental standstill. With the arrival of more favourable conditions, ecdysone is again secreted, and development resumes. Because many insect species have more than one generation of progeny per year, the prothoracic glands do not cease functioning except at some stage in the life cycle of the brood that must overwinter.

DORMANCY IN COLD-BLOODED VERTEBRATES

Two kinds of dormancy can be distinguished in vertebrates on the basis of body temperature. Most vertebrates are poikilothermous, or cold-blooded, because the body temperature follows that of the environment and is not kept constant by internal (homoiostatic) mechanisms. The second group, the homoiotherms, maintain a constant body temperature regardless of the environmental (ambient) temperature; called warm-blooded animals, they include birds and mammals.

Fishes and amphibians. The metabolism of poikilothermous animals is most influenced by the environmental variables of temperature, nutrition, and photoperiod. Photoperiod, the daily length of light exposure, has a marked metabolic effect in both fishes and amphibians; fishes, however, remain active throughout the year, although the activity may be limited by temperature, as in those fish that rest on the bottom or in mud during cold periods. Brief superficial freezing and supercooling (without freezing) to temperatures below the freezing point of body fluids are experienced by resistant species, but it has not been established that fishes that have been frozen solid can become active when thawed. In the Arctic, no fishes are found in lakes that freeze solid in the winter. Because most fishes do maintain some kind of activity year round, they cannot be said to become dormant in the sense in which the word is used in this article.

In addition to light and temperature, another environmental stress imposed upon fish is drought. Lungfishes, as represented by the African lungfish (*Protopterus*), burrow deeply into the mud when their water supply is diminished. They surround themselves with a cocoon of slime and remain inactive. Their gills are nonfunctional during this period of dormancy, and they use a lunglike air bladder for respiratory purposes. They rely on fat reserves as an energy source, and in order to conserve water, they excrete urea rather than ammonia. This is because ammonia as an excretory product is highly toxic; animals that excrete ammonia require large quantities of water to dilute it below toxic levels. Urea is a semi-solid substance of low solubility, and requires little or no water for its excretion. (Desert animals and many insects excrete urea.)

Protective niches

During periods of drought or cold, amphibians seek protective niches in which to remain dormant until the return of favourable environmental conditions. Overwintering of frogs and salamanders frequently involves their aggregation in large numbers in a moist terrestrial niche, such as a rotting log, the mud on banks or bottoms of marshes and ponds, or in springs. The more terrestrially oriented amphibians, such as toads, may pass the winter in solitary burrows on land. During dry seasons, frogs may be dormant in a mud cocoon.

Reptiles. *Effects of temperature.* Because reptiles depend on external sources of heat to keep warm, they survive during periods of low temperature by seeking a place where the temperature will not fall below freezing, except temporarily. The commonest niche for reptilian dormancy is almost always found underground at a depth dependent on the thermal conductivity of the soil relative to the minimum temperature reached. This factor alone can control the distribution of reptiles. None can survive in the Arctic or Antarctic in places in which the subsoil is permanently frozen; and relatively few can exist in areas near these regions, even if suitable sites for dormancy were available, because the short summers would prevent the completion of life cycles. Although the distribution of snakes at high latitudes or altitudes is limited, the adder has been found at 3,300 metres (10,000 feet) in the Swiss Alps and as far north as the Arctic Circle. The Himalayan pit viper has been found at an altitude of 5,000 metres (16,000 feet).

Dormancy in reptiles may display a circadian rhythm, a seasonal one, or both; it is a state of torpor directly induced by low temperature. When the adder, for example, experiences temperatures of about 8°–10° C (46°–50° F), it begins to search out suitable niches in which to rest. Its dormancy ends on the first sunny days after the maximum temperature has reached 7.5° C (45.5° F). Because these conditions vary, the adder's period of dormancy extends from 275 days in northern Europe to 105 days in southern Europe and is about two weeks in the United Kingdom, where the Gulf Stream provides warmth.

Reptiles also normally become dormant during the hottest parts of summer, but the physiology of summer dormancy is quite different from that of winter. As already mentioned, winter dormancy is a state of torpor, induced by a low temperature, that becomes more pronounced as the temperature falls. There is, however, a wide range between the animal's normal, active (coenothermic) temperature and the lowest temperature at which it can exist. At high temperatures, on the other hand, there is a much narrower range between the coenothermic temperature and temperatures that cause death. In other words, reptiles can tolerate colder temperatures much better than they can tolerate higher ones. For this reason, during hot weather they must seek refuge underground or in cool, shady places, where they remain physiologically active but must forego all normal activity because of the restricted nature of the cooler niche. Desert reptiles, in particular, exhibit such temperature responses daily.

During its dormancy, the amount of water needed by a reptile is less than at other times and is normally supplied by water produced from the metabolism of the animal's own stored food reserves, particularly fat. In areas in which alternating wet and dry seasons occur, reptiles maintain a longer period of dormancy during the dry season. This behaviour may be related more to the lack of available water than to temperature, because in such areas the onset of the seasonal monsoons elicits a period of increased reptile activity.

Winter dens of reptiles

Because there is only a limited number of suitable sites available for dormancy, several snakes, usually of the same species, may be found in each niche. As many as 100 or more snakes have been taken from one winter den. Occasionally, lizards and toads may also be found in the same den, but stories of snakes that share denning sites with small birds and mammals have been difficult to substantiate. It is much more usual to find that the entry of snakes into the burrow of a prairie dog or some other warm-blooded animal is followed by the evacuation of the original occupant.

Effects of latitude. Changes in latitude not only alter the lengths of the dormant and active periods of reptiles but also affect their circadian rhythms because of the changes in the proportions of night to day. Many species of snakes, including the adder, are normally active in the early evening. In the northerly latitudes (*e.g.,* northern Europe, such as Scandinavia and Finland), where the length of the active season is reduced by as much as two-thirds, these

snakes are active throughout the day and are able to take advantage of every warm hour in order to complete the necessary portions of their life cycle. Even this increased activity during the shorter summer season, however, does not compensate for the latitude. Growth and development slow to such a point that sexual maturity is delayed, and the reproductive period requires two years rather than one; young are produced only every other year instead of every year, as at lower latitudes.

DORMANCY, HIBERNATION, AND ESTIVATION IN WARM-BLOODED VERTEBRATES

Hibernation and estivation

The term hibernation is often loosely used to denote any state of torpor, inactivity, or dormancy that an organism might exhibit. Properly speaking, however, use of the term should be confined solely to warm-blooded homoiotherms; *i.e.,* birds and mammals whose feathers or fur serve as insulation to reduce heat radiating from the body and aid in the maintenance of constant body temperatures, which normally are independent of those of the environment. Because warm-bloodedness gives animals an internal physiological stability, they are less dependent on many environmental restrictions, particularly those limitations imposed on organisms by ambient temperatures. For example, only two species of reptiles are found north of the Arctic Circle, but great numbers of birds live and breed there. Warm-bloodedness also signifies a high metabolic rate, a factor that undoubtedly influences normal learning, which depends heavily on the frequency and recency of experiences. Because periods of lowered metabolism interrupt continuous learning experiences, they may explain in part why birds and mammals are so much easier to train than any other animal. The benefits of warm-bloodedness require the expenditure of large amounts of energy through the year and make a heavy demand on available food supplies.

The term hibernation is also used to delineate the dormant state only during winter. In arid regions a reverse phenomenon is seen in which the animal becomes torpid during the hot, dry, barren summer; such hibernation is called estivation. As a means of avoiding environmental stresses, hibernation and estivation are not common devices among warm-blooded animals and they are far less common among birds than among mammals.

Some warm-blooded organisms exhibit thermic instability, a heterothermous condition that allows their metabolic rate to be reduced, with a commensurate reduction in body temperature. Heterothermy is a transitional state between cold-bloodedness and warm-bloodedness; the animal is awake and moving during its temperature fluctuations. The body temperature, although not as constant as in humans, is not so low as to force the organism into deep hibernation. Among mammals, two monotremes, the spiny anteater and the duckbill platypus, are thermally unstable; many of the marsupials, including the opossum, the pouched mouse, and the native cat (a weasel-like spotted marsupial of the family Dasyuridae), are also unable to maintain a fixed body temperature.

The true hibernator not only possesses adaptations that enable it to respond as a homoiothermous animal during certain periods of the year but can also adapt to stressful environmental situations and become essentially a poikilothermous animal during other periods. An animal exposed to food shortages, low temperatures, or lack of water, for example, may "turn off its thermostat" and hibernate until the environment becomes more favourable. Unlike poikilotherms, however, hibernators still retain a measure of temperature control and can change their metabolic levels as required. They can arouse themselves to full activity, whatever the environmental temperature, whereas the arousal of a poikilotherm is dependent upon increased environmental temperatures.

Changes in preparation for hibernation

During the period prior to hibernation an animal must make a considerable number of gradual physiological and metabolic adjustments that appear to be correlated with temperature, light, and the availability of food. No one set of conditions applies equally to all hibernators: some store food, others do not; some become excessively fat, others gain a more moderate amount of weight. Generally, as the season advances and as the hibernator becomes progressively more prepared for hibernation, there is an increase of fat deposition and a general readjustment of body temperature, metabolism, and heart rate to lowered levels of activity.

Although no single factor or condition can be said to determine when an animal will go into hibernation, specific changes include an increase in the quantity of magnesium in the blood and a reduction in the activity of endocrine glands, such as the pituitary, thyroid, and adrenals. A reduction in gonadal activity has also been observed; hibernation does not occur when the gonads are in an actively functional state. Perpetuation of the species requires that the animal be warm and active during the mating and pregnancy periods.

There appears to be a relationship between sleep and hibernation; available evidence suggests that hibernation is entered into from a state of sleep. If hibernation is to be considered a form of sleep, then it must be considered a remarkably complex one. Hibernation and sleep are somewhat similar in that essential body processes continue during both periods at a lowered level. In sleep, the heart beats less rapidly, and breathing is slower; the body produces less heat, necessitating that a sleeping person be protected from the cold.

Hibernation in birds. *Temperature variations.* Birds normally have higher and more variable temperatures than do mammals. Whereas mammalian temperatures normally range between 36° and 39° C (97° and 102° F), avian temperatures range between 37.7° and 43.5° C (99.9° and 110.3° F), with the majority between 40° and 42° C (104° and 108° F). Although the nesting temperature of most passerine species (perching songbirds) is about 40.5° C (104.9° F), primitive bird species—like primitive mammals—have lower temperatures than do the more advanced species. The kiwi, for example, has an average coenothermic body temperature of 37.8° C (100° F). In general, the temperatures of small birds fluctuate more than do those of large birds. The temperature of a house wren (*Troglodytes*) may fluctuate 8° C (14° F) in 24 hours, that of a robin (*Turdus*) fluctuates about 6° C (11° F), and that of the domestic duck only about 1° C (2° F).

The circadian period of activity and rest in birds is accompanied by a temperature cycle. Birds active in the daytime have their highest temperatures late in the afternoon and their lowest in the early morning. Nocturnal species, however, such as owls and the kiwi, have their maximum body temperatures at night, when they are most active. Seasonal temperature variations are also found in birds, and, like mammals, certain birds exhibit thermic instability. Although some are capable of maintaining a highly stable body temperature, others have a fluctuating body temperature. A torpid poorwill (*Phalaenoptilus nuttallii*) is an example of a bird that demonstrates both thermic instability and true hibernation. Its coenothermic body temperature is relatively constant; it can, however, through the influence of a thermoregulatory centre (the hypothalamus) in the floor of the brain, become essentially poikilothermous. Under such influence, its body temperatures approximate those of the environment.

Energy conservation. Considering that hibernation and estivation are devices to avoid such factors as stressful extremes of temperature, lack of water, unavailability of food, or lessened photoperiod, they also must be energy-conservation devices for the animals concerned. Even short periods of torpor can conserve energy. The efficiency of this energy-conservation system can be demonstrated by comparing the smallest bird, the hummingbird, which exhibits circadian torpor, with the shrew, the smallest mammal, which remains active throughout a 24-hour period.

Oxygen consumption

Oxygen consumption is an indicator of metabolic rate, and at an environmental temperature of 24° C (75° F) during the day, an awake but resting hummingbird consumes about 14 millilitres of oxygen per gram per hour. At dusk, the rate drops first to a sleeping level and then plunges to a torpid level of about 0.8 millilitre of oxygen per gram per hour. Just before daybreak, the bird awakens for another activity period. The hummingbird has the highest metabolic rate and the greatest metabolic range of

any vertebrate. The shrew, in contrast, consumes about the same amount of oxygen as the hummingbird does during the day and even increases the amount slightly at night.

The hummingbird uses about 10.3 calories (units of heat energy) during each 24-hour period if it sleeps at night without becoming torpid but only 7.6 calories if it becomes torpid. As it wakes from the torpid state, its temperature increases about 1° C (2° F) per minute to a maximum; the entire process takes less than 30 minutes and sometimes as little as 10 minutes. The energy required to warm the tissues of the hummingbird is relatively small; a hummingbird that weighs four grams expends only 0.114 calorie to warm its body from 10° to 40° C (50° to 104° F). This is only 1/85 of the total 24-hour expenditure of energy of a hummingbird in nature.

The behaviour of the hummingbird can be contrasted to that of a larger bird, such as the poorwill, which is a nocturnal, insect-catching bird. During an average 24-hour day, the poorwill has brief periods of activity at dusk and just before dawn, the total of which is scarcely more than an hour. The temperature of the poorwill during these periods of activity, which are correlated with the bird's feeding habits, is between 40.5° and 43.1° C (104.9° and 109.6° F). Between periods of activity, the bird rests quietly, and its body temperature drops 1° to 3° C (2° to 5° F).

During periods when a supply of flying insects is not available, the bird hibernates in depressions in rocks or other suitably protected places, to which it returns each year. When hibernating, the bird's temperature is frequently within 1° C (2° F) of that of the environment; as a result, the energy saved is great. A poorwill whose body temperature is 5° C (41° F) has a metabolic rate only 3 percent of its metabolic rate at normal body temperature. Because the poorwill is a larger bird than a hummingbird, it may take more than an hour for it to emerge from hibernation.

Hibernation in mammals. It takes longer for larger animals than for smaller ones to go into hibernation because heat must radiate from the body before the temperature can be lowered. Thus, it would require a considerable amount of time for large birds or mammals to go into and emerge from hibernation each day, as do bats and hummingbirds. A 200-kilogram (440-pound) bear, for example, would need 5,100 calories to warm from 10° to 37° C (50° to 99° F). Unlike the hummingbird, which uses only 1/85 of its total daily energy expenditure to emerge from hibernation, the amount expended by a bear would be equivalent to its full 24-hour energy budget. Even if there were enough time in 24 hours for a large animal to enter into and emerge from dormancy, therefore, it would be metabolically extravagant, thus defeating a purpose of hibernation.

Actually, the most common misapplication of the term hibernation is in relation to the bear, which is not a true hibernator. Its body temperature, which normally averages 38° C (100° F), drops during its winter lethargy to about 34° C (93° F), seldom getting below 31.2° C (88.2° F). Hence, a bear's temperature during the winter does not approximate that of the environment. This is indicative of winter rest rather than true hibernation. During this inactive period, the bear sleeps but is, nonetheless, warm and capable of activity when stimulated, unlike a true hibernator. Moreover, it is also during this period when females give birth to cubs that suckle and are maintained by maternal warmth until they emerge from the den in the spring. This behaviour is in contrast with that of the Arctic ground squirrel, whose normal temperature is the same as that of the bear but whose temperature during hibernation drops to near freezing and, in some cases, to a degree or two below 0° C (32° F).

Although certain mammals are said to hibernate, they do not necessarily enter a state of deep hibernation during winter. Instead, for weeks at a time they may be inactive and lethargic in behaviour, with a slightly depressed body temperature. The chipmunk (*Eutamias*) is an example of what has been termed a shallow hibernator, as are bears and raccoons. Superficial hibernation, apparently a compromise between the minimum energy requirements of a deep hibernator and the high energy expended by an animal that remains active during the winter, saves energy without the stress of hibernation. The animal can thus conserve food, while still being able to escape from predators and such dangers as flooding of its burrow. The main energy source during the winter in this shallow hibernation state is food stored in the winter nest. There are instances, however, of shallow hibernators, such as the chipmunk, that enter a state of deep hibernation, particularly if without food.

Superficial hibernation

True mammalian hibernation. Omitting the thermally unstable mammals, the true mammalian hibernators are those whose lowered body temperatures approximate that of the environment and those who require extensive and complex physiological changes to turn from a warm-blooded animal into an essentially cold-blooded one for an appreciable length of time. Only three orders of placental mammals display such behaviour: the Insectivora, as exemplified by the hedgehog; the Chiroptera, the bats; and the Rodentia, including the marmot, hamster, dormouse, hazel mouse, and ground squirrel.

A typical mammalian hibernator, such as the Arctic ground squirrel, finds a protected environmental niche—in this case, a burrow beneath the surface—and builds a nest of grass, hair, and other materials to provide still further insulation. The usual hibernating position is one of being curled up in a ball with the extremities tucked tightly against the body so there is a minimal surface-to-volume ratio. After the temperature of the animal has dropped near that of the ambient temperature, it appears to be dead: its respiration is imperceptible, about three irregular breaths per minute; it does not react to outside stimuli in an observable fashion; nor does it react to being handled and uncurled, although such handling will trigger wakening mechanisms.

The internal organs, such as the digestive tract and the endocrine glands, are almost totally inactive. Because the process of hibernation necessitates the mobilization of all body resources, it places great demands on the tissues, all of which are directed toward the problem of maintaining the animal's metabolism at the minimal level necessary for life during the hibernating period. This means that all activity not immediately germane to the process of living at the lowest possible metabolic level ceases. Even bones and teeth deteriorate during hibernation. The hibernator apparently is balanced on a very narrow line between the maintenance of life at a level that makes recovery from hibernation possible and a reduction of metabolism to a level that will lead to death. Evidence obtained from tissues indicates that the process of hibernation is a precarious method of survival at best and one from which many animals do not awaken. As a mechanism of species survival, hibernation seems effective; for the survival of the individual, however, it is an uncertain and dangerous process.

The hibernator does not remain in a continuous state of hibernation from the time it enters in the fall until it emerges in the spring. Hibernating Arctic ground squirrels, for example, awaken at intervals of every three weeks or less. During this time the animals may move about and sometimes emerge from the burrow. These periods of arousal are more frequent at the beginning and end of a hibernation period than in mid-hibernation; and the lower the temperature at which an animal hibernates, the fewer the awakenings.

During the period of hibernation about 40 percent of the total body weight is lost, an average of about 0.2–0.3 percent per day. One period of arousal and wakefulness consumes more heat and energy than many days in hibernation. About 90 percent of the total heat production and weight loss during hibernation takes place during the arousal periods; only 10 percent is required to maintain the animal in hibernation. Thus, in the case of an unusually long or hard winter, the animal may be called upon to use all of its available energy sources in periodic arousals; it then enters one final hibernation period from which it does not awaken. Animals that store food in the nest have a chance to renew their energy requirements by eating when they awaken periodically.

Types of mammalian hibernators

Entrance into hibernation. Hibernating mammals can be divided into four major groups according to the way they enter hibernation. One group is exemplified by the golden hamster; it waits a variable time of from one to three months in the cold and then enters hibernation in one major temperature reduction. This is accomplished when the biochemical and physiological preparations have been sufficient to lower the animal to a level at which it is receptive to the hibernating stimulus, which causes the abandonment of the temperature differential between ambient and body temperatures.

A second group, of which the pocket mouse (*Perognathus*) is an example, prepares for hibernation relatively rapidly, waiting only a few days before becoming torpid in one major temperature decline. The third group, which constitutes most of the mammalian hibernators, includes ground squirrels and marmots. These animals wait only a few days before entering hibernation and then go through a series of steps of torpor and arousal, each one at successively lower body temperatures, until the level dictated by the stage of preparation for hibernation is reached.

The fourth group, which includes most of the bats, becomes inactive in the poikilothermous manner; that is, the body temperature follows the ambient temperature. Even though the bat seems ready to hibernate at any season, survival during hibernation depends upon more adequate preparation than is necessary for the transitory periods of torpor. Bats not only exhibit true hibernation during the winter but also have natural periods of hypothermia (subnormal temperature), which are unrelated to hibernation, during the remainder of the year.

The woodchuck, the dormouse, and the California ground squirrel enter hibernation in successive stages, with a complete or nearly complete awakening between each one. In the woodchuck, an initial decline in temperature is followed by an arousal. During the second decline there is a lower and more pronounced fall in body temperature, followed by a less pronounced rise. This process continues until the body temperature is essentially the same as that of the environment.

Heart rate and circulation. The body temperature of a hibernating mammal is affected by changes in respiration, heart rate, and oxygen consumption; all are apparently mediated by a part of the nervous system. The heart rate decreases prior to a decline in body temperature. In the woodchuck, the rate may drop from 153 to 68 heartbeats per minute within 30 minutes. In the California ground squirrel, the heart may beat as slowly as once a minute at 5° C (41° F). In contrast, the hearts of non-hibernators generally will not beat at all at temperatures below 10°–20° C (50°–70° F).

Preparation of the Arctic ground squirrel for hibernation

As an Arctic ground squirrel prepares for hibernation, its heart rate and its blood pressure decrease. Both may be detected before a decrease in body temperature can be noted. When the animal enters hibernation, temperatures of both the heart and abdominal regions are identical, indicating an even blood flow between the anterior (front) and posterior (rear) parts of the body. As the body temperature drops, the resistance to blood flow in the peripheral parts of the circulatory system increases because of the increased viscosity (resistance to flow) of the chilled blood and the constriction of the distal arterioles (small arteries) of the body. This peripheral resistance maintains blood pressure at relatively high levels in the deeply hibernating squirrel, even when the heart beats only three or four times a minute.

Neural changes. The nervous system of hibernators also is acclimated; certain specific structures and pathways are seemingly maintained to regulate and coordinate metabolism as temperatures drop. This adaptation of the nervous system enables changes in the environment to be perceived, even when the animal is torpid. In the Arctic ground squirrel, measurements of the general electrical activity of the brain indicate a 90 percent reduction when the animal is in hibernation, at which time brain temperatures approximate 6° C (43° F). During hibernation, both the peripheral nervous system (all the nerves outside the brain and spinal cord, which constitute the central nervous system) and the spinal cord have an increased sensitivity to certain stimuli; in addition, the areas of the brain that regulate temperature as well as cardiac (heart) and respiratory function remain active at ambient temperatures, below which the mammalian nervous system normally ceases to function.

Changes in the circulatory system involving constriction (narrowing) of posterior vessels and the favouring of anterior circulation allow the brain temperature of hibernators to remain a few degrees warmer than the environmental level. This enables the temperature of the brain to remain constant despite fluctuations in the temperature of the skin.

Endocrine activity. The male sex hormone testosterone stimulates reproductive activity. The golden hamster will not hibernate if injected with more than five milligrams of a hormonal preparation. Hibernation is also prevented if the animal is fed or injected with thyroid hormones or thyroid-stimulating extracts. The latter would seem to implicate the thyroid as another endocrine gland that plays an important role in hibernation. There is, in fact, a seasonal progression and regression of thyroid activity in hibernators; maximal activity occurs in the spring and minimal activity in the fall. And because hibernation does not take place in the absence of the adrenal glands, it appears that a minimal adrenal activity is also necessary for hiberation and survival.

Importance of timing

The importance of timing in the annual rhythm of activity and dormancy can be demonstrated: when hibernators are exposed to cold temperatures in spring and summer, they react as do all homoiotherms by increasing their thyroid activity and metabolic rate to maintain normal body temperature. But if they are exposed to cold temperatures in the fall, the thyroid activity and metabolic rate of hibernators are lowered. In some species, a combination of decreased food and lower ambient temperature is required to reduce activity of the thyroid gland and to produce hibernation, although cold alone is sufficient in ground squirrels and the dormouse.

Although hibernation does not take place during periods of gonadal activity or stimulated thyroid activity, it can occur during increased activity of the pituitary gland. This would suggest that there is a dissociation of cellular growth and hormone synthesis that is normally controlled by hormone secretion of the pituitary and its target organs. Thus, the triggering mechanism for the resumption of normal endocrine activity apparently resides elsewhere than in the pituitary. The function of the hypothalamic region of the brain in regulating appetite, fat deposition, water intake, and diuresis (increased excretion of urine), as well as in the control of temperature and sleep, would appear to make it a key area in directing life processes of the hibernator. Furthermore, the fact that the hypothalamus regulates the pituitary and other endocrine glands not only supports this thesis but also indicates that this area of the brain is the prime, or master, regulator of the entire hibernation process.

Reproductive cycles. The Arctic ground squirrel may spend more than half its life in hibernation (see Figure 23). It thus must be able to breed, rear young, maintain its home burrow, and prepare for the period of hibernation during an activity period of less than six months. This requires considerable adaptation of both metabolic and behavioral patterns. Prior to entering hibernation in late September or early October, there is a renewal of sexual activity in the testes of males, and, throughout the period of hibernation, they continue to grow. On the Arctic slope in early May, the male ground squirrel emerges from its burrow. As it utilizes the remaining fat and eats the stores of seeds and other food still in the nest, the male reaches a period of reproductive readiness. Mating takes place in the middle of May, and the young are born in the middle of June, after a gestation period of about 25 days. By the middle of July the young are above ground and eating the green Arctic vegetation, which they continue to eat until the onset of hibernation. By October, both the young of the year and the adults from the previous year weigh nearly 1,000 grams (2.2 pounds).

In the bat, the reproductive cycle is interrupted by hibernation. Gonadal activity in the male reaches its maximum

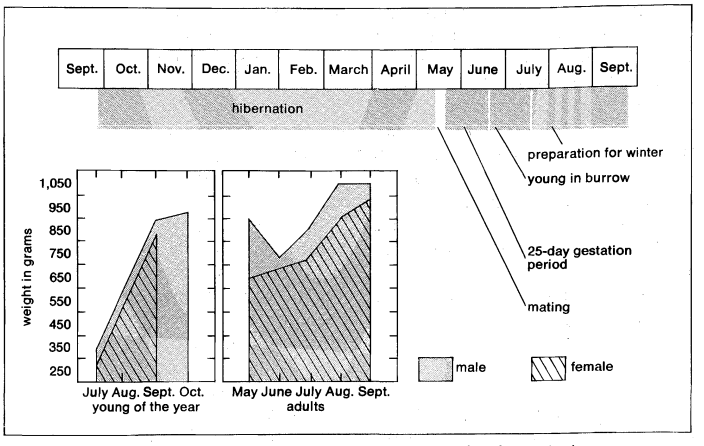


Figure 23: The annual cycle (top) in the life of a typical mammalian hibernator, the Arctic ground squirrel. The graphs show typical weight patterns of young and adult males and females during the aboveground activity period (see text).
From W.V. Mayer, *Hibernation*

in the fall, when copulation with the female occurs. The animals then hibernate, and the production of sperm in the male ceases. The sperm deposited in the female are stored in her reproductive tract throughout the period of hibernation; fertilization occurs the next spring, when the eggs are ovulated (released from the ovaries) within a few days after awakening from hibernation.

Delayed fertilization

The only exception to the general hibernation–reproduction pattern of bats is the vespertilionid bat (*Miniopterus*), in which there is no delayed ovulation and fertilization. In this species the eggs are ovulated soon after copulation, in the fall, and fertilized immediately. During the ensuing period of hibernation embryonic development is initiated and slowed, but it does not actually cease. The young are born in the early summer, soon after hibernation ends. The introduction of hibernation during pregnancy makes the gestation period several months longer than in non-hibernating tropical members of the same genus.

Cyclical reproductive activity has thus become adapted to the shortened activity season available to the hibernator. But although the annual sequence of reproductive events is known, the external stimuli that regulate the reproductive cycles of bats and other hibernators are not known. More knowledge is needed concerning the endocrine and nervous mechanisms that presumably regulate reproductive processes internally. It has been suggested that the pituitary–gonadal relationship influences the hibernating cycles as well as the reproductive cycle, hence both the latter and homoiothermism are controlled by a common mechanism. Such a suggestion is attractive in that the mechanism solves the regulation problems, but more needs to be known of the way in which hibernation directly or indirectly modifies the action of endocrine and neural mechanisms that direct the reproductive cycle.

Protection from disease and radiation. Hibernating organisms have a certain degree of resistance to infectious diseases that appears to be attributable to at least three factors, all of which are related to temperature. One is the fact that the lowered temperature of the host and the commensurate slowing of its metabolic processes prevent the multiplication of parasites to a greater extent than they affect the host's defensive mechanisms. Second, lower temperatures are more harmful to the development of a disease organism than to the host, as has been shown with the parasite *Trichinella spiralis*. In bats hibernating at 5° C (41° F), only larvae have been recovered from the intestines; but mature adult worms have been recovered from the intestines of bats kept at 35° C (95° F). The third factor is that the influence of low temperature on the chemical composition of the host tissues may also affect infectious organisms.

Hibernation also seems to protect animals from radiation. When ground squirrels are irradiated with radioactive cobalt while hibernating, they are found to be more resistant to the effects of the radiation than are squirrels irradiated while warm and active. This resistance, which is apparent over a wide range of doses, suggests that protective mechanisms function in the hibernating animal. In both hibernating and non-hibernating animals, repair processes within cells occur the first day after irradiation; however, when the metabolic requirements of cells are small, as in hibernation, the injured cells seem to be more capable of repair, and survival is greater. The large metabolic requirements imposed on injured cells of warm and active animals appear to render them incapable of an adequate repair response.

Awakening from hibernation. The process of awakening in the Arctic ground squirrel takes about three hours. There is a rapid rise in heartbeat and a decrease in peripheral circulatory resistance; the area around the head and heart warms more rapidly than the posterior part of the animal. This differential vasodilatation (widening of the blood vessels) in the anterior part of the body is a unique and vital part of the awakening process. The concentration of active circulation in this region results in a high blood pressure and an efficient and rapid warming. If a drug is administered during awakening that causes vasodilatation throughout the body, there is a marked drop in blood pressure even though the heart may almost double its rate; thus, the heart cannot maintain a high blood pressure at this time if all blood vessels are dilated. Later during the arousal process, after the anterior part of the body has been warmed, the posterior part of the animal warms rapidly.

Differential vasodilatation

Despite the deterioration of glands and tissues and the drastic reduction of all metabolic activity during hibernation, within 24 hours after arousal, all the squirrel's physiological processes are essentially normal. This rapid repair and recovery mechanism is one that requires further study. (W.V.M./Ed.)

Reproductive behaviour

Reproductive behaviour in animals includes all the events and actions that are directly involved in the process by which an organism generates at least one replacement of itself. In an evolutionary sense, the goal of an individual in reproduction is not to perpetuate the population or the species; rather, relative to the other members of its population, it is to maximize the representation of its own genetic characteristics in the next generation. The dominant form of reproductive behaviour for achieving this purpose is sexual rather than asexual, although it is easier mechanically for an organism simply to divide into two or more individuals. Even many of the organisms that do exactly this—and they are not all the so-called primitive forms—every so often intersperse their normal asexual pattern with sexual reproduction.

BASIC CONCEPTS AND FEATURES

The dominance of sexual reproduction. Two explanations have been given for the dominance of sexual reproduction. Both are related to the fact that the environment in which an organism lives changes in location and through time; the evolutionary success of the organism is determined by how well it adapts to such changes. The physiological and morphological aspects of an organism that interact with the environment are governed by the organism's germ plasm—the genetic materials that determine hereditary characteristics. Unlike asexual methods, sexual reproduction allows the reshuffling of the genetic material, both within and between individuals of one generation, resulting in the potential for an extraordinary array of offspring, each with a genetic makeup different from that of its parents.

Sexual reproduction and adaptation

According to proponents of the so-called long-term theory for the dominance of sexual reproduction, sexual reproduction will replace asexual reproduction in the evolutionary development of an organism because it assures greater genetic variability, which is necessary if the species is to keep pace with its changing environment. According to proponents of the short-term theory, however, the above argument implies that natural selection acts on groups of organisms rather than on individuals, which is

contrary to the Darwinian concept of natural selection (see EVOLUTION, THE THEORY OF: *Natural selection*). They prefer to view the advantages of sexual reproduction on a more immediate and individual level: an organism employing sexual reproduction has an advantage over one employing asexual means because the greater variety of offspring produced by the former results in a larger number of genes being transmitted to the next generation. The latter view is probably more nearly correct, especially in violently fluctuating and unpredictable environments. The former theory is probably correct when viewed in terms of its advantage to individuals that are spreading in geographic range, thereby increasing the likelihood of encountering different environments.

Natural selection and reproductive behaviour. Natural selection places a premium on the evolution of those physiological, morphological, and behavioral adaptations that will increase the efficiency of the exchange of genetic materials between individuals. Organisms will also evolve mechanisms for sensing whether or not the environment is always permissive for reproduction or if some times are better than others. This involves not only the evolution of environmental sensors but also the concurrent evolution of mechanisms by which this information can be processed and acted upon. Because all seasons are not usually equally conducive, individuals whose genetic backgrounds result in their reproducing at a more favourable rather than less favourable period will eventually dominate succeeding generations. This is the basis for the seasonality of reproduction among most animal species.

Natural selection also results in the evolution of systems for transmitting and receiving information that will increase the efficiency of two individuals' finding each other. These attraction systems are usually, but not always, species specific (see EVOLUTION, THE THEORY OF: *Species and speciation*). Once the proper individuals have found each other, it is clearly important that they are both in a state of reproductive readiness. That their sensory receptors are tuned to the same environmental stimuli is usually sufficient to achieve this synchrony (proper timing) in the lower organisms. Apparently, however, this is not enough in the more complex organisms, in which the fine tuning for reproductive synchrony is accomplished chiefly by a process called courtship. Another evolutionary necessity is a mechanism that will guide the partners into the proper orientation for efficient copulation. Such mechanisms are necessary for both internal and external fertilization, especially the latter, where improper orientation could result in a complete waste of the eggs and sperm.

Need for genetic variability

In most organisms, the period of greatest mortality occurs between birth or hatching and the attainment of maturity. Thus, it is not surprising that some of the most elaborate evolutionary adaptations of an organism are revealed during this period. Natural selection has favoured an enormous variety of behaviour in both parents and offspring that serves to ensure the maximum survival of the young to maturity. In some animals this involves not only protecting the young against environmental vicissitudes and providing them with adequate nutrition but also giving them, in a more or less active manner, the information they will need to reproduce in turn.

EXTERNAL AND INTERNAL INFLUENCES

As mentioned at the beginning of this discussion, the anatomical, physiological, and neurological aspects of reproduction and behaviour are dealt with in other articles. It is useful here, however, to consider briefly the external and internal factors that initiate reproductive behaviour.

Environmental influences. Light, usually in the form of increasing day length, seems to be the major environmental stimulus for most vertebrates and many invertebrates, especially those living in areas away from the Equator. That this should be such an important factor is quite reasonable in an evolutionary sense: increasing day length signifies the onset of a favourable period for reproduction. In equatorial regions, where changes in day length are usually insignificant throughout the year, other environmental stimuli, such as rain, predominate.

Superimposed on day length are usually several other factors, which, if lacking, often override the stimulating effect of light. Many insects, for example, will not initiate a reproductive cycle if they lack certain protein foods. Many animal groups have an internal cycle of cellular activity that must coincide with the external factors before reproduction can occur; a familiar example is the estrous cycle in most mammals except primates. Females are sexually receptive only during a brief period when they have ovulated (released an egg from the ovary).

Hormonal influences. Although the exact way by which light affects the reproductive cycle is still disputed, it undoubtedly varies from group to group. In birds, light passes either through the eyes or through the bony tissue of the skull and stimulates the development of certain cells in the forepart of the brain. These cells then secrete a substance that stimulates the anterior pituitary gland, which is located at the base of the brain, to produce an array of regulatory substances (hormones), called gonadotropins, that are carried by the blood to the gonads (ovaries and testes), where they directly stimulate the development of eggs and sperm. The gonads, in turn, produce the sex hormones—estrogen in the female and testosterone in the male—that directly control several overt aspects of reproductive behaviour.

Insect pheromones

Unlike the higher animals, the gonads of insects apparently do not themselves secrete hormones. Instead, stimulation by the corpus allatum, an organ in insects that corresponds in function to the pituitary gland, causes the secretion of liquid substances on the body surface. These substances are transmitted as liquids, or, even more significantly, as gases, to the recipient, in which they are usually detected by olfaction or taste. Such substances, which are called ectohormones, or pheromones, may serve as the major regulation and communication system for reproduction as well as other behaviour in insects.

In the absence of all other stimuli, many types of sexual behaviour can be induced simply by an injection of the appropriate gonadal hormone. Conversely, removal of the gonads usually inhibits most sexual behaviour. The apparent failure of complete hormonal control over reproductive behaviour has been a subject of much investigation and dispute. There is much evidence that many types of reproductive behaviour are or can be controlled solely by neural mechanisms, bypassing the hormonal system and any effect that it might exert on the nervous system to produce behaviour. Several types of reproductive behaviour controlled solely or almost solely by neural mechanisms are involved in or triggered by the processes that are initiated by courtship.

MODES OF SEXUAL ATTRACTION

The chief clues by which organisms advertise their readiness to engage in reproductive activity are visual, auditory, and olfactory in nature. Most animals use a combination of two modes; sometimes all three are used.

Visual clues. The appearance of many higher vertebrates changes with the onset of reproductive activity. The so-called prenuptial molt in many male birds results in the attainment of the nuptial plumage, which often differs radically from that possessed by the bird at other times of the year or from that possessed by a nonreproductive individual. The hindquarters of female baboons become bright red in colour, which indicates, or advertises, the fact that she is in estrus and sexually receptive. Such changes in appearance are less common in the lower animals but do occur in many fishes, crabs, and cephalopods (*e.g.*, squids and octopuses).

The role of aggressive behaviour

Often associated with changes in appearance are changes in behaviour, particularly the increase in aggressive behaviour between males, often a prime feature in attracting females; such changes have interesting evolutionary implications. In certain grouse, for example, females are most attracted to males that engage in the greatest amount of fighting. No doubt, fighting in some groups of mammals also serves this function as well as others (see above *Aggressive behaviour*).

In many animals the rise in aggression takes the form of territoriality, in which an individual, usually a male, defends a particular location or territory by excluding from it

all other males of his own kind. Occasionally, other species are also excluded when it is to the advantage of the defending individual to do so. Territorial behaviour involves many functions, not all of which are directly concerned with reproduction. For purposes of advertising, however, territoriality probably reduces the amount of interference between males and also makes it easier for females to find males at the proper time.

Auditory clues. The fact that sound signals can travel around barriers, whereas visual signals cannot, accounts for their widespread use in indicating sexual receptiveness, especially in frogs, insects, and birds. Like visual signals, a sound for advertising purposes usually encodes several pieces of information; for example, the signals usually reveal to the receiver the caller's species, its sex, and, in some cases, whether or not it is mated. The vocalizations of one type of frog also reveal the number of other males located nearby. This information, a critical clue for females, is a measure of how good the habitat is for depositing eggs. The sounds produced by the wings of mosquitoes attract females and are species specific. Humans have taken advantage of this signal by using artificial sound generators to eradicate certain mosquitoes. Advertising signals also serve to repel other males; a classical example is the territorial song of many songbirds.

Olfactory clues. Researchers have now become aware of the enormous amount of information that is passed between animals by chemical means. Well known are the urine, feces, and scent markings employed by most mammals to delimit their breeding territories and to advertise their sexual state. Males of a number of mammals are capable of determining if a female will be sexually receptive simply by smelling her urine markings. A substance in the urine of male mice, on the other hand, actually induces and accelerates the estrous cycle of females. A female gypsy moth is able to attract males thousands of metres downwind of it simply by releasing minute quantities of its sex pheromone each second. It has been calculated that one female silkworm moth carries only about 1.5 micrograms (0.0015 gram) of its sex attractant, called bombykol, at any given moment; theoretically, this is enough to activate more than 1,000,000,000 males, surely more than exist in any one place at any time. The sex attractant of barnacles, which are otherwise rather sessile (sedentary) organisms, causes individuals to aggregate during the breeding period.

The potency of pheromones

One other possible channel of communication occurs in a few fishes, namely electric discharge. Evidence now accumulating suggests that weak electrical fields and discharges in the Mormyridae of Africa and the Gymnotidae of South America represent the major mode of social interaction in these families.

COURTSHIP

Synchrony is the major factor in achieving fertilization in the lower animals, particularly in aquatic forms. In most of these groups, the eggs and sperm are simply discharged into the surrounding water, and fertilization occurs externally. The parents may never meet, so to speak. It might be assumed that this procedure would be roughly the same in the higher animals, with perhaps more overt behaviour to achieve synchrony, and that, after the two individuals found each other, fertilization would proceed fairly quickly. This is usually not the case, however. Although fertilization in the higher terrestrial forms involves contact during copulation, it has been suggested that all of the higher animals may have a strong aversion to bodily contact. This aversion is no doubt an antipredator mechanism: close bodily contact signifies being caught. Since females are in an especially helpless situation during copulation, they are particularly wary about bodily contact. In addition, males are particularly aggressive during the breeding period, which further increases the uncertainty of both individuals. These difficulties were solved by the evolution of a collection of behaviours called courtship. Courtship has been defined as the heterosexual reproductive communication system leading to the consummatory sexual act.

Courtship behaviour has many advantages and functions, not the least of which is the reduction of hostility between the potential sex partners, especially in species in which the male actively defends a territory. The major aspects of such behaviour seem to be appearance, persistence, appeasement, persuasion, and even deception. Because courtship behaviour involves the transmission of information by means of signals, it is useful to define at this point a peculiar and important group of social signals called displays.

A social signal may be considered any behavioral pattern that effectively conveys information from one individual to another. The term display has been restricted by some authorities to social signals that not only convey information but that, in the course of evolution, have also become "ritualized." In other words, such signals have become so specialized and exaggerated in form or function that they expressly facilitate a certain type of communication. The visual, auditory, olfactory, tactile, or other patterns by which organisms advertise their readiness to engage in reproductive activity provide examples of displays. Clearly, the kinds of displays utilized by organisms depend on the sensory receptors of the receiver. Whereas higher vertebrates tend to use visual and auditory displays, insects tend toward olfactory and tactile displays.

Types of display

In animals in which the male takes on a wholly different appearance during the breeding period, natural selection has eliminated from the female's appearance the "aggressive badges" of males that provoke fighting. It is not without significance that the appearance of the adult female in many species is much like that of the juvenile; this implies to the male a friendly, nonaggressive relationship. When one male approaches another that has intruded into the former's territory, the outsider may either return the aggressive display or flee. Females, however, usually quietly back up slightly and then slowly move forward again. With each approach the male's hostility lessens toward this appeasing, increasingly familiar individual. Often, as in many birds, the females resort to displays that resemble the food-begging behaviour normally seen in the young. Males frequently respond to this display by actually regurgitating food. Male spiders of some species offer the larger and more aggressive females food as bait, and copulation occurs while the female is eating the food rather than her potential mate. Mutual feeding displays, often with nonedible items, are engaged in by a number of insects and birds. In the courtship behaviour of several birds, extremely elaborate displays are utilized to hide the bill from the potential partner, because the bills of these birds are their chief weapons. Some aspects of nest building have been incorporated into the displays of such birds as penguins. Early in the relationship between the individuals, one or both may offer the other stones that are placed in a pile. The actual nest is not constructed until much later, however.

The common denominator in courtships is that the displays resemble functional behaviours that are appropriate to friendly, bonded situations, such as those between parents and between parents and their offspring. The degree of elaborateness of the display is governed by a number of factors. One is to prevent cross-mating between different species, an occurrence that usually results in the waste of the eggs and sperm. Any specific aspect—*i.e.*, one or more displays—used by an organism in species discrimination is called an isolating mechanism. In many species, the majority of the displays between individuals are a series of identity checks.

Species discrimination

Another factor that has an impact upon the complexity of displays is the length of time that the pair bond will endure. Brief relationships are usually, but not always, associated with rather simple courtship activity. In a number of insects, birds, and mammals, the males display on a common courtship ground called a lek or an arena. Females visit these courtship areas, copulate, and leave. The males do not participate in any aspect of parental care; the bond lasts but a few seconds. Yet despite the brevity of this relationship, in no other courtship system is there the development of such elaborate and almost fantastic displays in both the movements and appearances of the courting males.

POST-FERTILIZATION BEHAVIOUR

Various types of behaviour ensure that a maximum number of fertilized eggs or young will survive to become reproductive adults. Clearly, the number of eggs produced and their size represents a balance achieved by natural selection. This balance conforms to some optimum compromise between producing many eggs containing little food for the development of young or fewer eggs with more provisions.

There has been considerable controversy about the factors that limit the number of offspring an organism can produce. It has been suggested that, among animals in which the offspring are dependent on the parents for varying lengths of time, clutch or litter size has been adjusted through natural selection to the maximum number of offspring that the parents, on the average, can feed. There are, on the other hand, organisms that do not practice parental care and produce millions of eggs. According to one school of thought, these species have such a high fecundity (productivity) because the eggs and larvae suffer a very high mortality rate. Hence, it is necessary for such animals to produce thousands, even millions, of eggs just to obtain a few reproductive adults. An opposing school of thought, however, says that such species have high mortality rates because of their great fecundities. By similar reasoning, low death rates would be the consequence of low fecundity.

Protective adaptations. A number of adaptations have evolved to protect the eggs and larvae of species not attended by adults. In one such adaptation, the eggs or larvae are distasteful, inedible, or apparently harmful to potential enemies. The eggs of the jellyfish *Bougainvillia,* for example, contain stinging cells on the surface that deter predators. Many female butterflies deposit their eggs on plants that contain poisonous compounds, which the larvae incorporate into their bodies, making them distasteful. When disturbed many insect larvae, especially those that are camouflaged, give a so-called startle display; several caterpillars, for example, raise their heads as if to bite or their hindparts, in the manner of a wasp, as if to sting. Others suddenly present striking colour patterns previously hidden. Most of these displays have been shown experimentally to be effective deterrents against predators.

Caring for offspring. Animals that do not care for their young must provide for the nutritional needs of their offspring. One way of doing so is by producing an egg with a sufficiently large yolk supply that the young, when hatched, are already at an advanced, almost independent state. A peculiar example of this is found in the incubator birds (Megapodiidae), which cover their large eggs with soil and debris to create a mound of considerable depth, effectively providing heat for the developing eggs. After a very long incubation period, the young emerge as fully feathered miniature adults and are capable of flying in 24 hours. Before sealing the nest that they make for their eggs, many insects, such as certain solitary wasps, stock the nest with food. In a more bizarre manner, other solitary wasps place one egg in the body of an insect or spider previously paralyzed by the wasp. Upon hatching, the larva eats the still living host.

Social parasitism

Social parasitism, another fascinating aspect of post-fertilization behaviour, is found in certain insects and birds. In this case, the true parents do not care for their eggs or offspring; rather, they place them under the foster care of other species, often, but not always, to the detriment of the foster parents' offspring. In certain parasitic species of cuckoos, the females are divided into groups, or gentes, each of which lays eggs with a colour and pattern unlike those of the other groups. The females of each group usually select a particular species as the host, and, more often than not, the eggs of the parasite closely resemble those of the potential foster parent. This mimicry has evolved because many host species throw eggs not resembling their own out of the nest. Some young cuckoos also exhibit a behaviour called backing, in which they push out the other nestlings and monopolize the food supply.

Parental care. Among the organisms that remain with the eggs or offspring, one particular behaviour is striking—that of nest construction to keep the eggs and larvae in one spot and to protect them against predators as well as such environmental factors as sun and rain. The placement of a nest usually serves an antipredatory purpose, as in birds that put their nests near those of social wasps or stinging ants. Although they are not normally thought to do so, many mammals, particularly rodents and carnivores, construct special nests, dens, or burrows solely for reproductive purposes.

A number of fishes build nests made of bubbles that not only hold the eggs together but also provide the oxygen necessary for the developing embryos. Other fishes, particularly those that live in oxygen-poor waters, display elaborate fanning behaviour to keep the water moving around the eggs. In some fishes, the female incubates the egg in her mouth, thus providing protection against predators as well as constant aeration. The fry (young) of some of these mouthbreeders travel in a school near the parent. When danger approaches, they flee into the parent's mouth and later swim out after the danger passes.

Birds have the problem of keeping the eggs at an optimum temperature for development of the embryo. With the onset of egg laying in many species, the feathers of the lower abdomen are lost, and the skin in that area becomes thickened and highly vascularized (filled with blood vessels), forming the so-called brood patches. Usually the female develops these patches, which serve to transfer more effectively to the eggs the warmth from the adult's body. It has been shown that, like much of parental behaviour in the higher vertebrates, brood patches and "broodiness" are controlled by several hormones, combined with visual and tactile stimuli. Chief among these hormones is prolactin, which also controls the production of pigeon milk, a cheeselike substance produced only in the crops of adult doves and pigeons and fed to the nestlings by regurgitation.

Although there are some outstanding exceptions, most young mammals are completely helpless at birth. This helplessness is most striking in the marsupials (*e.g.,* opossums and kangaroos), in which the young are born at a very early stage of development; they crawl through the mother's hair to the brood pouch, where they attach themselves to a nipple and their development continues for many more months.

Licking of newborn

An early characteristic behaviour in mammals following birth is that of the mother licking the newborn. This serves at least two functions—one is general cleanliness to avoid infections or the attraction of parasites; the other would appear to be purely social. If a newborn mammal is removed from its mother and cleaned elsewhere before she can lick it, she usually will not accept it. Thus, licking behaviour also serves, in some manner, to establish a unique relationship between the mother and her offspring. Another characteristic mammalian behaviour is the suckling response of the newborn. Although this behaviour has been claimed to be the perfect instinctive response, it apparently is not so in many species; the trial-and-error period during which the newborn discovers the nipple, however, is quite short.

In birds, especially those that nest on the ground, one of the first adult responses to the hatching of the eggs is to remove the conspicuous eggshells from the area of the nest. It has been shown experimentally that, in gulls at least, this is an important antipredatory measure. When birds hatch, they have the ability to stretch their heads and to gape for food in response to any mechanical disturbance, such as that produced when the parent lands on the nest. Later in development, they stretch and gape only when the parents appear. This is another type of adaptive, antipredatory behaviour, as it would be dangerous for the nestlings to gape and vocalize in response to any environmental disturbance.

Group care. The ability of an animal to identify its own offspring at an early stage is apparently not important in animals that nest or are solitary breeders; offspring in the nest belong to that parent. In colonially breeding species or in those where the offspring of different parents are likely to become mixed, however, natural selection has favoured the evolutionary development of behaviour that makes possible the recognition by the parent of its own

offspring, thereby avoiding the danger of expending energy on offspring that do not possess the parent's genes.

There is, on the other hand, the situation in which the offspring are cared for by individuals who are not the parents. This phenomenon occurs among the social insects in particular and also among several groups of birds and mammals; future investigations may show it to be even more widespread. In such birds as the anis, the effective breeding group consists of several females and males. One nest is constructed in which all the females deposit their eggs, and all individuals participate in the care of the resulting offspring. In certain jays (Corvidae), the offspring of one generation participate in the care of the offspring of the next or another generation, but the exact family relationships among the participants are not clear.

In the social insects, this type of parental behaviour apparently results from the peculiar genetic relationships between the individuals in most social-insect colonies (termites are among the exceptions). The female and, in the termites, both the male and the female can control by chemical means the kinds (called castes in ants and termites) and sexes of the offspring. An outstanding feature of such colonial insects as the honeybee is that the majority of the individuals produced by the queen are sterile; these are the workers, the individuals who care for and feed both the queen and her offspring, the sibs of the workers.

Group care in honeybees

The queen is diploid in genetic makeup; that is to say, half of her genes are derived from her mother and half from her father. The males (drones) are haploid; that is, they have only half the genes possessed by the queen, all of them derived from the mother. A queen produces eggs fertilized by sperm she has retained in her body from the mating flight; thus the individuals produced are diploid, but, unlike the queen, they are sterile. This sterility results indirectly from a chemical secreted by the queen, called the queen substance. It inhibits the workers from building special brood cells that give rise to sexually developed individuals. If the queen fails to secrete this substance because of age or death, the workers immediately construct special brood cells with a substance they secrete; called royal jelly, it is necessary for the development of a larva then destined to be a queen.

How can the evolution of sterility in workers and their care of offspring not their own be accounted for? One possible explanation concerns the coefficient of relationship (the number of genes on the average shared in common) among the individuals of a colony. Because of the peculiar haplo-diploid mode of sex determination, the workers (sisters) share all the genes from their father and, on the average, half of those from their mother. Since each worker receives half of its genes from the father and half from the mother, the average genes shared between any two workers (sisters) is three-fourths. But between mother (the queen) and daughter (a worker) this average is only one-half. The offspring (the sterile workers), therefore, may contribute more to their fitness (the maximum representation of their genes in the next generation) by caring for their sisters than by providing an equal amount of care to their "own" offspring, had they been fertile rather than sterile. A drone, on the other hand, has a coefficient of relationship with one of his sterile sisters of only one-fourth, but retains a relationship of one-half with his mother and daughters (future sterile workers). This explains why workers provide more care for their sisters than for their brothers, and why the workers eventually drive off the almost useless drones, which are relatively scarce (having resulted from unfertilized eggs), from the colony. Because sisters share more genes with each other than with their brothers, they maximize the chances of these genes surviving into the next generation by providing more care for their sisters.

This explanation of group care and extreme sociality does not account for all cases. Indeed, termites are perhaps the most extreme among animals in these respects but lack the haplo-diploid sex determination mechanism. In addition, several groups having this mechanism have not evolved extreme brood care and sociality. Other factors have to interact for these systems to evolve, but it is not yet clear what they are.

REPRODUCTIVE BEHAVIOUR IN INVERTEBRATES

Protozoans and sponges. Most protozoans (one-celled organisms) reproduce asexually, usually by fission (splitting in two); in some species, however, sexual as well as asexual reproduction occurs and may be complex. The colonial organism *Volvox,* which may be either of one "sex" or composed of cells of both sexes, produces true eggs and sperm. A chemical substance released by "females" induces the production of sperm packets; following the union of the egg and sperm, the parent colony dissolves, and the zygote (fertilized egg) is released.

Conjugation

Another form of reproduction in protozoans is conjugation, in which organisms such as *Paramecium* fuse together briefly to exchange nuclear products. This results in a reshuffling of hereditary characteristics just as occurs in true sexual reproduction in higher animals. In some species of *Paramecium,* there are mating types, and an individual is of one type or the other. Opposite types apparently recognize each other by a chemical (pheromone) that is released on their body.

In the lower metazoans (multicellular organisms), reproduction is also by both asexual and sexual means. As befits their sessile life-style and low population densities, sponges that reproduce sexually are usually hermaphroditic; that is, each individual is capable of producing both sperm and eggs, but often at different times to prevent self-fertilization. The sperm are swept by water currents into another sponge, where they are picked up by specialized cells called choanocytes and carried to the egg. Fertilization takes place when a choanocyte fuses with the egg. The free-swimming larval stage that is produced is of short duration, after which the organism settles on the bottom and becomes a new adult sponge.

Coelenterates. Hydroids, jellyfishes, sea anemones, and corals of the phylum Coelenterata, or Cnidaria, reproduce by a variety of mechanisms. A familiar coelenterate animal, the freshwater *Hydra,* usually reproduces asexually by budding, a process by which small portions of the adult structure become new, but genetically identical, individuals. Hydras are also dioecious; that is, each individual produces either sperm or eggs. In many temperate-zone species of *Hydra,* sexual reproduction occurs during the autumn; the fertilized eggs enable the species to survive the winter.

Most of the other hydrozoans are colonial organisms, often occurring in polyp and medusal (umbrella-shaped) forms. In a colony, reproductive individuals called gonophores develop into free-swimming organisms (medusae) that reproduce sexually. Fertilization can be either external or internal; if external, the eggs are shed directly into the water. Internal fertilization results in larvae that swim out of the parent and soon settle on a surface, where they develop into another hydroid colony.

Sea anemones and the polyps of corals reproduce both asexually—by budding—or sexually. In the sexual mode, sea anemones have both dioecious and hermaphroditic species. One interesting aspect of sea anemones, which undergo internal fertilization, is that they are among the first lower animals known to provide parental care. The larvae of sea anemones remain inside the adult until they are ready to metamorphose (change in form), at which time they swim from the parent's mouth and settle on its base, remaining there until they develop tentacles. When they have reached this stage of development, they move away from the parent's protection.

Flatworms and rotifers. The reproductive structures of flatworms (phylum Platyhelminthes) resemble those found in the higher groups. Such flatworms as the land and freshwater planarians are hemaphrodites. Although some species can reproduce asexually by splitting in two, most engage in copulation. Some freshwater planarians can produce both thin-shelled summer eggs, which hatch in a short time, and thick-shelled winter eggs, which are resistant to freezing and hatch in the spring. An apparently unique situation in many planarians is that nutrition for the embryo is supplied by the addition of separate cells to the zygote, after which the entire mass is enclosed in the shell; more commonly, the yolk is incorporated within the structure of the zygote itself.

In the rotifers (phylum Aschelminthes), small but abundant freshwater animals, reproduction is usually sexual, and the sexes are separate. Copulation occurs by injection of sperm anywhere in the body wall of the female. Many species found in temporary ponds and streams exhibit a peculiar reproductive behaviour that is well adapted to their transient environment: they produce different kinds of eggs at different times of the year. One egg type, called amictic, is produced in the early spring. These eggs apparently cannot be fertilized, and the embryo develops without fertilization (parthenogenesis); the result is females with a life-span no longer than two weeks. When the population reaches a peak in the early summer, a second type of egg is produced. If unfertilized, this egg, which is called mictic, results in males. As the male population increases, most mictic eggs become fertilized, resulting in the production of a heavy-shelled dormant egg with much yolk. The dormant egg survives the winter and gives rise to the amictic females of the next spring. Thus, despite the many generations produced in the summer by so-called sexual means, the reshuffling and recombination of genetic material occurs only once a year.

Parthenogenesis

Segmented worms. The marine worms of the class Polychaeta (*e.g.,* clam worms and lugworms of the phylum Annelida) provide the first examples of a kind of courtship behaviour involving both visual and chemical displays initiated by some rather subtle environmental stimuli. Most polychaetes reproduce sexually, and there are two distinct sexes in most species. Either by transformation or budding, many polychaetes produce a reproductive form (epitoke). At a certain time of the year, the epitokes swarm to the ocean surface and engage in mass shedding of eggs and sperm. Some female epitokes of clam worms (*Nereis*) produce a chemical substance called fertilizin that attracts the male epitokes and stimulates the shedding of sperm. Male epitokes of a polychaete found in the Atlantic Ocean emit a flashing light; females emit a steady light. The light may serve to attract male and female and to aid in species discrimination. The swarming of the palolo worm *Palola* in parts of the South Pacific is apparently triggered by an annual and a lunar cycle; the epitokes separate from the parent (atoke) in October or November, during the last part of the lunar cycle.

The class Oligochaeta (phylum Annelida) contains a diversity of both aquatic and terrestrial worms, among which is the familiar earthworm, *Lumbricus.* Although some aquatic oligochaetes reproduce asexually, the majority are sexual, and all of these are hermaphrodites. At mating, two oligochaetes lie side by side so that the head of one is opposite the tail of the other. Sperm then pass reciprocally into small sacs, where they are temporarily stored. This transfer is more complex in the earthworms, however, because the respective male pores are not in direct opposition; each individual forms a temporary skin canal through which the sperm flow to their respective sacs for storage. The body of oligochaetes has a swollen girdle-like structure, the clitellum, which serves an important function in reproduction. After the eggs have matured, a mucous tube, secreted from the clitellum, slides along the body as the worm moves backward. The stored sperm are discharged into this tube, as are the eggs when the tube slides along the section containing them. As the worm literally passes out of the tube, a mucous, lemon-shaped cocoon forms around the now-fertilized eggs. This cocoon serves as a kind of primitive nest, in which the young hatch.

Many leeches (class Hirudinea), all of which are hermaphrodites, have copulatory behaviour much like that of earthworms. Cocoons are formed in a manner similar to that described above, but in some leeches the cocoon is transparent and remains attached to the parent in which the eggs were developed. After hatching, the young leeches remain attached to the "mother" until they become independent. One African leech gives birth to live young and even possesses a special incubating chamber in its body for the developing embryos.

Mollusks. The animals in the phylum Mollusca (*e.g.,* clams, snails, and squid) display a diversity of reproductive behaviour. The majority of the amphineurans (chitons) and pelecypods (*e.g.,* clams, oysters) are dioecious—*i.e.,* individuals are either male or female. Because most species simply shed their eggs and sperm directly into the sea, individuals tend to form dense aggregations during the breeding period. The environmental factor that triggers the release of eggs and sperm has not yet been established with certainty, but, at least in a few species, after one individual has shed its sex products, the others follow in a kind of chain reaction that is clearly chemical in nature. In some mollusks, however, such as the European oyster, the eggs are retained and brooded.

The snails and slugs include hermaphroditic as well as dioecious species. Copulation in the hermaphroditic land snail *Helix* is preceded by a curious courtship involving a bizarre tactile stimulation. When the two partners come together, each drives a calcareous dart (the so-called love dart) into the body wall of the other with such force that it is buried deep in the other's internal organs.

The love dart

To avoid predators, some arboreal slugs copulate in mid-air while each partner is suspended by a viscous thread. In the slipper-shell snails (*Crepidula*), which are rather sessile, all the young are males; their subsequent sex, however, is determined by their nearest neighbour. They remain males as long as they are near a female but change into females if isolated or placed near another male.

Remarkably advanced courtship behaviour in the cephalopods, particularly the squids, involves complex visual displays of movement and changes in colour pattern. Males signify that they are ready for breeding by assuming a distinctive zebra-striped pattern, displaying their fourth arm in a flattened manner, and approaching other individuals with a jerky motion. This fourth arm in squids

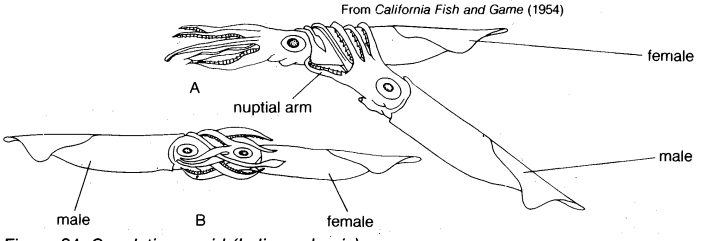


Figure 24: *Copulating squid (Loligo vulgaris).*
(A) Position when spermatophores are transferred to mantle cavity. (B) Position when spermatophores are transferred to sperm receptacle.

and the third arm in octopods, called a hectocotylus, is structurally modified for carrying spermatophores, or balls of sperm. The male cuttlefish (*Sepia*) places the spermatophores in a pocket near the female's mouth, from which the sperm subsequently make their way to the tubes that carry eggs (oviducts). In no squid studied thus far do either of the sexes care for the fertilized eggs, which are laid on vegetation. This is not the case with octopuses, however; at least in *Octopus vulgaris,* the female broods her large number of eggs (about 150,000) for as long as six weeks. During this period she aerates the egg clusters and keeps them free of detritus, exhibiting remarkable behaviour for an animal that produces so many eggs. Brood care such as this is usually associated only with organisms that produce a small number of eggs.

Arthropods. *Crustaceans.* With a few exceptions, barnacles are the only hermaphroditic members of the class Crustacea in the phylum Arthropoda. This is in agreement with the theory that a sessile mode of life tends to be correlated with hermaphroditism. Thus, it is not important for the organism to be near an individual of the opposite sex, but simply to be near any individual of the same species.

Some barnacles are parasitic and have undergone a radical degeneration in form. One, *Sacculina,* is an example of the way in which the reproductive necessities of one species can profoundly affect the reproductive behaviour of another—in this case, the host. Several cells from a larval barnacle penetrate a crab's body and migrate through the bloodstream until they reach the lower portion of its stomach. The cells then send rootlike projections throughout the crab's body. When the crab molts, the barnacle

protrudes a large bulbous portion of its body through the ventral (bottom) surface of the crab. If the crab is a female, its broad shell protects this structure, which contains the barnacle's reproductive organs. The body shape of the male crab, however, is much narrower and does not provide such protection. If the host is a male, therefore, the barnacle first consumes the host's testes; at its next molt, the crab assumes the shape of a female. Should the parasite be removed, the crab regains a male appearance and regenerates its testes.

In the copepods (*e.g.,* sea lice, *Cyclops*) and the amphipods (*e.g.,* beach fleas), the sexes are mostly separate, copulation is brief and without elaboration, and the female of both groups broods the fertilized eggs. The eggs of copepods are usually attached in two clusters to the rear of the female; many amphipods have a special pouch on their ventral surface for brooding the eggs. Many copepods and some amphipods are parasitic on fish and on such marine mammals as whales.

In the crustacean order Decapoda, which includes shrimp, crayfish, lobsters, and crabs, the sexes are separate, fertilization is mostly internal, and egg laying usually occurs shortly after copulation. In terrestrial crabs, however, the females of which migrate to salt water to expel the eggs, the sperm are stored, and fertilization and egg laying are delayed for several months after copulation.

Fiddler crabs of the genus *Uca* and several other decapods show territorial behaviour, an act that is not very common among invertebrates. As in many groups in which males defend territories, male crabs often differ in appearance from the females. Males are much more brightly coloured than the females, and one of their front claws is greatly enlarged; the mostly dull-coloured females have two small front claws. Depending on the species, males perform either simple or complex rhythmic dances in front of their sand burrows. The waving and vertical movement of the large claw is apparently species specific.

The necessity of spermatophores

As in squids and octopuses, the sperm of primitive terrestrial arthropods—millipedes, centipedes, springtails, and silverfish—are often transferred from males to females in structures called spermatophores. During the transition from an aquatic to a terrestrial mode of life, spermatophores became necessary, particularly for those species that had not developed copulatory organs for direct transmission of sperm. Because sperm transfer in these animals is often complicated and takes considerable time, the delicate sperm would be in danger of drying up, were it not for the moisture contained in the spermatophores. It would appear, therefore, that all species that exhibit indirect sperm transfer in which spermatophores are utilized have not achieved complete independence of water.

Males of most primitive soil-dwelling arthropod species place sperm drops on threads in damp locations or use threads or chemical products to guide females to externally placed spermatophores. Most male millipedes have secondary genital appendages called gonopods, by which they transfer the spermatophore directly to the genital opening of the female. One millipede actually uses a "tool" in sperm transfer; the male rounds a fecal pellet, places a drop of sperm on it, and, using its legs, passes the pellet back along its body to a point opposite the female's genital pore. Paired body projections then are used to inject the sperm into the female, and the pellet is dropped. Males of the common bark-inhabiting millipede *Polyxenus* transfer sperm by spinning thin threads on which they place sperm drops; they then construct two parallel thicker threads on which they place a pheromone to attract the female. This chemical and tactile guidance system causes the sperm to become attached to the female's vulva (the external part of the female's genital organs). Males eat the sperm not picked up and replenish it with fresh sperm.

Arachnids. The arachnids (*e.g.,* spiders and scorpions) exhibit the earliest pattern of classical courtship behaviour during which rather ritualized movements are involved. In the true scorpions this behaviour takes the form of the *promenade à deux,* in which the male holds the female by her front claws and apparently stings her in a joint near the base of the claw. The ensuing dancelike pattern apparently results from the male seeking a suitable surface upon which to deposit his spermatophore. After he deposits the spermatophore, the male drags the female over it, releasing her after the spermatophore has passed into her genital pore.

Behaviour of male spiders

As mentioned above, many male spiders have a particular problem in approaching the aggressive and predatory female in order to deposit a spermatophore. The hunting behaviour of most spiders is adapted to react to the slightest movement or vibration of the web, causing the spider to rush forward and bite its prey as quickly as possible. Thus, it is not surprising that male spiders have evolved fairly elaborate display movements and patterns to convey their identity. Many males are quite strikingly coloured, providing additional information about their identity. Some males approach the female only at night and vibrate her web in a highly characteristic manner, different from that caused by the struggling of a trapped animal.

Insects. One puzzling aspect about the courtship behaviour of insects is its sporadic nature. Most insects should exhibit behaviour involving approach, identification, and copulation. Yet, whereas male fruit flies (*Drosophila*) often have elaborate displays preceding copulation, male houseflies and blowflies (*Musca*) simply fly at any object of the proper size and attempt to copulate with it. The reason for these differences in behaviour may be that some insects do not require courtship. Males of some butterflies and moths, for example, simply wait by the pupa and copulate with the female immediately after she emerges.

It is more likely, however, that the majority of insects

By courtesy of the Commonwealth Scientific and Industrial Research Organization, Division of Entomology

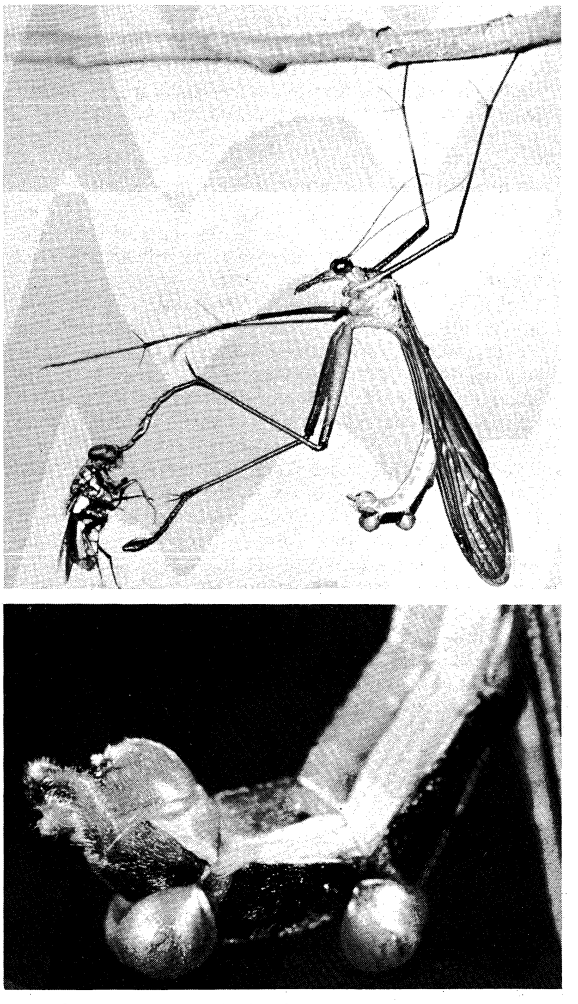


Figure 25: *Reproductive behaviour in mecopterans.* (Top) With his hind tarsi a male mecopteran, *Harpobittacus australis,* captures an insect. Suspending himself he everts two abdominal sacs that release a pheromone that "calls" females. The female, who approaches upwind, is presented with the prey, and copulation follows. (Bottom) Protruding sacs as they appear in magnified view of male abdomen.

have fairly elaborate displays, but man is unable to sense them. The pheromones are, in fact, rather elaborate displays used as sex attractants by many insects; such sensory mechanisms are not usually perceived by man. It has been experimentally demonstrated that the reproductive behaviour of some butterfly species depends heavily on visual clues; similar experiments with other species have failed to show such behaviour. It must be realized, however, that insect vision is quite different from that of vertebrates. Most insects have vision that is sensitive to ultraviolet light, which man and the other vertebrates cannot normally perceive. Butterflies may appear to have identical wing colour patterns under normal light, but, when viewed under ultraviolet light, the patterns differ drastically. Thus, insects that mimic each other in order to appear identical to a vertebrate predator actually possess an unbreakable code by which each species is able to distinguish its own kind.

A reproductive behaviour that is usually misunderstood by those who have observed it is the copulation process in dragonflies. The actual copulatory organ of the male is located close to the thorax, not, as in most insects, near the tip of the abdomen. After a male alights on a plant and transfers sperm from the terminal genital opening to the copulatory organ, he seeks out a female and grasps her behind the head with claspers on his abdomen. Although the two fly in a tandem position, actual copulation occurs only when they alight, and the female bends her abdomen to receive the sperm from the male's organ. Colour, pattern, and movement are important in species recognition. In experiments, it has been found that artificial models acceptable to male *Platycnemis* dragonflies must consist of a female head, thorax, and one wing; the model also must be moved from side to side about once every four seconds to be effective. Complete aerial mating in insects is rare, but it does occur in mayflies, houseflies, ants, wasps, and bees.

Sound displays

Among the cicadas, crickets, and some grasshoppers, females normally mate after they have been attracted to a male by vocalizations of the latter, which, in most cases, are species specific. It has been demonstrated that deafened female grasshoppers do not permit copulation. In many crickets, the specific stridulations (noises) that occur after each copulation keep the female near the male until he is ready to produce another spermatophore. These stridulations also prevent the female from removing the spermatophores before insemination has been completed.

Even some butterflies incorporate sounds into their reproductive displays; in some manner, the butterfly *Ageronia* makes a loud cracking sound when engaged in courtship. Many other insects may incorporate sound into their reproductive displays, perhaps utilizing sounds beyond the sensitivity of the human ear.

Research has revealed that olfactory displays are widespread in insects. The sex attractants for this purpose are usually volatile pheromones. Among certain species of butterflies, such as the queen butterfly (*Danaus gilippus*), the males possess "hair pencils" that project from the end of the abdomen and emit a scent when swept over the female's antennae during courtship behaviour. Copulation does not occur in the absence of this chemical display.

During some stage of their development, a number of insects are either external or internal parasites on a wide variety of animals, including other insects. A particularly bizarre pattern is found in the stylopids, which belong to the order Strepsiptera. Though seldom seen, these insects may be common internal parasites of wasps and bees. The abdomen of the adult females, which never leave their hosts, consists of a bag of eggs that is concealed in the host. The forepart of the parasite, which projects from between abdominal segments of the host, is usually concealed by the host's wings. The females of one stylopid group are apparently unique among animals in having two genital openings—both in the head—in the form of membranous windows. The larvae emerge through these openings, crawl onto a plant, and seek another host. When the host molts its old cuticle (hard skin), the larvae penetrate the soft body. Females extend their heads through the host's abdomen and mature within the host. The males, however, leave the host, pupate in the host's cast-off cuticle, and emerge several days later as adults. The male stylopid then seeks out a host insect and taps it on the side of the abdomen. If no female is present, the male leaves; if a female is present, she somehow signals her presence. The male then inserts his abdomen under the host's wing and enters the genital window of the female.

It is in the orders Isoptera (termites) and Hymenoptera (bees, wasps, and ants), however, that the reproductive behaviour of insects attains its highest level of sophistication. Although dung beetles and some other insect species brood their eggs and care for the young, extreme insect sociality, with its peculiar brood-care system, is found only among the isopterans and the hymenopterans. The principal criterion for such behaviour would appear to be that the female must remain with her brood until after they begin to hatch. Although the phenomenon has been intensively studied, the explanation for the evolution of extreme brood care in ants, many wasps and bees, and termites remains one of the more challenging problems in biology.

Most colonies of social insects reproduce in two ways: either sexual individuals are produced that mate and start new colonies, or the colony breaks up after reaching a certain size. Some species reproduce in both ways. In the first case, the chances of finding new sites are maximized by providing as many individuals of different sexes as possible, each equipped with appropriate guidance mechanisms. In the second, members of the parent colony explore the environment and establish a new colony where suitable.

Another example of reproduction in social insects is that practiced by many ants. Most larvae in an ant colony develop into wingless, sterile workers. Some, however, may get more food (a point that is controversial) and grow more rapidly. These do not pupate when the other larvae do; instead, they become king-sized individuals that eventually metamorphose into sexually mature males or females with wings. Their sex, like that of the wasps and bees, depends upon whether or not the egg was fertilized by the queen.

Mating swarms

The winged sexual forms, or alates, are produced at certain times during the year and swarm in mating flights to establish a new colony, which may actually be no more than a few hundred feet from the old colony. Actual copulation may occur either during flight or after landing on a surface. For most species of ants, it is not known whether a male will copulate with more than one female or if a female will copulate with more than one male. After copulation, the female seeks a location for a new nest and loses her wings within three to five days. Generally, two months are required to rear the first daughter workers. Some females carry a live mealybug with them on the mating flight and take it to the new colony site, where the mealybug's offspring provide the honeydew to feed the ant's initial offspring. Generally, however, the female ant does not provide food for her first offspring; instead, the larvae eat many of the first 100 or so eggs. This egg cannibalism decreases when there are sufficient workers to feed the larvae.

REPRODUCTIVE BEHAVIOUR IN VERTEBRATES

Fishes. The reproductive behaviour of fishes is remarkably diversified: they may be oviparous (lay eggs), ovoviparous (retain the eggs in the body until they hatch), or viviparous (have a direct tissue connection with the developing embryos and give birth to live young). All cartilaginous fishes—the elasmobranches (*e.g.*, sharks, rays, and skates)—employ internal fertilization and usually lay large, heavy-shelled eggs or give birth to live young. The most characteristic features of the more primitive bony fishes is the assemblage of polyandrous (many males) breeding aggregations in open water and the absence of parental care for the eggs. Many of the species in this group, such as herrings, make what appear to be completely chaotic migrations to their breeding areas. Actually, however, each of these huge spawning aggregations is made up of small, coordinated parties consisting of one female and one or more males. On the other hand, a number of fishes are monogamous, form pairs, and care for the eggs or young. In courtship behaviour, in which they utilize all potential

stimuli including sound, chemical, and electrical stimuli, the range and complexity of their displays are not exceeded by any other vertebrate group.

Hermaphroditism in fishes

Although the sexes are usually separate, hermaphroditism is much more common among the bony fishes than in any other group of vertebrates. The reasons for this condition are both physiological and ecological. Whereas the developing gonads of all other vertebrates have an outer and inner layer of tissue, those of bony fishes have a simple origin that lacks any male or female elements. In terms of the evolutionary process, this type of development is likely to be more adaptable to pressures that favour hermaphroditism. When, because of one or several interacting factors, a population density reaches a low point in some species, reproduction may be limited to a low probability of contact with another sexually active individual. In such situations (*e.g.*, very deep sea habitats, tide or stream pools) the evolution of even temporary self-fertilizing hermaphrodites would have the greatest advantage.

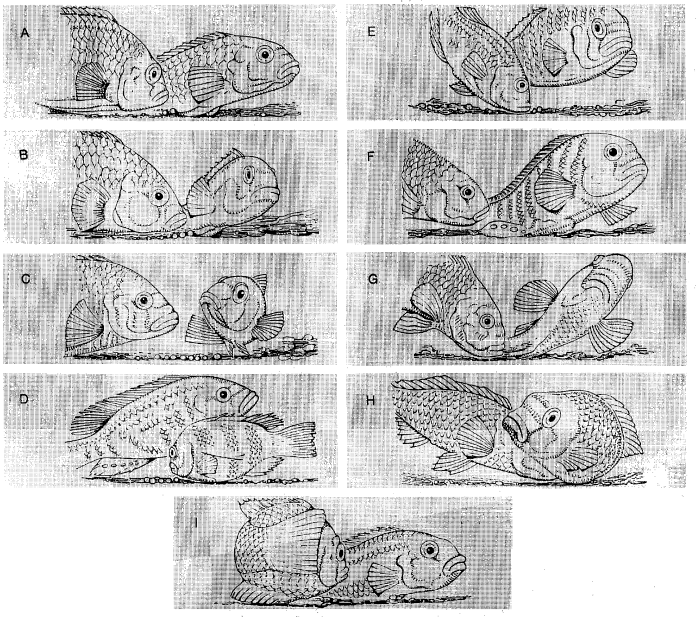


Figure 26: *Phases in the spawning sequence of Haplochromis burtoni.*
(A–C) Female on right deposits batch of eggs; (D,E) female collects eggs in her mouth, where she broods them; (F–H) male emits sperm onto bare surface while female attempts to take up dummy eggs (conspicuous spots near the base of the anal fin of the male) into her mouth; semen enters her mouth in the process and fertilizes the eggs; (I) female deposits another batch of eggs.

One form of hermaphroditism fairly common in bony fishes is the protogynous type, in which the individual functions first as a female and later as a male; it is much more frequent than the reverse situation (protandrous hermaphroditism). The selective reasons for the predominance of the former are presumably associated with the relationship between smaller body size in females and the greater energy requirements needed to produce eggs. In addition, in some promiscuous mating systems, it may be selectively advantageous to be a male when the body size is large and the individual experienced, rather than small and young. Most sea basses, parrot fishes, and wrasses have this sort of hermaphroditism.

Amphibians. Although true viviparity has been described in the African frog *Nectophrynoides,* most amphibians lay eggs. Some salamanders, however, retain the eggs within their body and give birth to live young. Courtship displays in frogs are almost entirely vocal, although in salamanders they may involve tactile, visual, and chemical stimuli. In the European newt *Triturus,* for example, in which mating takes place in the water, the male places himself in front of a female with his back to her. Suddenly, he executes a leap, directs a current of water at her, faces her, and bends his tail forward alongside his body; by waving his tail, he sends toward her a gentle current of water that probably carries a chemical stimulant. If the female responds by approaching the male, he turns and faces away, whereupon she touches his tail and he deposits a spermatophore, which she takes into her cloaca, a common passageway into which waste products and reproductive cells are discharged.

Most frogs and salamanders do not show brood care, but there are exceptions. In the European midwife toad the male rather than the female carries the sticky eggs on its hindlimbs. In a number of Neotropical frogs, the male carries the eggs under a flap of skin on its back. In some species, the young (tadpoles) cling to the back of the male by using their sucker-like mouths.

Lack of brood care

Reptiles. Reptiles are the first vertebrates that, in an evolutionary sense, have evolved an egg that is truly independent of water. Indeed, many snakes and lizards have even gone beyond this stage and have attained complete viviparity. It is difficult to generalize about reproductive behaviour in the reptiles because the various groups differ from each other in the sensitivity of their receptor organs. In many turtles, for example, the males are territorial and are very aggressive during the breeding period. Courtship behaviour involves mainly tactile stimuli, but olfactory clues are also important. It has been recorded that the wood turtle (*Clemmys*) actually emits a low whistle during courtship. Turtles usually bury their eggs and do not brood them.

Lizards appear to use almost every sensory mechanism in their reproductive activities. The nocturnal geckos employ vocalizations, in addition to tactile and olfactory stimuli. Skinks such as *Eumeces* rely heavily on olfactory clues. Lizards of the large family Iguanidae, on the other hand, are almost entirely diurnal creatures and utilize, in the main, visual displays, some of which are the equal in complexity to any known among the vertebrates. Many, such as the anoles, are equipped with a throat flap (dewlap) that is often brightly coloured and specifically marked; it is utilized both in courtship and territorial defense. The skinks and a number of other lizards are known to guard their eggs.

In general, the reproductive behaviour of snakes is not well known. The tongue is apparently an important sense organ for receiving olfactory and other chemical stimuli. The males of some snakes have characteristic skin papillae (nipple-like projections) on the throat; the fact that they rub the papillae over the female's body suggests that tactile stimuli are also important to reproduction. In boas, the rudimentary pelvic bones serve as "claws" for lifting the hind end of the female and for producing a vibration that is said to be important in the process of copulation. Some snakes, the pythons in particular, incubate and guard their eggs.

The bellowing roars of male alligators serve to establish breeding territories and apparently also to attract the females. Female crocodiles remain in the vicinity of their nest and will defend it vigorously.

Birds. Although all birds lay eggs, it is curious that they do so, because the time of highest mortality in most birds usually occurs during the egg-laying period. Apparently, birds lack some adaptation that would permit them to become viviparous.

Most birds build a nest and incubate their eggs, but the incubator birds and such brood parasites as cuckoos are among the exceptions to this rule. Many females that lay a fixed number of eggs are referred to as determinant layers. The pigeons and doves are outstanding examples of this behaviour; for some as yet unknown reason, they never lay more than one or two eggs. Other species are often referred to as indeterminate layers because, in the absence of a suitable stimulus, they continue to produce eggs. More often than not, this stimulus is the presence in the nest of a certain number of eggs. Such behaviour is clearly adaptive—if eggs are lost for some reason and if other environmental stimuli are present, the missing eggs are replaced. The distinction between determinate and indeterminate layers is often blurred, for many indeterminate layers will not replace more than one or two missing eggs.

Number of eggs laid

The duration of egg incubation varies from as little as

nine days in some tropical perching birds to as long as 80 days in some albatrosses. In most species that form pairs, both individuals incubate and feed the young, but the female usually has the greater share of the burden. Among the exceptions to this behaviour pattern are the tinamou (partridge-like game birds), ostriches, some gallinaceous species (*e.g.,* pheasant, grouse, turkeys), and phalaropes. In the phalaropes, the role of the sexes is largely reversed: the females are more brightly coloured than the males and, not surprisingly, are the aggressive ones in courtship and in territorial defense; incubation is carried out solely by the male, but the female aids in feeding the young.

Because many birds begin incubation with the laying of the first egg in the clutch, the eggs hatch at different times. This strategy is often employed by species whose food supply for the young may vary in abundance over a fairly short period. Hence, should food suddenly become scarce, only the smallest chick or chicks will starve rather than the entire clutch. Species in which the young hatch in a relatively well developed, almost independent state tend to have very large clutches, as in many gallinaceous birds. In this case, it might be said that the ultimate size of the clutch is regulated by the abundance and quality of the food available to the female as she produces eggs. The same explanation also accounts for clutch size in parasitic birds—*i.e.,* those that lay eggs in the nests of other species.

Breeding density

The breeding densities of birds vary from one pair in many square miles, as in some birds of prey, to such species as the fulmar, which forms colonies numbering as many as 250,000. Some colonies of the African weaverbird (*Quelea*) have been estimated to exceed 1,000,000 individuals.

One interesting aspect of reproductive behaviour in birds, possibly peculiar to them and to some mammals, is that many courtship displays are learned, or at least perfected through practice, from the parents. An example is the learning of birdsongs. It has been shown in some cases that when chicks are switched from the nest of one species to that of another, they learn some and perhaps all of the songs of the foster parents and do not develop their own species' vocalizations. When mature, such birds often prefer to choose as mates individuals of the same species as their foster parents' rather than those of their own species.

Courtship stimuli in birds are mostly visual and auditory, but it is possible that odour may be important in some petrels and shearwaters. As previously mentioned, most birds form pairs. In these and in many that do not, the males engage in communal, or lek-type, displays on a common courtship ground, such as the familiar strutting grounds of turkeys and many grouse. In addition, there are the incredibly bizarre communal dances of the birds of paradise; the jungle-floor dancing of the cock of the rock; the pasture display grounds of the shorebird, the ruff; and the forest arenas cleared for displaying purposes by the tiny manakins. Many of these display areas are used for many years; in some manakins, for example, certain cleared arenas have existed continuously for at least 30 years. In most lek species, the males are usually brightly coloured, and the females are rather dull in appearance. An exception occurs in some hummingbirds, the so-called hermits, in which both sexes are rather dull in coloration and in which the males group together in singing assemblies.

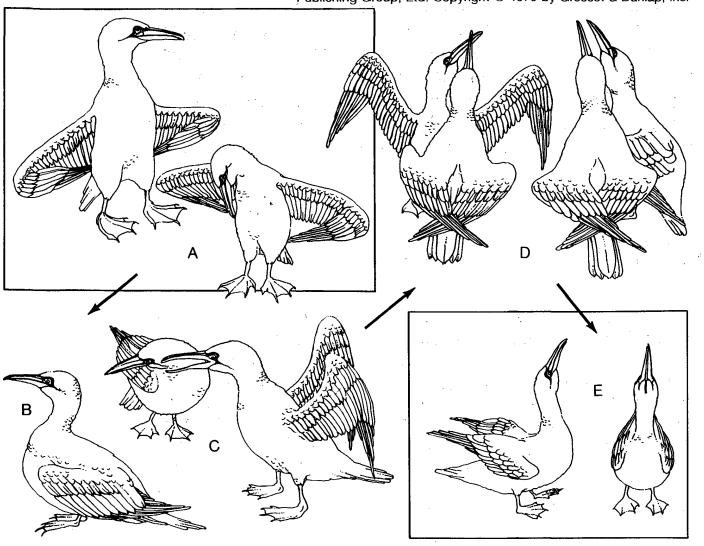


Figure 27: *Courtship behaviour in the North Atlantic gannet.* (A) Bowing; (B) male advertising; (C) facing away; (D) mutual fencing; (E) sky pointing.

Mammals. Most mammals give birth to live young. The outstanding exceptions are the egg-laying monotremes of Australia, the platypus (*Ornithorhynchus*) and the echidnas (spiny anteaters). In the duckbill platypus, a brief courtship involving a chase in the water precedes copulation. The two eggs that are produced are placed in a burrow and hatch in eight to 10 days. In the reproductive behaviour of the spiny anteater (*Tachyglossus*), the female apparently lays her single egg directly into her pouch.

As already mentioned, another general aspect of reproductive behaviour in mammals is the estrous cycle, knowledge of which is essential to an understanding of the mechanisms involved in the reproduction of any mammalian species. Females are usually responsive to males only during that portion of the estrous cycle when they are in heat; that is, when one or more eggs have broken out of the ovary and are in the process of descending to the uterus. The factors causing this event vary greatly, but in some such as rabbits and cats, copulation itself is the main stimulus. In general, however, those mammals, particularly the large ones, that live in temperate areas—*e.g.,* bears, dogs, wolves, foxes, seals, and some deer and antelopes—have one estrous cycle per year. Mammals that live in warmer zones, such as some areas of the tropics, tend to have more than one estrous cycle per year. The sexual cycle in males, the height of which in some forms is referred to as the rut, is, not surprisingly, usually correlated with that of the females. The males of many species of domestic mammals, however, seem to be capable of copulating at almost any time of the year.

Estrous cycle

Another general aspect of mammalian reproductive behaviour is that they do not normally form pairs. Exceptions occur in certain carnivores and in some primates, in which parental care is divided between the sexes. As in many insects, the courtship behaviour of most mammals does not appear to be elaborate; but, just as in the former group, most mammals (humans are an exception) have an acute sense of smell. It is possible, therefore, that many of the chemical attractants wafted into the air by receptive females are actually courtship displays that are more complex than has been realized. This is not to say, of course, that visual, auditory, and tactile displays do not occur. Many deer and antelopes, for example, have rather complex ritualized visual displays employing such movements as strutting and arching of the heads, as well as conspicuous colour patterns. Males in many species discharge urine on females as a preliminary to copulation. Tactile and auditory displays have been shown to be important in aquatic mammals, such as porpoises and whales.

In addition to a number of mammalian pheromones, other odour effects occur in mammals that, aside from their simple advertising value, have an important influence on reproductive behaviour. It has been shown that, when a recently impregnated female mouse is exposed to the odour of a male other than the one with which she has mated, implantation of the egg in the uterus often fails; as a result, there is a rapid return to estrus. The odour of a strange male may signify to a female rodent an unfavourable situation in which to raise young, inasmuch as a number of male rodents attempt to attack offspring not their own. Although it is not yet certain, there might be an adaptive explanation for this behaviour. The population fluctuations of rodents have attracted much attention, and, perhaps correctly, studies have focussed on the ecological parameters of these fluctuations; for example, it has been demonstrated in the laboratory that certain behavioral mechanisms involving odours exercise profound control over the reproduction and population levels of rodents. It has also been shown that the odour of mice can stimulate the production of hormones that cause a decrease

Additional effects of odours

in the reproductive capacity of other mice. In another study, estrus was suppressed and many pseudopregnancies developed when four or more female mice were grouped together in the absence of a male. These results offer a partial explanation for the reduction of population growth in rodent colonies with high population densities.

EVOLUTION OF REPRODUCTIVE BEHAVIOUR

There is a popular tendency to think of primitive animals (in a phylogenetic or descent sense) as lacking "elaboration"; *i.e.*, that the animals of earlier geological periods had simpler displays or perhaps lacked crests or pheromones or elaborate communal displays in comparison with their present-day counterparts. There is no a priori reason for this belief. The fossil record indicates that the societies of which these animals were a part were as diverse and complex as those in which their relatives now live; certainly their display repertoires should have been equally complete. This is not to say, however, that the primitive forms of reproductive behaviour used the same displays for courtship as do the modern forms.

Displays. It has been pointed out that, in general, animals have relatively few displays; in addition, it has been deduced that the relative stability of displays is a dynamic equilibrium—that is, new ones are gained and old ones are lost at about the same frequency. Displays are lost when they no longer convey a selective advantage to the individuals using them; that is, when they are no longer effective in promoting the behaviour that seeks to maximize gene survival in the next generation.

New displays, on the other hand, generally arise by ritualization of previously existing behaviours or functions; that is, when a selective advantage accrues to those individuals who, to convey information, use certain behaviours or functions in a manner that is either partly or totally different from their original purpose. Pheromones, for example, are usually derived from compounds that are natural breakdown products of body metabolism, such as the compounds in urine. Thus, urine, as the precursor of these chemical sex attractants in insects, functions for display purposes, which is far removed from its basic excretory function.

Darwin's theory of sexual selection

Darwin proposed a theory of sexual selection to account for the presence in animals of displays and functions that apparently were not related to survival. He pointed out that two general concepts were involved. First, the evolution of such characteristics as the larger size of males in many species and the development of horns and antlers in mammals could be accounted for by their usefulness in fights between males for their sexual possession of females. This concept has been termed intrasexual selection. For such colourful male structures as the plumes of birds of paradise and the tails of peacocks, Darwin suggested that they resulted from the cumulative effects of sexual preference exerted by the females of the species at the time of mating. This second concept has been termed epigamic selection.

A displaying male has been known to convey information about his relative fitness; that is, his ability, with respect to other displaying males, to maximize the survival of his genes into the next generation. Both the brightness of his coloration and the frequency with which he struts say something about the effectiveness of his genes to produce a "healthy" individual. Once this correlation takes place, selection favours those females who are able to choose the "most fit" males. Correspondingly, sexual selection intensifies the signals up to the point at which any further elaboration of those signals would result in a loss of fitness. When selection goes beyond this point, the male, because of his elaborate ornamentation and other displays, is more likely to suffer from predation before he has the opportunity to reproduce.

Sexual selection. The discussion concerning courtship displays leads naturally to the concept of sexual selection. Why do the males of some species possess elaborate displays? Why, in fact, do some species "elect" to utilize one mating system, say a monogamous one, while others "choose" a polygamous one? It has been suggested that many courtship displays and mating systems, particularly those involving polygamous systems with communal displays in a common courtship area, have an epideictic function—that is, they provide information as to the number of like individuals in a locality. The animals then act according to the information received, often by reducing their reproductive output. Because this concept implies that natural selection is acting for the good of the species rather than for the good of the individual, it has been called group selection. This concept has provoked considerable controversy for two reasons: first, there is no known mechanism by which group selection can function; second, as mentioned earlier, the pertinent behaviours involved can be more simply explained in terms of Darwinian selection dealing with individuals rather than groups.

Polygynous mating systems

In a number of polygynous (mating of one male with more than one female) and promiscuous species, adult females outnumber adult males, sometimes by a factor of five or more. It has been erroneously suggested that this sexual imbalance is the cause of the polygynous mating system, in which one male has several female partners. It has been demonstrated, however, in all polygynous species so far studied, that the ratio of males to females is 50:50 at the time of birth; in many cases, this ratio persists until the cessation of parental care. Therefore, it is the polygynous relationship that causes the imbalance, not vice versa: because sexual selection is the dominant factor in a polygamous and promiscuous species, it results in a greater mortality of males than of females.

Because one male can impregnate many females, thus lowering the selective value of an individual male, females are more valuable than males in an evolutionary sense. It can be seen, therefore, that sexual selection always favours a polygynous and promiscuous system unless it is disadvantageous to the females, as it is in most birds. In most mammals, however, polygyny is the dominant mating system because the male is not needed for parental care. Therefore, monogamy is favoured over polygamy only when some environmental resource (food, for example) is limited and when the maximum survival of young requires the care of both parents. As in all other aspects of reproductive behaviour, the type of mating system that is employed by a species is the result of natural selection.

(N.G.S.)

BEHAVIOUR OF ANIMALS IN GROUPS

Characteristics of social behaviour among animals

Social behaviour among animals takes many forms. The American naturalist and artist John James Audubon observed one of the largest social groups that man has ever known, in the fall of 1813 near Henderson, Kentucky. The species was the passenger pigeon (*Ectopistes migratorius*), once incredibly numerous but hunted to extinction by the end of the 19th century. Audubon wrote:

> The air was literally filled with pigeons; the light of noonday was obscured as by an eclipse; the dung fell in spots not unlike melting flakes of snow.... The people were all in arms.... For a week or more, the population fed on no other flesh than that of pigeons.... The atmosphere, during this time, was strongly impregnated with the peculiar odour which emanates from the species.... Let us take a column of one mile in breadth, which is far below the average size, and suppose it passing over us without interruption for three hours, at the rate mentioned above of one mile in the minute. This will give us a parallelogram of 180 miles by 1, covering 180 square miles. Allowing 2 pigeons to the square yard, we have 1,115,136,000 pigeons in one flock.

SOCIAL AND NONSOCIAL BEHAVIOUR

The largest social organizations ever known are those of desert locusts; the pigeons were second; and present-day

China is probably third, although some pelagic fish schools may be next. Persecution by man is reducing all the large social organizations except his own: the bison and the anchovies off California and Peru have fared poorly compared to smaller groups.

Large numbers or crowding do not in themselves constitute social behaviour. It is usually true, for instance, that a fish that produces a million eggs tosses them out less socially than does a fish that produces a single young and cares for it much more, and that a polygamous bird is less social than a faithfully monogamous one. Overcrowding leads to many social abnormalities. Crowded cats, for instance, develop a "despot" and "pariahs," and there is an almost continuous frenzy of spiteful hissing, growling, or fighting. Crowded rats display, in addition, hypersexuality, homosexuality, and cannibalism.

Effects of crowding

Animals sometimes are brought together by some localized attraction or scarcity, as are moths around an electric light, animals at a water hole in the dry African savanna, birds and bees at a fruit-bearing tree, or iguanas crowding to nest on islands free of predators. To determine if a grouping is social or not, it is necessary to examine the distribution of the animal within the limits of its needed habitat. Most animals require a certain sort of habitat—woodland for a squirrel, or nearly bare ground for a horned lark. Within the correct habitat, the animal also requires certain resources, such as food and water and nesting or roosting sites. The needed habitats and resources collectively form the "niche" of the animal. If an animal's niche is locally distributed, the animal may be found clumped, even if it is not particularly social. If the niche or habitat is patchily or irregularly distributed and the animal cannot move easily from one patch to another, it is said to live in a "coarse-grain" environment or to have a "coarse-grained" niche. Such animals often seem social when they are not. If the niche or habitat of an animal is rather uniform, so that the animal can move about and find what it needs in many places, it is said to live in a "fine-grain" environment or to have a "fine-grained" niche. Such animals often seem solitary when they actually are reacting to each other and hence are social. They tend to be solitary because they do not need to follow others to get to the right environment.

Within a fine-grain environment, or within one "grain" of a coarse-grain environment, animals may occur in groups even when they are not social. A "random" pattern of distribution, in which animals wander without regard to each other, brings asocial animals together at times. An even distribution is much more common, as in territories of many songbirds in which each pair occupies its own plot of ground; these are actually social animals, in the sense that each interacts with its neighbours so as to keep them at a certain long distance. "Clumped" distribution of animals is also common; this usually is truly social in the sense that each animal interacts with its neighbours so as to keep them at a certain short distance. Regularity of the short or long distance an animal keeps from its neighbours is thus as much an indication of sociality as is grouping; only the theoretical (and probably nonexistent) animal that completely ignores its neighbours is truly nonsocial. Time is also a factor in spacing; truly social animals have a tendency to move to the correct distances from each other and to maintain those positions over specifiable time periods, such as for the morning hours each day.

FACTORS INVOLVED IN SOCIAL INTERACTION

Social interaction in time and space is sometimes shown to the ethologist by patterns of following and leadership, although neither of these is necessarily social; following occurs in a dog tracking a rabbit, and leading is seen in an anglerfish luring a smaller fish for dinner by dangling a fleshy protuberance on the snout. In the Eurasian red deer (*Cervus elaphus*), an old female leads the does and fawns about. Spanish merino sheep have been bred to follow each other, while the Scottish highland sheep have been bred to be more independent; following may thus vary within a species according to its needs or to the environmental pressures on it.

Leadership and dominance

Following need not mean there is a leader. The first bird in a migratory "V" of geese or pelicans is not continually the same, and the leaders of a school of fish change every time the school changes direction.

Leadership is sometimes, but not always, associated with a dominance hierarchy or peck order, which may or may not be a sign of social behaviour. Peck order, first noted in bumblebees, means that *A* pecks *B*, *B* pecks *C*, etc., but it does not necessarily mean that *A* leads *C* or *B* or that the three are interacting constructively. The dominant central males of a baboon troop tend to influence the direction taken by the rest of the troop around them; but the bullying dominant stag circling around a herd of red deer has to follow wherever the females decide to go, and his initiative is limited to running straying females back to the herd.

Drawing by Christian D. Olsen

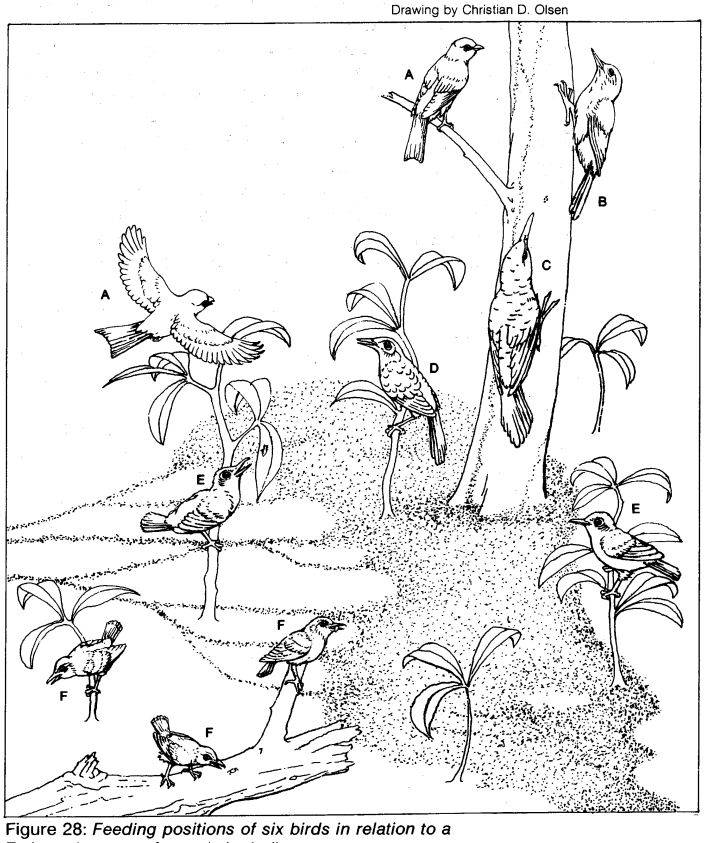


Figure 28: *Feeding positions of six birds in relation to a T-shaped swarm of ants (stippled).* (A) Gray-headed tanager, *Eucometis penicillata.* (B) Plain-brown woodcreeper, *Dendrocincla fuliginosa.* (C) Barred woodcreeper, *Dendrocolaptes certhia.* (D) Ocellated antbird, *Phaenostictus mcleannani.* (E) Bicoloured antbird, *Gymnopithys bicolor.* (F) Spotted antbird, *Hylophylax naevioides.*

Dominance hierarchies may occur whether or not there is social behaviour in the usual sense. Among several species of birds that follow swarms of army ants in Panama to catch insects flushed by the ants, the big ocellated antbird (*Phaenostictus mcleannani*) is dominant, the medium-sized bicoloured antbird (*Gymnopithys bicolor*) is next, and the small, spotted antbird (*Hylophylax naevioides*) is chased by all the other members of the flock (Figure 28).

Crowding almost any two solitary animals together will produce a dominance hierarchy, in which one animal becomes boss or kills the other. This is a major cause of deaths in zoos and aquariums, but it is not necessarily social behaviour. Some biologists even say that dominance hierarchies are evidence of antisocial rather than social behaviour and are expressions of inadequacy in overcrowded social systems. It is certainly true that most peck orders appear in unnatural situations, such as among chickens in a henyard or animals in a cage. In most animals, the absence of a dominance hierarchy, rather than the presence of such, in a crowded context is a sign of a high development of social behaviour.

A further interaction that the ethologist watches for in

Division of labour

social animals is division of labour. Any two animals will, of course, divide up food or any other resource between them. Indeed, no two animals in nature ever have precisely the same niche; if two species have similar niches, they will tend to develop different ways of doing things, or one will exterminate the other. Ecologists call this the "competitive exclusion principle." Man attempts to reduce insect competition by various methods of "control." He kills off other animals by cutting down the trees in which they live. Most animals are less destructive and tend to divide up the world rather than to exterminate other species. Evolution, before the advent of man, seems to have produced continuously more kinds of animals and a greater division of niches, except in periods of environmental disaster. The three antbirds mentioned above tend to specialize in large, medium, and small prey; to the degree that these birds take different foods, they are cooperating better with each other.

Division of labour also occurs within a species. Males and females of some woodpecker species forage in different places in the trees, taking different types of food. Some animals use the same resource but do different things to get it. The male huia (*Heterolocha acutirostris*), an extinct bird of New Zealand, apparently used his short straight beak to open logs while the female, with her long curved beak, removed the insect larvae from the long tunnels he exposed. Male and female birds often build nests together, dividing the labour equally in some cases.

Social insects have more elaborate divisions of labour. Animals that cooperate by division of labour tend to use varied resources and to find more uses for them.

Division of labour seems to be a passing phase in the evolution of most animals other than man. It is evidently of little advantage to most animal societies. The honeybee, the leaf-cutter ants, and other animals with division of labour nearly always occur only where there is little competition from more specialized animals. Few of the most advanced insects or other animals show division of labour. In habitats such as the coral reef and the tropical forest, division of labour tends to be among different species rather than within a species.

Social interactions of various types are more important in determining degree of sociality than are most of the above characteristics. The only interactions between animals that are seldom considered social are behaviours in which animals take something needed by others. The question arises, however, as to how to classify the parasitic relationships of a fetus in the mother and the male anglerfish (*Photocorynus spiniceps*) on his mate, or the fact that killing by predators may help animals to avoid overgrazing their habitats. One could speak of communication of feint and chase in the interaction between moose and wolves. When a bird chases another off its territory, it uses communication and interacts with the other bird very strongly.

Humans often consider chimpanzees or bees as more social than desert locusts, because the locusts have rather simple interactions. Male hummingbirds and birds of paradise, however, have their elaborate plumages and social displays because the females get together with them so seldom that they might not recognize a suitable mate without the displays. It seems generally true that elaborate rituals evolve where social bonds are most fleeting or likely to be disrupted. To determine sociality, one must look at the total spectrum of social interactions as well as at their diversity and productivity, rather than just at a single feature.

Altruistic behaviour

Altruistic social behaviour is often found among animals. An altruistic animal is one that expends some of its energy helping another without direct benefit to itself, be it a mother bear protecting her cubs against their hungry father or a bird giving an alarm call that warns its neighbours of a hawk. An alarm call may help the bird itself, of course, by startling the predator or warning it that this alert bird will be hard to catch.

Psychologists have found that a rat or monkey will slow its rate of pressing a lever for food if that lever also gives an electric shock to a nearby rat or monkey. Rats will take turns sitting on a platform so that others can feed without being interrupted by electric shock. Rats or pigeons can be trained to cooperate in getting food.

Since the altruistic animal always loses something by its behaviour, the question arises why altruism exists. One answer is that, as the evolutionist Charles Darwin suggested, when an animal protects its offspring, it helps its kind to survive the process of natural selection. When porpoises help an injured relative to the surface where it can breathe, they seem to be following a pattern of behaviour that can be accounted for by evolution. Their altruism clearly helps the group and therefore becomes part of the genetic endowment. Altruism, significantly enough, is usually limited to an animal's relatives. Most social animals, such as penguins, feed only their own young. When the individual animal loses more than it or its relatives gain, as when female seals nurse young not their own, the question arises whether this serves the survival of the larger group or the species. Under some conditions, the survival of the group may be more important even than survival of the individual, as when the honeybee dies defending the hive. The worker honeybee, which is not able to reproduce, is in the biological sense not an individual so much as an extra limb of a collective animal.

Reciprocal altruism, in which a benefit is later returned to the benefactor, need not be between related animals and may not even seem altruistic. Alarm calls of birds often alert entirely unrelated kinds of birds, which later may return the favour. An act that seems selfish in the short run is sometimes altruistic in the long run, or vice versa, in the case of maladaptation. Wasps, ants and termites that cannibalize or dominate nestmates at times of food shortage may better keep the colony from starving. Individual ants and bees are often lazy, spending most of their time resting or wandering aimlessly, but these unemployed individuals form an easily mobilized reserve in times of danger.

Individual and group recognition are often important aspects of social structure. Ovenbirds (*Seiurus aurocapillus*) in North America recognize neighbouring males by their songs and react aggressively mainly to songs of strange males. Animals that have long parental bonds often show individual recognition. Herring gulls (*Larus argentatus*), for example, recognize chicks or mates by slight differences in voice or appearance. The larger or more ephemeral the society, the less there can be individual recognition between distant individuals and the more important becomes recognition by group characters, such as the "nest odours" of social insects. Ants, bees, and termites often attack strangers, or even members of their own colony that have been experimentally removed for a few days or washed. Many kinds of parasitic insects (beetles, flies, butterfly caterpillars), however, provide food or scents that gain them entry to a nest, then prey on larvae there.

Other internal characteristics of societies are age structure, birth rates, and death rates. A young wasp or termite colony has few old animals, a mature colony has more, and a declining colony or one that is producing reproductive forms has few young. The old colony has a lower percentage of foraging workers than does the young colony, and has a lower birth rate and higher death rate as a consequence; but only the old colony produces reproductive forms.

Nomadism and migration

Societies also perform movements, such as nomadism and migration (see above *Migratory behaviour*). Army ants wander nomadically after prey. Wildebeest and locusts of Africa emigrate to green areas of local rains; flocking or solitary birds migrate back and forth to escape winter or drought; anadromous fishes, such as salmon, move to the sea for food and to rivers to spawn; catadromous ones, such as eels, do the reverse.

Migrants are often placed at a disadvantage compared to residents, for the latter can take the regular food supplies and leave only ephemeral sources for migrants. Migrants that follow army ants for food in Central America are subordinate to residents and succeed only when residents are absent. The migrant can turn its world from a coarse-grained one to a fine-grained or dependable one by migrating from one patch to the next, and by force of numbers migrants sometimes displace residents.

Societies can make use of seasonal environmental changes by migrating locally, such as sowbugs clustering to estivate in hot weather or ladybugs hibernating in masses. Other societies show food storage; *e.g.,* harvester ants (*Messor*) and wood or pack rats (*Neotoma*). Honey ants (*Myrmecocystus*) have a "replete" caste that bloat their abdomens with stored honey and hang from the roofs of underground chambers until tapped by other workers.

Societies show population fluctuations, from extinction to explosion. Mass emigrations in some, such as lemmings and squirrels, occur mainly after population explosions or periodic extirpations of food supplies. Some populations, such as many protozoans, worms, and insects, normally undergo violent fluctuations, but are resistant to extinction. These are animals in which the adults emigrate or both adults and young emigrate. Animals in which the adults normally occupy a fine-grained habitat and only the young move, such as most higher animals, normally have relatively stable populations, but are easily killed off by a new predator or temporally unpredictable events.

A major external characteristic of the more complex societies is that they construct things, or modify their environments. The elaborate air-conditioned castles of some termites, the path systems of feral house cats, and the patterns of singing at dawn among birds of ephemeral or coarse-grain habitats, all are "structures" created by animals.

Cooperation and competition are major aspects of animal social behaviour. Social facilitation, as when yawning spreads through a pride of lions or chickens eat more rapidly together, shows that cooperative competition can be social. Social animals, which live close together, often interfere and fight more with each other, especially in early stages, than do solitary ones. A fair percentage of the communication between social animals involves "agonistic" or threat–submission behaviour. If this behaviour results in a more adaptive dispersion of the animals, it has been altruistic. Grouped goldfish and other animals survive heat, cold, metallic ions, and other "pollution" better than do isolated animals; but if too many animals aggregate, they pollute each other. Social groups therefore have an optimum size and density, based upon an equilibrium between advantages and disadvantages.

Linking the internal and external characteristics of social systems are flows of energy and materials. Social systems use energy in building structures, or information-rich systems. Physically, one can measure the success of a social system by how efficiently and extensively it uses energy and materials and converts them to physical, biological, or cultural structures. Success in the short run can lead to disaster in the long run, as when elephants or humans destroy an African forest and then must starve or emigrate. Energy and material flows involve an interspecific web, the ecosystem, and must be measured over the long run and in general as well as locally or in the short run.

Social behaviour, therefore, must include interactions between different kinds of animals. A flock of sandpipers in flight is not less social if it includes two species rather than one. If species cooperate, a flock of two species can even be more social than a flock of one species. At African waterholes, baboons keep the lookout while associated antelope are good at scenting predators. Symbiosis is social, even though the late biologist Traian Savesculu of Romania was half right when he joked "Symbiosis is like marriage—a mutual exploitation." Like all social organizations, it is both cooperative and competitive.

Social behaviour may thus be defined as "more or less diverse and constructive interactions among two or more animals." Social behaviour is usually constructive, productive, and adaptive; but it sometimes persists for a time after the evolutionary basis for it is gone. For instance, many kinds of hawks in the eastern United States have almost disappeared, but the flocks of small birds that were formerly their prey still form as vigilante groups.

(E.O.W./Ed.)

COMMUNICATION AS A SOCIAL PROCESS

Social behaviour commonly includes communication, which may be defined as the sending of a message through a medium to a receiver so as to change the status and perhaps the behaviour of the receiver. In general, communication is employed by animals to attract or repel other individuals of particular groups and to establish and maintain distinct forms of social organization that range from relatively simple pair and family bonds to the highly structured troops of some primates and the complex colonies of social insects.

Sources of information

The information involved in animal communication can come from many sources; any facet of the environment perceived is considered information. In linguistic communication the primary function of words is to convey information. Similarly, animals (including man) have modes of behaviour that, in the course of evolution, were selected for their value in providing vehicles for conveying information. During the evolutionary process some of these vehicles also retained more direct functions, but many became specialized for a communicative function alone. These communicative acts, known as displays, include various posturings and movements; sounds; particular ways of making contact among individuals; the release of specialized chemicals called pheromones; and even electrical discharges. Displays have been studied as important means for transmitting information in animal communication. There are, of course, other information sources in animals, some of which have also undergone evolutionary specialization toward a communication function. Among them are what may be called badges—*i.e.,* attributes that are merely structural and nonbehavioral in nature: the red breast of the robin, the red underside of the breeding male stickleback fish, and the mane of the male lion. Many other sources of information can be found in the repeated forms of interaction that develop during prolonged relationships between two individuals and in individual expectations about the nature of the roles in which they encounter others, both familiar associates and strangers. The activities of individuals who interact socially provide a constant and usually rich information source, but, in the study of nonhuman communication, the bulk of systematic research thus far has been directed toward displays and badges; it is, therefore, these highly specialized categories that are of the greatest concern here.

The functions of communication. Because the complexity of social interactions makes experimental manipulation difficult, human understanding of the role of signalling in the social life of animals remains largely based upon inference. It is difficult to repeat an example many times with rigid control of all variables except the one being investigated, and attempts to structure the testing situation to simplify the form of interaction often obviate the interaction. Displays are universal among animals of any degree of structural complexity, however, so that they would not have been evolved and retained if they lacked important functions. But the function of a display is likely to differ, depending upon the individuals involved. A small bird seeing an approaching hawk, for example, may utter a vocal display indicating the high probability that it (the communicator) is, or soon will be, engaged in an attempt to escape. Other small birds, upon hearing this vocalization, may seek cover immediately. Hence, the function of the vocalization is to give them a better opportunity to remain alive and not to increase the immediate chances of survival of the communicator—indeed, its chances for survival may slightly decrease. The display functions for the communicator in that it protects individuals whose continued existence provides a benefit to him greater than the cost of using the display. These individuals may be his offspring or associates whose similar responses to the environment will provide him future protection and, through their alertness in the future, make it possible for him to spend less time scanning his surroundings for predators.

From the ways and circumstances in which displays are used and from the apparent responses of recipients, it is possible to enumerate the general functions of animal communication. First, displays guide animals to one another, thereby enabling one to advertise its presence and behavioral predispositions to potential recipients. Displays enable individuals in a group to respond selectively to particular associates at appropriate times.

Selection of information

Second, communication permits animals to identify one another. Individuals can thus select information of importance to them—usually from members of their own species and often particular individuals. Special cases exist, however; members of different species that normally coexist in the same environment may attend each other's signal. Thus, the maximum alarm communicated by one songbird when it discovers a falcon or accipitrine hawk in its environment is attended by all other songbirds species in the area. In addition, by facilitating identification, communication acts at a premating level to help maintain reproductive isolation among species.

Third, communication reduces the amount of actual fighting and fleeing among animals, an excess of which could disrupt social encounters. In functionally aggressive encounters, such as territorial or dominance disputes, this reduction is achieved by threat displays that often lead to some form of capitulation by one opponent before fighting occurs. In less aggressive circumstances, communication enables animals to appease and reassure one another that each is not likely to be initially aggressive in his present state. Fourth, communication aids in synchronizing the behaviour of individuals who must come into appropriate physiological states in order to breed. This is necessary within pairs and, in some species, among whole colonies of pairs.

Fifth, displays enable individuals to use each other to monitor the environment, not only on a relatively long-term basis but also on a very immediate basis. Thus, in species that spend much of their time living in compact social groups, such as flocks, coteries, or troops, an indication by any one individual that it is fleeing precipitously—often a vocal display in addition to the flight itself—usually correlates with the presence of a dangerous predator and leads to evasion, hiding, or alertness on the part of the other members of the group.

Finally, communication facilitates the maintenance of special relationships between individuals by making available information about the readiness of each to engage in certain activities. The maintenance of individual relationships in cohesive groups is furthered by communication, which keeps members aware both of the behaviour of associates whom they may not be able to see and of the readiness of associates to change their activities. For example, vocal displays usually precede flight by a member of a resting family of geese, and the family then tends to depart as a unit. Within some types of relationship, display behaviour also aids in eliciting general classes of responses; for example, offspring usually signal to arouse various forms of care-giving behaviour from their parents.

The functions in which communication appears to be used vary considerably among different species; each has specialized features, some quite remarkable. It has been demonstrated, for example, that vocalizations and other sounds made during hatching by chicken-like birds (*i.e.*, members of the order Galliformes) influence the rate of hatching of sibling chicks, so that all members of the brood can leave the nest simultaneously. It has been suggested that birds migrating in flocks may use signals in order to inform each other of their position in the night sky, so that the individuals in the flock can perhaps compensate for small individual navigational errors.

Dialectal differences in birdsongs

One interesting aspect of birdsong is the occurrence of dialectal differences (regional variations) among populations of a single species living in different areas. Several such changes that are known to occur between adjacent populations of the South American rufous-collared sparrow (*Zonotrichia capensis*) correlate with relatively major habitat changes. Very few dialectal changes occur over an enormous range on the Argentine pampas, but in this case the habitat of the species also changes little. The habitat changes markedly in the Andes mountains over short distances, however, as elevation rapidly increases, and, concurrently, many more dialectal changes occur there in birds' songs. The suggested function of the correlations between display and features of the habitat is that they provide markers that identify populations adapted to different local conditions; such markers would permit more appropriate selection of mates than would otherwise occur, at least in the marginal areas between populations. It has been suggested that a similar functional explanation may be involved in the evolution of human dialects.

Signals have evolved, the primary function of which lies in communication between, rather than within, species—particularly in cases in which identifying markings or displays of dangerous or distasteful animals provide information to potential predators or in those in which innocuous species mimic the signals. Other species, such as some forest falcons (*Micrastur*) of the New World tropics, apparently employ vocalizations as a kind of lure to attract prey species for capture; in this case the information is of use only to the individual providing it. An American ethologist, Martin Moynihan, has shown that elaborately specialized means of communication have evolved in bird species that join to form large mixed foraging flocks. These signals attract individuals to the flocks and help to maintain the cohesiveness of the assemblages as they move through the trees.

Modes of information transfer. The evolution of animal behaviour and structures toward a communication function has yielded a mode of communication adapted for each externally oriented sensory receptor system—*e.g.*, organs of vision, hearing, and taste. Each mode, although specialized, has limitations with respect to such properties as energy utilization; the ability to surmount environmental obstacles; the ease with which the source of the communication can be located; the persistence or transitoriness of the signal; and the available range of physical complexity. These differences have been exploited during evolution.

Sound. Because sound disseminates and fades rapidly, a given unit of information does not remain to interfere with, or garble, succeeding units. In addition, sound can be varied with regard to pitch, clarity or harshness, duration, loudness, and rate of repetition, with each variable providing greater range of ability to encode. One advantage of sound as a medium of communication is that vocal displays can be uttered by animals who need to keep their appendages free for other activities and can be received by individuals who need not face the communicator in order to receive the signals.

Location of sound

It is usually a simple matter for an animal with two ears to locate the source of a sound, although some modifications (described below) can help to conceal the location of the transmitter from potential dangers. Virtually all of the animals for which sound is important are bilaterally symmetrical and hence have paired hearing organs. Sound is a superb means of encoding information that must pass around environmental obstacles, such as trees or other vegetation. Apparently, some animals utilize frequencies that are particularly good at bypassing obstacles; this appears to be the case at least in the vocalizations of forest birds. Because the highest frequencies are obstructed in the forest and attenuated relatively rapidly by wind and air in open habitats, they apparently are not selected for use in the communication of at least the majority of vertebrate species.

The most obvious examples of the use of sound in displays are the vocalizations characteristic of most of the better known air-breathing vertebrates (*i.e.*, reptiles, birds, and mammals). Many nonvocal means of producing audible displays exist, although none match the potential for elaboration found in vocalizations. Many invertebrates produce sounds by rubbing one body part against another (stridulation); this technique is also used by fishes and is, in some ways, comparable to the chest beating done by the male gorilla (*Gorilla gorilla*). Gorillas also beat upon the ground and upon other objects in their environments; alarmed beavers slap the surface of water with their broad tails. Some vertebrates have elaborated on this sort of behaviour. Many woodpecker species, for instance, seek out certain dead limbs or even the tin roofs of buildings on which to produce their drumming displays. The North American ruffed grouse (*Bonasa umbellus*) produces a sound like the beating of a low drum by beating air toward its chest with its broad wings. Many other birds use specialized wing or tail feathers to produce sounds during display flights—such as the "winnowing" flight of snipe.

Other forms of vibrational signalling

Some forms of vibrational signalling are not perceived as sound, at least not by the relevant participants. Thus, although the sounds employed during the social interactions of honeybees are audible to man, it is likely that the bees perceive them primarily as vibrations through receptors in their feet. Some other displays of this type are not audible to man. Males of some web-building spiders, for example, approach females for mating very cautiously, signalling their presence and identity by strumming on the females' webs.

Vision. Visibly encoded information provides for much easier pinpointing of source than either sound or chemical signals, although visible displays are also much more easily obscured by structures in the environment. Ease of locating the source of a signal is often extremely important as, in a gull colony, for example, in which a large number of individuals are present in small space and it is important for the recipient to identify the individual that is displaying. The sight of the communicator also provides information about his orientation and so functions (like the human signal of pointing) in selection of a relevant recipient. To avoid the problem of being obscured by the environment the communicator is often able to select a display position that makes him more easily seen by the most relevant recipients. When the latter are at a distance, for example, the communicator is likely to display from a highly placed station, or, in the case of many grassland bird species that have no high perches available, to perform a flight display above the vegetation. Visible displays in many species of social birds and in at least some primates (such as baboons) are often combined into relatively complex assortments that are thought to convey unusually precise information.

Unlike sounds, which are usually very transient and can be difficult to maintain, visible posturings can sometimes be maintained with relative ease, although they usually interfere with the communicator's ability to engage in other forms of behaviour. Many animals have surmounted this problem by the evolution of badges—morphological specializations, such as bright patches of skin, fur, or feathers; horns; casques; and crests. Badges convey information about the general identity of the communicator (*i.e.,* species, sex, age) and some information about his physiological state. Animals have also evolved ways to utilize sources of information that supplement displays and badges. Some species provide information of some highly relevant samples of the environment; the honeybee, for example, uses a drop of nectar in the dance at the hive to indicate the identity of the food source. Courting males of many bird species feed their mates or provide them with bits of nest material. Unmated male weaver birds make nests at which they display and which are subsequently used for breeding; male bower birds build "bowers" (a variety of display structures), and male manakins clear leks (special display arenas) that serve only a communicative function. The use of constructions in some cases extends to supplying information in the absence of the communicator, as in territorial marking. Rabbits and other mammals use dung heaps (both visible and scented) for this purpose, and bears scratch marking posts.

Despite their flexibility, visible means of signalling have disadvantages in addition to being easily obscured. They may be too easily located, drawing undesirable attention of predators and other inappropriate recipients to the communicator. Moreover, the signals are available only if the recipient looks at the source directly; this hinders his freedom to do other things simultaneously.

Chemicals. Many species have evolved special chemical products that are released under particular circumstances. Some of these substances are used as defense mechanisms against predators; they apparently function primarily by being distasteful or even injurious, but some may serve to warn the predator that its intended prey can harm it, thus eliminating the need for actual contact. A variety of chemicals called pheromones are used for communication within a species. Fishes and other aquatic animals secrete certain pheromones into the water; moths release pheromones as sexual attractants into the air to be wafted downwind; and various highly social insects mark surfaces with pheromones or spray them into the air, sometimes secreting them onto their own body surfaces so that chemicals can be directly tasted by other individuals. Scents are used by many species to mark territories, as in the well-known urination patterns of domestic dogs.

The role of pheromones

Much research has been done on the use of pheromones by species of ants. It has been found that, depending on circumstances, different pheromones are secreted from different glands; for example, one type is secreted when the ants are laying trails, another when they are indicating alarm. The rate of release and the response thresholds of the different pheromones have been adjusted during the course of evolution so that the fading time of the odour correlates with the functions performed. Alarm pheromones, for example, which cause the ants to congregate,

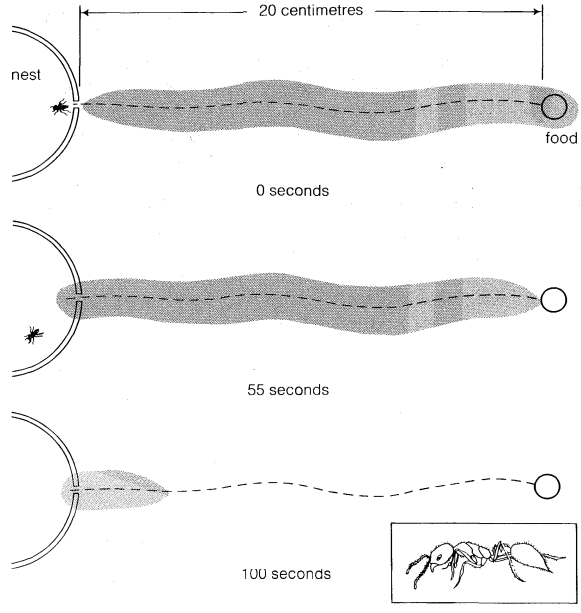


Figure 29: Odour trail of fire ant worker made by exuding a pheromone along its extended sting(s). This experimental situation demonstrates the rate of disappearance and active space of the trail.

diffuse very rapidly and fade quickly, unless continuously renewed by additional secretion from newly attracted individuals who also take alarm. Pheromones used for group identification and for simple assembly (in the absence of alarm), fade much more slowly. Even though pheromones are less persistent than certain structures used for marking, they are used primarily for their persistence—at least in the case of species that have other means of communication available. Extremely rapid fading of the odour is a problem, however, as is the recovery rate of the recipient's chemoreceptors. Pheromones are probably an inadequate means of communication in social events that change rapidly.

Touch. Specialized patterns of touching—*i.e.,* tactile displays—are often overlooked in studies of animal communication unless they involve distinctive movements. Although a recipient animal may be aware of a distinctive touch, an observer may not; nonetheless, many forms of touch are evident. The remarkable dances of honeybees are customarily performed on a vertical comb in the darkness of the hive. Recipients of these displays follow the communicators closely, maintaining contact with their antennae; thus, much of the information is probably received as changing tactile patterns.

Grooming

In many mammalian species, the members of social groups engage in bouts of grooming (called allogrooming when performed on another individual). Although visible to group members other than those in the interaction, allogrooming probably functions largely as a tactile display. Specialized touches with the hands are now suspected to be precopulatory signals in female rhesus monkeys (*Macaca mulatta*). Individuals of the South American monkey

Callicebus moloch rest together in trees with their tails intertwined, a tactile display that probably serves a function similar to that served by allogrooming in social groups of baboons and macaques.

In some cases the communicator leaves information available for individuals it will never meet. Female wasps of at least one species are able to indicate to their offspring the direction in which to seek egress from the cells in which the eggs are laid. The information is stored in the geometry and texture of the walls constructed by the females to seal the nest chambers.

Electrical energy. A number of fishes that live in muddy waters produce regular patterns of electrical discharges as an active sensory system (active in the same sense as a bat's sonar scanning of the environment). There is good evidence that some species respond to electrical discharges of individuals of the same species, and that some aspects of the discharges, including their cessation, may function as displays.

The role of displays. *The display repertoire.* The individual animal may have a repertoire of up to about 40 displays. For most relatively social adult fishes, birds, and mammals, the range of repertoire size for different species varies from 15 to 35 displays. Further, most species have evolved displays adapted for the different sensory modality of their recipients; as a result, their repertoires comprise sets of displays (*e.g.,* both visual and audible) that overlap considerably in the information they convey. A bird can often convey information about the probability of flying by fluttering its wings and by uttering a particular call and may often do both simultaneously. Recipients sometimes are able both to see and to hear the communicator; at other times, however, he may be obscured behind foliage, or, if visible, the sounds may be masked or distorted by wind noise or by the calls of other birds. The redundancy typical of communication systems probably exists primarily to facilitate information transfer in such noisy environments that may unpredictably interfere with any one form of transmission.

Limitations to the number of displays

In view of the immense value of language to man, it is remarkable that evolutionary pressures have not produced a greater elaboration of communication in animals other than man. This implies that further elaboration would impair rather than facilitate communication. Although the nature of this impairment is still a matter of theoretical conjecture, a promising explanation is that of Moynihan, who argues, in part, that it is difficult for a species to evolve the ability to make rapid and finely tuned responses to rare events. Rapid escape from a surprising stimulus (a relatively rare event) has frequently evolved but it is not a finely tuned response. As the size of a display repertoire increases, appropriate circumstances for the use of certain displays would become increasingly rare; Moynihan proposes that the number of displays in an animal's repertoire should not exceed that at which its rarer displays have little usefulness. This number of displays cannot yet be predicted. Other factors also act to limit the number of displays. Among them is the need for each display to be sufficiently distinctive to make it unlikely that a recipient will mistake it for another. Beyond some point, as repertoire size increases distinctiveness requires increasingly elaborate displays that may be time-consuming to produce and require too much attention by the recipient.

The restriction in the size of the display repertoire undoubtedly has had considerable influence on the evolutionary selection of the kinds of information that are encoded. An American zoologist, W.J. Smith, has suggested that this restriction explains two striking characteristics of displays: the paucity of narrowly precise messages and the relative abundance and widespread occurrence of information about such general behavioral patterns as locomotion or social association. Each species has so few displays available to it that only a minimal number can be used in restricted situations. When practical, a display has selective advantages if it conveys information that can be of use in mediating a variety of social interactions. Some types of activities, such as attack and escape, are too critically important in most species to be indicated by ambiguous signals. Others, including many of the acts that bonded (*e.g.,* paired) individuals can perform in each other's company without endangering the fabric of social organization, may be safely left for a recipient to predict based on its experience in similar contexts, provided it has the information that the only likely activities are bond-limited ones. There is strong selection pressure toward both the use of general messages (in situations in which immediacy of correct response is not of overriding importance) and similar messages (for species with a wide overlap of social habits).

The information content of displays. The study of animal displays was once primarily concerned with the motivational states of the animals using the displays. It was recognized that, if an animal uses a display only when in a certain motivational state, then the display informs a recipient that the communicator is in that state; the recipient thus should be able to predict, at least in part, the subsequent activities of the communicator. Because only certain aspects of the communicator's motivational state are common to all uses of a given display, however, these aspects are the only ones that can be said to be encoded in the display. Thus, a recipient must obtain other information in order to establish the presence of other motivational aspects. It is more useful, therefore, to study the behaviour (rather than the motivation) of communicating animals in order to correlate specific behaviour patterns with specific displays. Because communication is a social process, the most productive course of study concerns the nature of the information that can be inserted into social interactions by displays. This information helps to determine the course of the interaction and is concerned primarily with making the behaviour of the interactants more predictable.

Identification

It is important for the recipient to be able to identify the communicator with some degree of precision; if he identifies himself, through displays or some other means, as belonging to a category of individuals important to the recipient, the latter will pay attention to him, insofar as is practical or necessary. Of prime importance is specific identity; the most important communication occurs among members of the same species. Thus, most of the flow of communication among members of other species is irrelevant. Within a species it may be useful for the recipient to know the sex and perhaps the age of the communicator; even the general physiological state (*e.g.,* breeding or nonbreeding condition) may be significant. All of this information, successfully communicated, narrows the range of behaviour that can be expected by the recipient of a display. For bird and mammal species with complex social behaviour, some features of the form of a display may be peculiar to the individual. A female passerine bird (a perching bird) newly arrived from spring migration to the region where she will breed encounters many individuals of several different species. By using certain vocalizations (usually called song), some of the birds identify themselves as males of her species and also communicate to her that they are in breeding condition and probably the information that they are unmated. This information enables the female to predict their responses to any approaches she might make to them. Also, after she begins to form a pair bond with a male, she must be able to identify him.

All displays studied thus far identify the communicator to some degree. The amount or kind of identifying information differs according to the social requirements of the species.

As already noted, identification is not only encoded by animals in displays but also in badges. The strikingly distinctive breeding season plumages of drake waterfowl and of males of many other bird species attest to this. Female birds of many species are less distinctively coloured and it is usually they, not the males, who select a mate; male birds of such species apparently accept any female who attempts to form a bond, but females respond only to males with the correct plumage. The evolution of unusual specializations apparently has been necessary in only a few cases. Gulls, for instance, are able to distinguish species by a combination of eye colour and its degree of contrast with the white head.

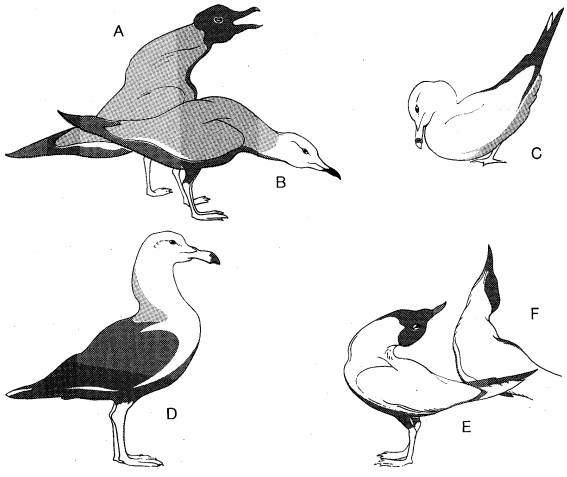


Figure 30: *Visual communication in birds: some display postures that are widely distributed among gulls and terns.* (A) Oblique posture during long call display of the Galápagos dusky gull (*Larus fuliginosus*). (B) Low oblique posture during the new call of the Peruvian gray gull (*L. modestus*). (C) Choking posture in choking display of ring-billed gull (*L. delawarensis*). (D) Upright posture of Belcher's gull (*L. belcheri*). (E) Climax of head-tossing display of Franklin's gull (*L. pipixcan*). (F) Erect posture of gull-billed tern (*Gelochelidon nilotica*).

From M.H. Moynihan, *American Museum Novitates* (1959). The American Museum of Natural History

Location

Closely related in many ways to information concerning identification is information specifying location. An animal may need to transmit its message without fully revealing its location. It may, for example, use a visible display to a nearby recipient while otherwise concealed by vegetation from potential and undesirable recipients that would perceive vocal displays. In some cases, however, animals provide information about their location in proportion to the need of the recipients. A small bird on fleeing from a hawk, for example, may utter a high-pitched warning vocalization that carries little information about its own location. Natural selection has favoured characteristics of this vocalization that make its source difficult for a two-eared predator to locate. Many small birds hear the sound and are warned that one of their associates has found desperate reason to flee; they need not know his exact location or even why he is fleeing. The hawk, who needs primarily the location information, is denied it, largely since he is unable to locate the source of the vocalization and often has not yet seen the fleeing communicator. It is significant that the begging vocalizations of many nestling birds are similar to the warning call; the reason also is similar—to prevent predators from discovering the location of the nest, which the nestlings' parents, of course, already know.

In certain cases a vocalization can vary in ways that increase or decrease its locatability. One example, which has been studied in detail, is the call given by males of the Panamanian frog *Engystomops pustulosus* as they await females at spawning ponds. The call of a lone male contains fewer clues to his location than does the call of a male in a chorus of other males. The former probably can rely on the persistence of any passing female to find him, but, because he is alone he is more vulnerable to predators and so must give as little information as possible about his location. On the other hand, the male within a group gets some protection from predators by virtue of being among other frogs; not only is the approaching predator likely to encounter one of the other frogs first but, collectively, they are more likely to detect the predator. After the chorus has attracted a female to the immediate region, however, any male that is less easily located than his fellows is at a disadvantage; under these circumstances each male maximizes the ease with which he can be located.

Behavioral information

Much of the remainder of the information encoded by animals in their stylized displays is behavioral—it indicates the likelihood with which a recipient can expect a communicator to engage in different types of activities.

The bared teeth, growls, curled lip, and bristled hair of a watchdog at the approach of a stranger indicate that the animal may attack. The same display components may also indicate that it may not attack, that something, perhaps fear, is holding its aggression in check, however tenuously. Much of the displaying done by animals seems to indicate the probability of attack or escape. Such agonistic behaviour patterns must be controlled in the establishment and maintenance of organized social units such as pairs, families, and troops. A species may encode information about antagonistic behaviour in more than one display. The green-backed sparrow (*Arremonops conirostris*), for example, uses a vocalization that sounds like "chuck" when it is less likely to attack than to escape, but, when both attack and escape are equally probable, it utters either of two calls, which Moynihan has described as "medium hoarse notes" or "hoarse screams." Although the last two vocal displays indicate equal probabilities of the two acts, they do not carry exactly the same information. The "hoarse screams" are used when attack or escape is very likely to happen, and the "medium hoarse notes" are used when the bird's indecision between the two alternatives makes it less likely that either will occur. The two displays thus carry information not only about attack and escape behaviour but also about the probability that some act will occur. In fact, information about the probability of a specific act is apparently encoded in all displays, but not always in the way described above. The eastern kingbird (*Tyrannus tyrannus*) encodes the information that it may attack in a vocalization with a sound similar to "zeer." The likelihood that it may attack when using this call varies, increasing as the abruptness and harshness of the vocalization increases.

Of the information widely encoded among diverse animals in displays, little is apparently as narrowly defined as attack and escape signals. Although the encoding of information regarding attempts by a communicator to engage in social play would seem useful in view of the fact that many mammals (and even some birds such as the hobby falcon, *Falco subbuteo*) play socially, information that an approaching communicator is not likely to be vicious but intends only to play is apparently encoded only by a few primates.

Adults of many vertebrate species appear to lack common reasons for coming into contact with their fellows other than aggressively. Many animals thus have difficulty in approaching one another closely enough to copulate. The displays of only a few species provide information that an individual is approaching specifically to copulate; in fact, the general information provided by the individual in other displays apparently is often sufficient, given the right contextual circumstances, to indicate the intention to copulate. So far as is known, the information provided by most birds and mammals through their displays is in large part of a general nature.

The importance of general information

In many birds and mammals, approach in order to copulate is restricted to individuals who have established a pair bond. In this relationship each partner adopts, and recognizes as appropriate in the other, behaviour patterns that characterize certain roles, special restrictions, and special privileges. Such behaviour patterns are regular in occurrence and apparently expected to some degree by each individual. The complexity of this behaviour varies considerably among species with different degrees of social complexity—*i.e.,* the individuals may be merely temporarily mated or permanent members of a troop of numerous individuals—but the appropriate behaviour will, for some relationships, include the possibility of copulation. Many species employ at such times displays that are used in a variety of behaviour characteristic of closely interacting, bonded individuals. These displays indicate that the communicator will probably select behaviour patterns appropriate to the bonded relationship, without specifying what the behaviour will be.

The somewhat more precise information that may be available among bonded individuals usually appears to specify that appropriate behaviour will not involve physical contact. In these circumstances, the communicator simply associates with the other individual, remaining

in his presence but doing nothing that requires bodily contact. Such information about association is apparently basic and is encoded by many species, at least of birds and mammals. A superficially similar sort of information that may also be widespread indicates that the communicator will engage in a variety of activities, from social encounters to nonsocial acts of individual maintenance (*e.g.*, foraging), but will neither attack nor attempt to escape from the individuals to whom the display is directed. He will thus behave nonagonistically.

Drawing by J. Adamska-Koperska based on M.H. Moynihan, *Journal of Zoology* (1966)

Figure 31: Tactile communication; tail-twining by titi monkeys *Callicebus moloch.*

General information that has been identified in displays of three mammalian species (gorilla, man, and the black-tailed prairie dog) appears to be so basic that it, too, may be widespread among social vertebrates. The communicator provides information that he will probably hesitate to engage in a certain social interaction. This information is usually provided when the communicator is engaged in an activity that would be hampered by interruption. How widespread this sort of information is among animals is not known.

Other kinds of general information known to be widely encoded are less directly social in nature. The best understood information involves the probability that the communicator will engage in some form of locomotion. One kind of display may be used when an unmated male bird moves from point to point along the boundaries of his territory, occasionally stopping to advertise. The same display may be used by a mated male as he approaches his rather aggressive and perhaps sexually unreceptive mate. It may also be used by either mate on approaching a predator near its nest, or later by the nearly independent fledglings when importuning increasingly unwilling parents. Nor does this exhaust the range of situations in which this one display is used. In all cases the display fails to indicate with any precision just what the communicator's probable activities will be, but it is well correlated with a behavioral indication of uncertainty over whether the communicator will initiate or cease flying, or change direction if in flight. In any case the activities of the communicator indicate what it may attempt to do, and the display appears to provide information as to whether or not this apparently probable behaviour will be performed. The information the communicator thus provides indirectly implies that there are factors acting to counter his obvious motivational tendencies, these implications have different social relevance in different circumstances.

Displays are also known that encode even more general information—exactly how general is not yet clear, but the following example is illustrative. The blacktailed prairie dog (*Cynomys ludovicianus*) has a combined visual and vocal display in which the communicator throws its foreparts vigorously into the air, directs its nose straight up, and utters an abrupt, two-part vocalization; the performance has been named the Jump-Yip. Jump-Yips are employed on many occasions, all of which have in common a certain probability (usually less than 50 percent) that the communicator will begin or continue to flee. Other forms of behaviour associated with the Jump-Yip differ greatly, depending on the circumstances; for instance, the communicator may greet or associate with its mate or another individual (such as a member of its family group); approach its mate sexually; challenge a neighbour aggressively; or do one of various nonsocial acts such as foraging or dust bathing. The display seems to function either as an indication that some form of behaviour may be interrupted or prevented by fleeing but that this other behaviour may be socially more important than the fleeing; or that the other behaviour may be negligible and the possibility of flight itself the critical feature of the situation for the particular recipient. The display is not used unless there is an alternative to flight, but it does not provide information about what that alternative is. Such information is usually clear enough contextually, however, The unspecified alternative might be anything in a prairie dog's behavioral repertoire. Jump-Yip

Of widespread occurrence in the displays of birds and mammals is information that the communicator, however strongly motivated to perform a particular act, lacks the opportunity to do so. Thus, when, in a social encounter, a juvenile gorilla tries to flee but finds itself cornered, it adopts a special posture indicative of its frustrated escape. An eastern kingbird or a black-headed gull (*Larus ridibundus*) performs a very stylized and spectacular aerial display when it is strongly motivated to attack but lacks a suitable opponent—such as an intruder into its territory. A Carolina chickadee (*Parus carolinensis*) that appears to be seeking a border encounter with a territorial neighbour employs a special form of song if the neighbour fails to appear in response to its challenges. Various songbirds of the New World tropics have vocal displays that are used only when one bird becomes separated from the companions with which it associates while foraging; the calls cease after reunion occurs. In short, it appears that if an animal is frustrated from engaging in activities of sufficient social importance to merit communication, it may be able to encode its frustration with respect to that activity.

The foregoing describes the types of basic information presently known to be widespread among adult birds and non-human mammals. Other types of information are known to be encoded by other animals: ants provide directional information in odour trails, honeybees in "dances." The fire ant (*Solenopsis*) odour trail appears to indicate specifically the route to a food source, but the honeybee dance is less specific—it can be used to provide information about food sources, water sources, and potential sites for new hives. Both cases reflect the evolution of communication within complex social organizations of invertebrates, based on evolutionary properties quite different from those of vertebrates.

Even among vertebrates, some systems of social organization differ markedly from those of most of the birds and mammals whose communication patterns have been studied. Among nocturnal frogs, for instance, breeding males congregate loosely and call to attract females. They attempt to mate with any individual that comes near them—whether it be a male or a female. A male or a female who has already laid her eggs usually utters a

distinctive call and is released; the encoded information appears to be that the individual that has been clasped does not belong to the class that will shortly lay eggs (*i.e.*, gravid females), a rather precise but entirely suitable sort of information. Precision in such narrowly specific messages may well be the rule in groups in which complexity of social behaviour is minimal.

Information needs of young

Immature animals dependent on adults often have social requirements quite different from those of adults and provide, in part, different information, although they too may have displays encoding attack, escape, and frustration information. At least among infant mammals, much displaying is done to indicate a need state—such as hunger, pain, need for social interaction—for which the infant must depend on others. Some displays do not specify the needs that are relevant—the most common forms of human crying, for instance, are rather imprecise. Others, such as the widely gaping mouth of a juvenile passerine bird, appear to correlate primarily with readiness to eat. Infants, because of their often limited behavioral repertoires, are somewhat difficult subjects for behavioral studies of the kind normally employed by students of animal communication. They provide an important challenge, however, as attempts are made to learn both what they encode in their displays and how this information is altered during development into adult forms of encoding.

Interpretation of information. The information exchanged during social interactions among animals allows them to select their subsequent activities with greater certainty that their actions will be more appropriate than would be possible without the new information. When an animal is offered information through a display, it can select a response that is based, in part, upon this information. Depending upon the abundance and the type of information, the response must be such as to permit flexibility—*i.e.*, a generalized response may be practical. On the other hand, receipt of abundant and varied information permits the individual to choose among less generalized, and probably less flexible, responses with a greater certainty of making the appropriate selection. Complete understanding of animal communication, therefore, requires knowledge of the responses that individuals make to the information communicated. This is the study of what Colin Cherry has referred to as the "meanings" that recipients draw from the circumstances in which communication occurs.

The communication process between two animals is summarized in the following paragraphs. Through displays, badges, and other aspects of his appearance and activities, one individual adds information to another individual's environment, in which much information is already available. The human observer is certain that communication has occurred only when the second individual selects a response, demonstrating the meaning the event has had for him. Unfortunately, the study of responses in social situations is technically very difficult.

The recipient animal certainly ignores most of the many sources of information available and weighs others as he selects his response. In most instances, his response is not directly available to the human observer, for the recipient's mental state or disposition may change without an overt indication. For most displays the response observed most commonly is indistinguishable from that of inattention. When the recipient does alter his activity immediately after exposure to a signal, his first acts may be common to many different kinds of events—he may simply face the source of information, move a little closer to it, or a little farther away.

Initial response

The first response of a recipient to a variety of signals conveying many possible types of information about the communicator (including high probability of attack, intent to escape or to engage in association, social play, or copulation) may be to approach the communicator. The predisposition of the recipient to react further is likely to depend on the signal, and further activity may not occur until the recipient is supplied with further information.

Modified displays. Although the nature of displays and badges and of the types of information that animals encode in them is important in the process of animal communication, it provides only part of the story; for example, a display is a vehicle for a particular type of information transfer, but the communicator, in the manner of its displaying, can modify some aspects of this information. The probability of attack, as encoded in the vocal display of a bird species, can be ascertained by the abruptness or harshness of a vocalization. The intensity of the communicator's involvement in the situation can also be indicated by variations in the rate at which the vocalization is uttered, or in its loudness. Such modifications provide information about the nature of the expected attack—*i.e.*, whether it will be vigorous or perfunctory—and about the communicator's state with the passage of time.

The position of the communicator as it vocalizes may indicate a particular recipient as the most relevant in a given instance. There is little reason to conclude that the orientation of the communicator has become ritualized as a part of the display. Sometimes an intermediate orientation becomes part of a ritualized display, and orientation must very often provide important information.

It appears that the information obtained by most animals about situations in the environment is either associated with stylized displaying or is implied by it with fair probability. For instance, a bird that utters a call when it sights and then flees from a dangerous aerial predator is not specifying the presence of a hawk, even though it is likely that it has been frightened by a hawk; under some circumstances (*e.g.*, if the bird has nestlings) a predator normally less dangerous than a hawk, such as a cat, is responded to with the call usually given for hawks. The bird advertises its fear in the presence of different stimuli at different times. It does not specify what the stimuli are, but instead, what its probable response is or will be. Although a bird does not specify directly that it has a territory, it regularly indicates different degrees of readiness to behave aggressively to inappropriate individuals of (usually) its own species—and so defines a region, which can be labelled its territory, that it will protect from intrusions. A forager bee does not state that it has found food at a certain place; rather, she describes a direction of flight (perhaps a flight she is likely to make again shortly) and, on request, provides a sample of the food. On the other hand, an ant that lays an odour trail only between

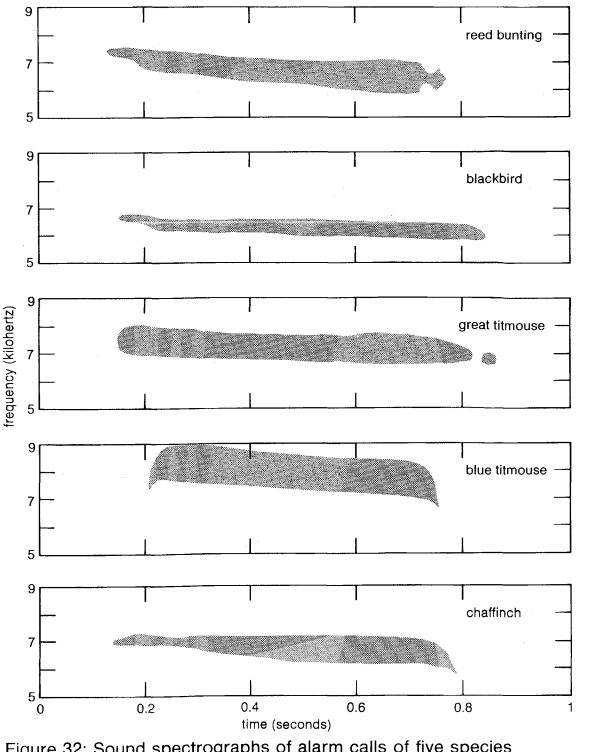


Figure 32: Sound spectrographs of alarm calls of five species of European songbirds showing similarity and broad frequency pattern. Each call is a single, simple, slightly descending note, lasting 0.6 to 0.8 second.

a food source and its home colony is providing information that food is present; such narrow specific use for a display, however, is the exception, at least among birds and mammals.

Evalution of communication. In the evolution of communication within a species, one normal constraint usually is that the exchange of information must be useful to both the communicator and the recipient. The behavioral patterns evolved by the communicator enable him to transmit information that increases the probability of a social response suited to his needs. A recipient evolves the tendency to respond to this information only when the response suits his needs, which often differ, at least superficially, from those of the communicator. When their needs are not compatible, lack of selection pressure for the recipient to respond appropriately usually removes any advantages for the communicator, or at best yields an evolutionarily unstable situation of misinforming, to which recipients are always counter-adapting. Yet, in certain relationships between individuals belonging to different species, particularly in predator–prey relations, selection pressures for providing misinformation are such that it is a widespread phenomenon. In Batesian mimicry, potential prey species advertise that they are actually unpalatable or dangerous to predators, which either learn to leave most of them unharmed or evolve avoidance responses. Other potential prey species, lacking defense mechanisms or unpalatability, may mimic the behavioral and morphological specializations of the unpalatable ones, thus surviving by providing predators with false information. The predators, of course, are under evolutionary pressure to develop means of distinguishing the true from the false information.

Evolutionary precursors of displays

The evolutionary process by which the transmission of information between members of the same species can become specialized has been studied by the pioneer European ethologists Konrad Lorenz and Nikolaas Tinbergen. They have established that displays are specialized activities that have evolved from precursors, or predecessors, of several types. Important among these are intention movements—*i.e.,* incomplete performances of acts, such as taking flight or turning away, which are usually performed by an animal not quite committed to a given course of action. Through the process of ritualization, components of some intention movements become exaggerated and divorced from their directly functional roles. The exaggeration is often evident in an increased conspicuousness (perhaps with the concurrent evolution of morphological badges, to which the display draws attention) or an increase in the conspicuousness of only some components, thus creating a difference between the display and its evolutionary precursor.

A second source of precursors in the evolution of displays lies in inconspicuous but complete movements, such as eyebrow-lowering movements that help protect the eyes; such movements have been incorporated into the facial expressions (frowning) of numerous primate species. Other movements, such as jabbing motions of attack, are ritualized by being aimed in a stiff and often repetitive fashion away from their customary targets; these are called redirected activities. Still other movements appear to be occurring outside their customary functional contexts, as if displaced. Called displacement activities, they remain perhaps the least understood, particularly since it is not always clear that they are as functionally irrelevant as they sometimes appear. Displays apparently derived from displacement activities often resemble the activities devoted to individual maintenance—self-grooming, feeding, and drinking. Such behaviour often occurs in close conjunction with other kinds of socially relevant activities and so is perhaps easily available for evolutionary specialization. Further, some maintenance activities, which are directly related to the physiological results of exertion (as when a bird ruffles its feathers to cool its body) and thus to active social encounters like chasing and fighting, are commonly exaggerated as components of display behaviour. Other autonomic (involuntary) responses, such as urination and defecation by thwarted or frightened mammals, may have been sources for the evolution of some marking displays.

The precursors of ritualization are more easily imagined for visible displays and perhaps tactile displays than they are for other forms, although there is no reason to believe that the evolutionary process is fundamentally different in any case. Chemical displaying seems highly specialized: the chemicals *per se* have in most cases probably originated from metabolic waste products, and the acts of releasing different chemicals (the displaying) may have evolved from precursors classifiable as individual maintenance activities or "autonomic responses." The evolutionary precursors of vocal displays are also conjectural, and many extant vocalizations undoubtedly arose from preexisting ones. Vocalizations must have arisen originally from some form of noisy, controlled breathing; Darwin's suggestion that the breathing patterns of terrified or sexually aroused animals would provide a source for specialization has not been bettered.

Evolutionary origin of vocalization

The evolutionary process of ritualization yields two somewhat distinctive classes of displays. As described above, much ritualization functions to yield a display distinctly different from the act from which it was derived. The act, perhaps a movement preparing a bird to take flight, remains in the behavioral repertoire of the species, serving its original functions, and, to be fully effective, the display must be distinguishable from it. There are, however, cases in which the form of the act is not altered, but its frequency of usage is in one of several ways. This is possible primarily in cases in which the evolutionary precursor is not maladaptive when done to excess (with respect to its original function). Cases in which increased frequency of performance occur are known primarily from social acts, whereas acts that are transformed in the process of ritualization may have social or nonsocial precursors. The best known example of ritualization through increased usage is what is known in mammals as allogrooming and in birds as allopreening—care given by one individual to the condition of the body surface of another individual. In highly social birds and mammals this occurs much more frequently than is necessary for cleansing of the fur or feathers, is done among animals that have bonded relationships, and is often expressed asymmetrically with respect to some feature of the social organization of a species—that is, in one species, subordinate individuals may groom dominant ones more than the former are groomed by the dominants, but in another species the reverse may be true. In addition to being a sanitary procedure, allogrooming apparently expresses the acceptance of bond-limited relationships by both the groomer and the groomed.

The evolutionary process of ritualization operates within a number of limits in addition to those imposed by the process of communication. Each species has a history, in which the origins of its attributes are, ultimately, products of genetic chance.

Closely related species evolve different solutions to the problems and opportunities of communication from those of more distantly related species. Each lineage has developed a working system based on the opportunities it has received, or, having failed to develop such a system, has become extinct. The products of ritualization are not ideal solutions to communication tasks; rather, they are practical ones. The similarities and differences of displays among species contain clues to phylogenetic relationships, although, like all other such clues, they are most safely used when their full functional significances are understood.

Nongenetic differences

By no means do all of the differences among the communication patterns of different species result from the different genetic peculiarities of the different lineages. Species differ in the nature of their social behaviour. Birds, for instance, may be paired through the nesting season and flock in the remainder of the year; be paired only briefly, followed by dispersal on individual territories; be paired for life (year round); or utilize various other types of social structure. The great range of patterns of social organization determines many things about the sorts of behavioral interactions occurring among individuals and, hence, the functions required of communication patterns. The form of social organization is likely to be set, in part, by characteristics of the species' habitat, and the habitat thus must indirectly influence the directions of ritualiza-

tion. Habitats also have a direct influence on ritualization of form, because they differ very much in the degree to which they obstruct information or mask it by noise. The environment of most species also contains other species, some of which communicate with similar displays, creating a need for specific distinctiveness in form.

A basic limitation exists in the nature of the sensory receptor organs available to different kinds of animals. Social insects make much use of chemical and tactile signalling, but visual signals are relatively more important to fishes and visual and auditory signals to birds. Among mammal groups there is considerable specialization, but on the whole mammals make considerable use of all sensory modalities in display behaviour.

(W.J.S./Ed.)

Types of animal societies

To understand social behaviour more fully, it is necessary to examine it throughout the range of animal life. W.C. Allee, in his classic book *The Social Life of Animals,* distinguishes two major types of animal societies. One is the parental, or familial, society, in which parent and offspring stay together for varying lengths of time. The other is the pair bond, or club, society, composed of individuals that come together from different families. This type was much emphasized by the 19th-century English philosopher Herbert Spencer because it corresponded to his social Darwinist ideas. The social Darwinist does not like to admit that a weak son can win out if he has powerful parents; but recent work with rhesus monkeys shows clearly that the son of a high-ranking mother tends to be protected by his mother and hence gets to the top of the hierarchy even if he himself is a weakling. Parental societies are very common.

PARENTAL SOCIETIES

Parental societies are found at all levels, from the cell to the monkey troupe. All animals provide for their young in some way. In every animal there is a period when the young is part of the parent and receives materials from the parent. Later, the young may partly or completely separate from the parent; in some animals, the more or less separate young is then helped by the parent, or helps it.

Parental behaviour among simple organisms. Even some of the simplest organisms show colonial aggregations of the parental type. Some viruses form inclusion bodies in the cells they attack; these bodies are thought to be colonies of daughter viral strands. Other viruses form ordered arrays.

Bacterial colonies

Bacteria, only a few steps up the evolutionary scale beyond viruses, also show parent–young colonies. Diplococci, which can cause pneumonia, are dot-shaped bacteria that have two daughter cells in each group. Streptococci form chains, and staphylococci arrange themselves in grapelike clusters. In all of these, and in a large number of other colonial bacteria, the offspring that are produced by a dividing parent generally stay together for some length of time.

Protozoa, a few steps beyond bacteria, also show parental sociality. Many reproduce by simple division and hence give the daughters help only before the split. Under difficult conditions, protozoans commonly form a protective "cyst" and divide within it. In such divided cysts 2, 4, 8, 16, 32, or even more daughter cells may associate until the cyst "hatches."

Some protozoans form definite colonies in addition to or in place of cysts. *Volvox* and many other slow-moving or sedentary colonial protozoans show differentiation or division of labour between cells of a colony. In *Volvox,* the forward cells have large eyespots and a few rear cells take care of reproduction (Figure 34, top).

Sponges and coelenterates

It is almost certain that sponges evolved from colonial flagellate protozoans. Sponges are integrated networks of cells, some of them amoeboid (amorphous) and some flagellate. It has been shown that if a sponge is strained through cloth so that the cells are separated, they will reunite and form new sponges so long as a flagellate collar cell can rejoin an amoeboid cell. The sponge is thus on the border between colonial organization and integrated multicellular organization (Figure 33, top). One advantage of integrated multicellular organization, with different types of cells performing different functions, was probably that the sponges could become much larger than the largest multinucleate or even colonial protozoans and thus could capture these protozoans. This type of organization also provides strength: some cells can hold on in swift currents, while some can secrete skeletons and others concentrate on food getting. Thus cooperation gave sponges and similar multicellular animals an advantage in competition with even the largest and most aggressive single-celled animals.

Douglas P. Wilson

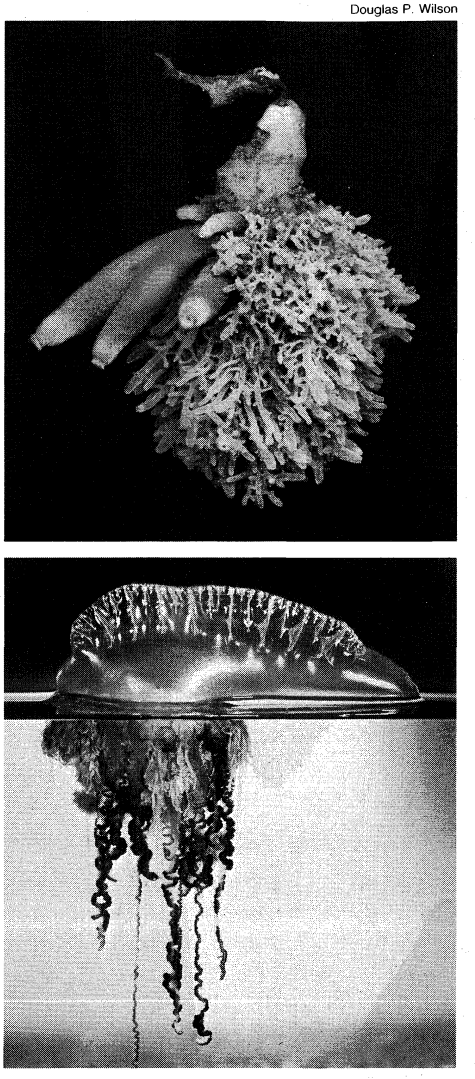


Figure 33: (Top) A colony of three species of sponges, approximately three-quarters life-size. At top of colony is small bread-crumb sponge (Halichondria); branching mass at right is *Leucosolenia complicata;* tubular sponges on left are *Scypha ciliatum.* (Bottom) Portuguese man-of-war (*Physalia physalis*), the body of which may be as much as 12 inches long.

The colonial organization of cells into protozoan colonies or into multicellular animals will be referred to below as the "colonial-1" stage (Figure 34, top), in contrast to the colonial organization of attached multicellular animals—the "colonial-2" stage. The colonial-2 stage is well developed in successful aquatic animals just above the sponges; the coelenterates (hydroids, jellyfishes or medusae, sea anemones). Coral reefs bear witness to the success of colonial-2 growth in other coelenterates (Figure 34, bottom). Many different types of free-swimming colonies, such as the dangerous fish-killing Portuguese man-of-war (*Physalia*), exhibit huge complexity on the colonial plan (Figure 33, bottom). In corals and other colonies, the original individual is linked to its offspring in a network. Food material and chemicals are often exchanged between

individuals over the tubes of this network. Often the different individuals in this network show division of labour. Sometimes, as in *Obelia,* there are only reproductive and feeding individuals. In *Physalia,* there are swimming individuals, stinging ones, and many others, including one that serves as a gas-filled float. The interdependence and communication in such a colony is so extensive that the colony seems almost to be an individual and is sometimes called a superorganism.

Despite the great success of sponges and coelenterates, the main line of evolution goes onward through colonial-1 animals. The link was probably wormlike animals somewhat like present-day flatworms. Most flatworms and higher worms show very little association between parents and young.

From flatworms to insects. The main line of evolution leading from flatworms to insects shows little parent–young cooperation or colonial development. The entoproct Bryozoa, or moss animals, form treelike colonies like those of corals and thus show type-2 coloniality; but it is not certain whether the Bryozoa actually belong to the line leading to insects. A few wheel worms or rotifers, such as the free-floating *Conochilus volvox,* form colonies. A few annelid worms, such as sybellid fan worms, bud off chains of individuals in a manner like the flatworm *Catenula.* This is a common method of breeding in some annelids; the special rear "worm," or "epitoke," breaks off and swims to the surface, where it releases sperm or eggs and dies, often in huge swarms of epitokes as in the Samoan palolo worm (*Palola siciliensis*).

Typically such worms as roundworms and earthworms strew their eggs about or attach them to something as soon as they are fertilized or brood the eggs only briefly. A few nemertine worms are viviparous—*i.e.,* they produce live young. The annelid worm *Ctenodrilus* is said to be truly viviparous, the nutrition of the young coming via maternal blood vessels. Most mollusks take little care of their young. One chiton, *Callistochiton viviparus,* gives birth to young that have undergone development in the ovary. In a few bivalves such as the European oyster (*Ostrea edulis*), the eggs develop in the gill filaments. Most squids release single eggs or chains of eggs, but some members of the octopus group stay near their eggs and remove debris from them. The paper nautilus, *Argonauta,* forms a paper nautilus shell and the mother takes care of her eggs in it. At times the male hides in the shell.

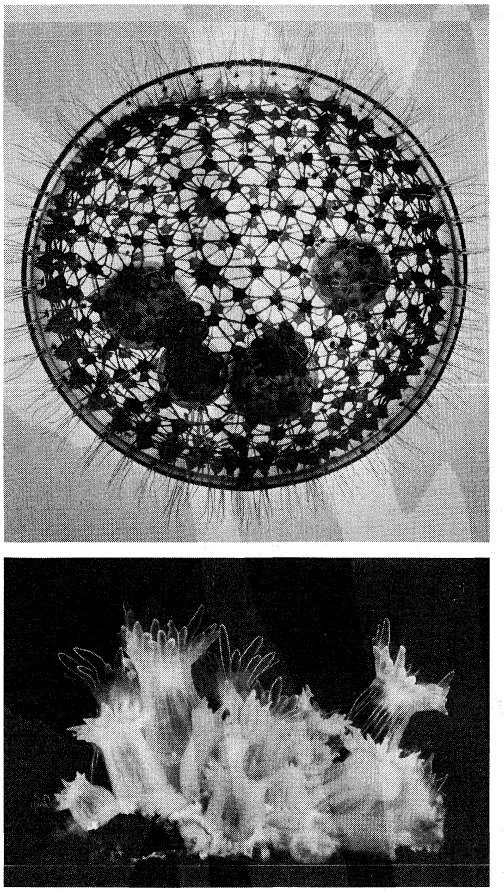


Figure 34: (Top) Glass model of protozoan colony Volvox (magnified about 40 ×). (Bottom) Colony of coral animals *Astrangia danae* (magnified about 1.7 ×).

Peripatus and its relatives, the onychophorans, are intermediate between annelid worms and arthropods and have well-developed parental systems. Some Australian forms lay eggs, but others keep the eggs inside until young hatch; many of these are viviparous, giving the young secretions from the uterus.

Not until one reaches such jointed-legged animals as crabs and insects (arthropods) does one find much extended association between young and their parents. A few scattered arthropods still have no parental care other than production of eggs. The female walkingstick casually drops eggs as she moves about. Many ostracods and copepods, and many of the edible shrimps, shed eggs into the ocean waters. Most arthropods, however, care for their young briefly.

Parental care among arthropods

Scorpions are all viviparous or ovoviviparous (eggs developing in the mother), and many carry their young about. The female pseudoscorpion of the leaf litter often builds a little nest, and the young get nourishment from her in a belly pouch. The female solifugid, or sun scorpion, makes a burrow for her eggs and then brings prey to the young after they emerge from the burrow. The female whip scorpion attaches eggs to herself and carries them until the young go through several molts; she dies as soon as they leave. Spiders generally weave a silken case for eggs and young. The female wolf spider carries her young on her back. Some young spiders build a family or community web together. The harvestmen and mites mostly lay eggs in the environment, but some mites carry eggs until they hatch. Some ticks deposit egg masses, and hundreds of young seed ticks may stay together, to the dismay of a human when such a mass drops on him and starts to spread. Sea spiders (pycnogonids) are strange, for the male takes the egg mass from the female and cares for the eggs until they hatch, or slightly longer. Most crustaceans briefly brood their eggs, or eggs and young, often in special pouches on the body of the female.

Millipedes often form a nest; the young *Spirobolus* later eat the material of that nest. Some female millipedes coil about the eggs for several weeks. Many centipede females brood eggs, but others do not. In symphylans, often considered a link to insects, the female carries eggs in cheek pouches.

Insect families

Social insects. Insects show the greatest development of family structure among animals. Most so-called insect societies are, strictly speaking, families. Sometimes they are called colonies, but the individuals are not directly attached to each other as in the "colonies-1" of protozoa and multicellular animals or the "colonies-2" of corals. They might be called "colonies-3" because they are "attached" by chemicals as well as by social behaviour. The young stay with both parents or with the mother and form social organizations of high complexity. Social behaviour of this type is known among the thrips (Thysanoptera), Zoraptera, book lice (Psocoptera), web spinners (Embioptera), termites (Isoptera), and roaches (Blattoidea) of the more primitive insects, and in some groups of the higher insects—the aphids and lace bugs of the Heteroptera, and especially, in the ants, bees, and wasps of the Hymenoptera. Some of these groups bear little resemblance to the families of vertebrates. A critical observer might say of insect societies that the parents enslave their first children or their sisters, frequently with "drugs," and thereby ensure better care of their later ones.

Societies of lower insects are simple. The female (or male in the case of mole crickets, *Gryllotalpa*), works hard to build a nest and to protect its first offspring. The offspring may reciprocate by helping to build a colony web under a stone or leaf or under tree bark, as in the web spinners and book lice. In wood roaches, such as *Cryptocercus*

Figure 35: Bivouac of army ants (*Eciton*) between trees, which are about 14 inches apart. Inset is detail of bivouac magnified slightly larger than life.

Carl W. Rettenmeyer

punctulatus of the southeastern United States, the young must stay with the adults, because all have symbiotic protozoans inside them that digest wood cellulose: at every molt the roach loses all its protozoa (because the linings of the fore- and hindgut are also molted) and must eat the feces of another roach or die.

In termites, the male and female lose their wings after a dispersal flight and dig a cavity in which they raise the first young. These young have their sexual maturation inhibited by chemical secretions from their parents; instead of reproducing themselves, they work hard to make more chambers and get food for the next young. They often get fecal material from each other full of symbiotic protozoa to digest wood and also use the fecal material to build houses. The more advanced forms masticate wood and grow fungus on the pulp produced in that way. The young have a division of labour, some being workers and some soldiers; there are also nasutes, which have snoutlike processes that eject a sticky substance used in warfare to protect the colony. Such colonies may become huge and build houses higher than a man's head. The first parents are not so much the leaders of the colony as egg-producing machines cared for by their first offspring. Eventually some of the slaves achieve their freedom when the chemical secretion from their parents runs short; they develop wings, fly off, and start new colonies of their own.

Some beetles show behaviour approaching the social. A British rove beetle defends its eggs and young against intruders. Dung beetles dig burrows and store dung for their larvae. The male and female burying beetle cooperate to dig away the soil underneath small dead animals; the female feeds her larvae on regurgitated food. Bark and ambrosia beetles dig tunnels in wood and grow fungal spores; the female feeds her young on pieces of fungus while the male keeps away other males.

Some moths and butterflies associate in the caterpillar stage. Social caterpillars, such as the tent caterpillars (Lasiocampidae) and the larvae of small ermine moths (*Yponomeuta padella*), make webs similar to those of colonial spiders but use them only to hide in rather than to catch prey.

Bees and wasps

The origins of social behaviour can be seen in bees and wasps. There are solitary bees and wasps, all of which prepare a protected place for the egg and later the larva. Some "gall wasps"—as in "gall aphids" and some mites—sting plants, which then provide fleshy galls for the young larvae. The parent often provides food for the larva. The tarantula-killer wasp will sting a huge spider and store it in a drugged state by the egg. The "parasitoid" hymenopterans lay an egg on or in a wandering caterpillar to parasitize it. Some bees or wasps return to put a new spider or other food source into the nest after the first food has been eaten, a process called progressive provisioning. From this it is only a short step to having a single female care for several young in a compound nest, as in *Polistes* wasps, and another short step to having sisters or young stay around the nest and help care for the later young. In some insects, such as wasps of the genus *Polistes,* this is done by having the first or strongest female harass or dominate the later or weaker ones. Their sexual growth is repressed and they cannot lay eggs as long as the dominant female is there. Chemical dominance, or drugging, is the next step; in the more social bees and ants, chemicals produced by the queen are actually needed by the workers, and exchange of food and drugs (trophallaxis) is regular.

Polymorphism and behaviour

Division of labour often occurs in ant and bee societies. Ants are often polymorphic, with small individuals working in the nest and medium or medium-large ones working outside; huge-headed individuals become protective soldiers or even use their heads as plugs to stop up the nest entrance to all besides members of the colony. Honeybees have division of labour by age—the youngest bees feeding larvae, older ones building the comb, and still older ones flying out for nectar and pollen and bee glue. Many of the polymorphic differences are apparently determined by food, as when the new queen bee gets royal jelly regularly, while the smaller workers get royal jelly for only a few days. Other differences are genetic, as in the case of the male ant or bee, which comes from an unfertilized egg.

These family societies of insects are diverse and successful. Termites and ants are among the most common tropical insects, bees and wasps among the common subtropical and temperate ones. The houses of termites—earth castles with shingled construction to shed rain and porous outer layers to control carbon dioxide and humidity—are equalled in their intricacy only by those of man. The honeybees communicate with chemicals and dances to tell each other the distance and direction of flowers.

The ferocious defense of the nest by wasps and hornets avails them little, however, against the onslaughts of marauding hordes of army ants (tribe Ecitonini) in the tropical forests of the Western Hemisphere and of hordes of driver ants (tribe Dorylini) in Africa. The army ants (see Figure 35) do not eat trees or people, as early stories would have it, but they tear apart arthropods. The driver ants, which have scissor-like mandibles that cut flesh, can tear apart humans if given the chance. These are probably the largest of the familial or "colonial-3" societies. It has been estimated that a large colony of the army ant *Eciton burchelli* includes 1,500,000 individuals, and the colonies of the driver ant *Anomma wilverthi* probably contain up to 22,000,000.

The leaf-cutter ants of the Western Hemisphere live in huge underground colonies. They, along with termites and a few beetles and moths, are agriculturalists. The leaf-cutter ants cut strips of green leaves and make a paste of them in which they grow fungus. Their underground chambers may reach several yards. It is hard to realize that such huge colonies are extended families.

From bryozoans to humans. In the other great line of evolution, which leads to man, the social use of the family has taken a different tack. Where the first line began with actively moving, wormlike individuals and ended with drugged, tiny individuals in huge families, the line that leads to humans begins with colonial, attached animals of the general appearance of corals but of the structure of worms and ends with social animals in which families play an important but relatively small role.

Wormlike animals

Some early wormlike animals evidently settled down on the floors of ancient oceans, and to protect themselves had

to develop shells (as in the lamp shells or brachiopods) or colonies with specialized defensive members (as in the moss animals or bryozoans). Brachiopods are solitary and shed their gametes into the seawater. Moss animals form colonies in which there is direct or partially impeded exchange of body fluids. Their societies show more division of labour than do termite colonies. There are feeding individuals, reproductive individuals, special whiplike individuals (vibracula), and bird-head individuals (avicularia) that hit or bite other animals settling on the colony.

It seems incongruous to suggest that active vertebrates developed from tiny wormlike animals living sedentary and colonial lives, but the future does not always belong to the strongest, biggest, or fastest animals of a given age. The moss animals, with their tiny encrustations or filamentous colonies, are internally much advanced over the more abundant corals.

One major side branch of this line of descent does lead to the nonsocial echinoderms—starfish, brittle stars, sea cucumbers. Few of these animals take care of their offspring, and even their gametes tend to be shed broadcast into the seawater. Some sea stars, brittle stars, and sea urchins of the Arctic and Antarctic brood their eggs. In some, as the brittle star *Amphipholis squamata,* the young are attached to the mother and get nourishment from her. Some sea cucumbers, equally divided among cold-water and warm-water forms, brood their eggs externally or, as in *Thyone rubra* of California, inside. A few sea lilies or crinoids brood eggs or young.

There is also little parental care in several of the wormlike side branches of this line of descent. Pogonophoran worms, which even lack a digestive tract, sometimes brood eggs in their tubes in the ocean mud. Arrowworms (Chaetognatha) occasionally carry eggs about, but most release them into the ocean waters where they swim. The arrowworms are successful predators, but the line to vertebrates leads for the most part through the colonial or sedentary filter feeders—the pterobranchs, the acorn worms, and the tunicates or sea squirts.

The pterobranchs are sedentary, colonial wormlike animals, with a central stalk in some colonies but no direct connection in others. The individuals of the latter wander in and out of the colony tubes. Pterobranchs are related to burrowing solitary acorn worms, the hemichordates. All these animals release their eggs and sperm rather casually into the ocean.

The sea squirts (Tunicata) are mostly soft spongelike masses that cling to rocks or pilings in the sea. Most shed eggs into the sea, but some brood eggs or young and release them partly grown. The young are either free-swimming tadpole-like animals or are budded from the adult to form a colony.

The line of descent up to this point is, curiously enough, closely associated with colonial animals, while the line that led to insects produced rather few colonial animals. It has been suggested that advances made during periods of coloniality may produce better free-living individuals and vice versa; the inference is that drastic new changes in a colonial animal can be perpetuated because of feeding by the rest of the colony and later be incorporated in a viable free-living combination.

Some biologists, including Darwin (with his vested interest in competition), have suggested that the sea squirts and all the other colonial animals are unimportant sidelines in evolution. They suggest that the mainline passed through nonsocial, competitive, free-living, wormlike and, later, tadpole-like animals. Wormlike animals led to animals like arrowworms, to the tunicate tadpole, and then to fishlike animals such as *Amphioxus.*

Certainly the next steps were through free-swimming animals with relatively little family life—cephalochordates (lancets) and the jawless fishes. Modern lampreys make a nest for external fertilization, but the hatching larvae make their own way to live as filter feeders in the muddy debris of stream bottoms.

Fish and amphibians

Many jawed fishes take little care of their eggs or young, but there are some major exceptions. Many sharks, skates, and rays give birth to live young, and some have placentas to nourish them. Some fishes make nests; the Siamese fighting fish (*Betta splendens*) and others make bubble nests at the surface, while sticklebacks (*Eucalia*) and a mormyrid fish (*Gymnarchus*) make weed nests, and such fishes as salmon dig spawning nests in the bottoms of streams. Stickleback males fan their eggs until the young hatch, making sure there is enough oxygen. Other fish, such as the black-chinned mouthbreeder, fan their eggs by keeping them in their mouths. Other mouthbreeder fish periodically spit out the young fry and take them back in until the young are feeding well. Mouth-breeders include a freshwater catfish (*Loricaria typus*) and some marine catfishes (*Galeichthys felis* and its relatives), plus cardinal fishes (Apogonidae). There are several fish of a dozen different taxonomic assemblages that bear their young alive, among them the surf perches (Embiotocidae) and the mollies (Poeciliidae). Some have a placenta-like arrangement to nourish the young until birth.

While care of young by the male is frequent in fishes, it is rare among invertebrates (where the sea spiders are the only major example). Care by the males frees the female to obtain more food and hence raise more young per unit of time, which may be necessary in the event that food is difficult to find.

Male care of the young is well developed in lungfishes, predecessors of land vertebrates. Most salamanders and frogs, on the other hand, are not very good parents. The male Surinam toad (*Pipa pipa*) presses the eggs into the back of the female, and the young of *Rhinoderma darwinii* go through development in the vocal pouch of the male. The live-bearing frog (*Nectophrynoides*) of Africa, in which the young hatch within the mother and remain with her for protection, is another exception to the general rule that amphibians release their eggs and care for them little.

Higher vertebrates

Among reptiles the best parental care is in alligators and some crocodiles, where the female makes a mound of dead leaves or sand and stays around to protect the eggs, to release them when they hatch, and to guard the young for perhaps as long as a year.

Birds are well known for parental care. Most build nests, incubate eggs, and care for young. The males commonly help the females in this. Often young birds stay with their parents a year or more, and numerous examples are known of the young of one brood helping to feed the young of later broods. Often the young help the parents defend a "group territory," as among Australian bell magpies (*Cracticus*) and Mexican jays (*Aphelocoma ultramarina*).

The most advanced parental society yet recorded among birds is that of the ocellated antbird, which follows swarms of army ants in order to capture insects they flush in the neotropical forests. This bird's young stay with their parents for several months, then go and find mates, but return to their parents periodically for several years. The

Irven DeVore

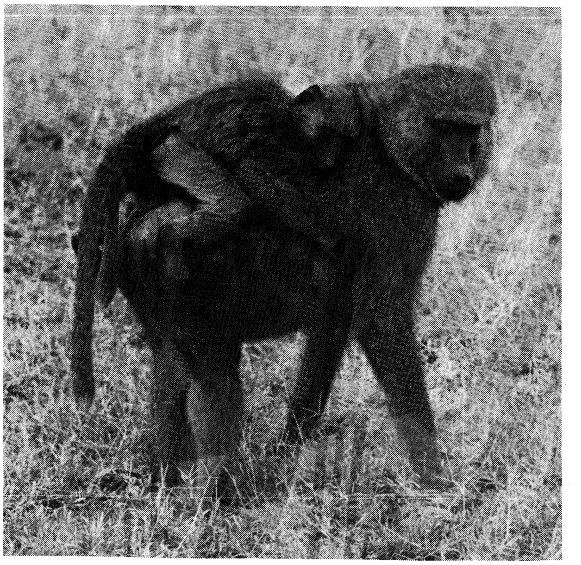


Figure 36: Mother olive baboon (*Papio anubis*) and young.

young bird and its mate are accepted as part of the extended family; they are not chased away as often as are unrelated birds.

An even greater organization of parental care is found in mammals, except that the males seldom help care for the young. The young are usually nourished before birth by the placenta of the mother, except in the egg-laying duck-billed platypus, the spiny echidna, and in marsupials. The young of marsupials are born prematurely and grow in the pouch of the mother for long periods. All mammal young, even the platypus and spiny echidna, must lap or suck milk produced by the mammary glands of the mother. This ensures that there is a strong family association between mother and young (Figure 36).

Commonly, groups develop around a mother and may be joined by other such groups and by males to form bands, troops, and herds. Troops of monkeys and apes are basically families or grouped families. These and wolf bands include males, which may help to raise young. The "extended families" of humans lead to tribes, states, and nations.

Nationalism as a force in human affairs is commonly related to mother and home and family, as well as to interlocking family relationships even more complex than those of ocellated antbirds or wolf societies.

SOCIETIES WITH SEXUAL BONDS

Nonparental social relationships fall into two categories, sexual bonds and nonsexual bonds. Normally, only the latter can involve members of two or more species. Sexual bonds lead in many animals to parental bonds, of course, but differ in that the bond is normally between offspring of different families. The reason for this is that the main advantage of sexual union is to combine the good genetic features of two different lines. Some young, of course, will have the bad features of both lines and will be eliminated—a wastage tolerated in nature as a necessary expense.

Most animals depend on elaborate behaviour patterns to bring the male and female and their gametes together at the right time, rather than using the seemingly more certain processes of asexual reproduction, virgin birth, hermaphroditic self-fertilization, or male parasitism. Many marine animals shed eggs and sperm into the water or into special nests but do so only when chemically stimulated by the presence of substances from the opposite sex. Others use their "internal clocks" to release gametes only at a certain time of day or year. Samoan palolo worms form special gonadal body sections by budding, then release them to swarm in huge numbers at the surface of the ocean on a schedule set by the Moon. The grunion (*Leuresthes tenuis*), a small Pacific fish of the silversides family, is famous for its males and females meeting to fertilize eggs high on the beach at the highest tide each month and at the highest waves of that tide as well. The synchronization of male and female requires them to have an internal lunar clock, such as that known for the colour changes of fiddler crabs.

The functions of courtship

Courtship must basically ensure two things: that the correct male and female get together at the right time with as little loss as possible; and that the offspring have the best possible chance to survive. Ensuring these two things has led to elaborate courtship patterns in animals. Where different kinds of animals that look, smell, or feel alike coexist, each individual must be especially careful not to hybridize with the wrong species. Otherwise it will waste its eggs or sperm in a union that will produce no young, or young that are malformed or maladapted for the world into which they emerge. Animals often develop complicated odours, colours, or voices as means of identification. There are many species of fruit flies of the genus *Drosophila,* and to avoid mismatings, each species has its own pattern of waving the wings by the male. The sight and sound of the wrong pattern of waving is enough to cause a female fruit fly to flee (see also above *Reproductive behaviour*).

Where there is only a limited breeding area or a patchy environment in which the good areas are restricted, strong male–female differences are advantageous. In these cases the males have to advertise, often with song, to help females locate good areas. Males of species whose courtship displays are performed in groups usually have to compete more directly with each other and thus tend to develop large size or striking colours, overpowering scents or voices, or other exaggerated features. Female domestic chickens, sage grouse, and baboons tend to copulate mainly with the lordliest and most dominant males. This is not so evident in other animals, in which the females tend to mate with the male holding the best territory. The dominant male is likely to be the oldest one, the one that has proved he can survive and hence is "fittest."

Supermasculinity

The success of the "supermasculine" or lordly males in a few species may be advantageous only to those males. A few lordly males often usurp the few suitable places to breed, as in male sea elephants (*Mirounga angustirostris*) on Pacific beaches and in male red-winged blackbirds in North American marshes. If the lordly male blackbirds are eliminated, however, other males come in and the females breed with them just as quickly.

The supermasculine males in these species, full of pomp and strutting, seldom care for their young. Baboon males, it is true, do at times stop fights between lesser animals and drive away leopards or other predators. But usually the female takes care of the young herself. It may even be an advantage if she is "ladylike" and unobtrusive, so that the lordly male may draw predators away from the young. Keeping the male away from the young may also allow the female and her young more food. If environments are limited in food, keeping excess males out of the breeding area clearly is an advantage.

Supermasculinity, as well as being correlated with female care of the young, is also associated with polygamy. The mating of several animals of one sex with a single individual of the other sex tends to be associated in birds and mammals with great differences between the sexes. Serial polygamy, or the mating of an animal of one sex with several of the other sex at different times, may also occur. Promiscuity, or the mating of each female with several males and each male with several females, tends in supermasculine animals to resemble serial polygamy. Monogamy, or the mating of each male with one female, tends to occur mainly in animals with little difference between the sexes.

Having many mates does not necessarily mean that an animal is more social than if it has only one mate. In most cases, the polygamous male spends much time driving away other males and little time courting his females. His females spend little time with him, because they are busy raising many young—with little care for each. Among African weaver birds (*Ploceus*), monogamous species of the forest have smaller clutches than do polygamous ones of the savanna.

The whole system of lordly males, ladylike females, polygamy or serial polygamy, and multiple young tends to occur mainly in animals with restricted and undependable sources of food and other necessities. One investigator found that males of the long-billed marsh wren (*Cistothorus palustris*) in Washington, where their marshes varied greatly in quality, had several mates; females went to males with good territories and left neighbouring males, with poor territories, mateless. Another investigator found that in Georgia, where the marshes were everywhere about equal in food supply and nesting cover, the long-billed marsh wrens were usually monogamous. The correlation suggested by many recent studies is this: sexual dimorphism and diethism (behavioral differences) arise in animals in which environmental opportunities are restricted due to undependability or local distribution.

NONFAMILIAL SOCIAL BONDS

Social behaviour also occurs among animals that are not necessarily related by parental or sexual bonds. A "flock" or "band" of animals may be formed of only one family in some cases, but often several families or individuals join together.

Spacing. As previously noted, social organization within a species may be shown not only by the presence of clumping or positive movement of individuals but also by even spacing resulting from negative movements away

from each other. Sociality is shown more by the presence of a definite spacing than by nearness.

Absence of spacing among simpler forms

There is little evidence for social spacing in protozoans and simpler beings, such as viruses and bacteria. Most recorded groups of unicellular or lower animals are probably parental or sexual groups. There seem to be few social interactions among microorganisms, but this apparent dearth of social behaviour may be an artifact introduced by the disturbance of observation.

Most colonies of sponges and coelenterates seem to be parental colonies or aggregations in favourable sites rather than nonfamilial societies. There is little information on nonfamilial social organization among colonial-2 animals, even among those that can move, such as the colonies of Portuguese man-of-wars or the planktonic rotifer or seasquirt colonies. Most worms and their relatives are not known to react to each other or to form social structures of the "flock" type. Little is known about mollusk organization; but the simple fact that mollusks do not pile up on top of each other suggests that they are capable of negative social reactions.

Barnacle larvae prefer to attach next to another member of their own species or on a place where one of their species has just been removed. They avoid settling on top of another member of their own species, even though they readily settle on a member of another species. They react in part to chemicals that are specific to their species. Barnacles, then, aggregate in colonies but within a colony space themselves out. If the multicellular animal is a colony-1, the coral colony a colony-2, and the bee society a colony-3, the barnacle society may be called a colony-4. Colonies-4, in which the interactions take place between unrelated, unmated, and unattached individuals, are regular in arthropods. In many cases the existence or nature of a colony-4 is not obvious or is problematical. Bees form colonies-3 and perhaps also colonies-4, for colonies of bees space themselves out in the environment and avoid establishing themselves too close to other active colonies; but there is no evidence that bee colonies avoid getting too far apart, although if they did then mating between bees from different colonies would become difficult.

Colonies-4 are known among other arthropods. Tube-dwelling amphipods form colonies but chase each other and avoid getting too close. They show the phenomenon of "personal space," or "individual distance," as surely as do swallows sitting on a wire, among whom a new arrival settling will sometimes cause shifting outward by individuals on both sides. Individual distance or personal space is a fairly sharply defined space around each individual that can be penetrated by another individual without hostility only after certain overtures.

Territoriality

The amphipod colonies also show the phenomenon of territoriality. Territoriality differs from personal space in that a territory is centered on some object outside the body of the animal itself. The male bitterling (*Rhodeus sericeus*), a European fish that lays its eggs inside a freshwater clam, will chase male intruders away from his clam even if the clam gets up and moves several feet. The male house finch (*Carpodacus mexicanus*) of North America chases other males away from his female, wherever she moves. More often, however, the external reference for territory is a fixed plot of ground, a nest hole, or some immovable set of objects. The animal may chase out intruders or tolerate them in the territory, but in his territory he is in charge. Work with bicoloured antbirds in Panamanian forests suggests that a territory may be defined as "an external referent in which one animal or group dominates others that become dominant elsewhere." A pair of bicoloured antbirds permits others on its territory, but as soon as the pair crosses a boundary line into another territory, it becomes subordinate to the pair of that territory.

Territoriality is known to exist among insects such as dragonflies and ants, some fish, a few frogs, some lizards, most birds, and many mammals. It probably exists, at least in chemical forms, in tube worms of a muddy beach or rocky shore, and perhaps even among sedentary protozoans.

Swarming. Personal space and territoriality are definite negative signs of coloniality-4, but there are also positive signs. Mutual repulsion is only part of sociality, for mutual attraction must also exist. Mutual attraction is found in many arthropods above the barnacles. Some of the best studied examples are the mating swarms of ants, flies, midges, and, especially, fireflies. Another example of mutual attraction is the migratory horde, of which the African migratory locust (*Schistocerca migratoria*) is the best studied example.

The synchronized communal displays of the fireflies of Thailand are among the most impressive exhibitions of the insect world. The gatherings of thousands of males on the mangroves of the coastal swamps have been described as a city of pulsating glitter, every male anticipating the flicker of his neighbours and flashing in unison with them. Communal mating displays of this type are common in birds such as manakins, sage grouse, ruffs, and birds of paradise, and are called lek displays. Presumably, the communal display enables females to find the males more easily. Seldom do bird leks approach the numbers or synchrony of a firefly lek in Thailand, but several male manakins sometimes cooperate in a synchronized cartwheel dance. It is difficult to explain why males competing with each other for mates should help one another, but the phenomenon has been well documented.

Communal displays

The legendary colonies-4 of the migratory locusts are far larger and more impressive than the migratory colonies-3 of the army and driver ants. When food is abundant, the locusts disperse widely and grow up in what is called the "solitaria" phase. As the locusts crowd and encounter each other, they begin to change colour and enter the "gregaria" phase, in which they look so different that they were once considered a separate species. The "gregaria" locusts behave differently too, for they are excitable and social. They begin to march over the ground as food supplies diminish. Finally they take wing in huge hordes and fly downwind to a low-pressure area where rains have recently fallen. Here they descend on crops and other vegetation, eating it to the ground before flying to the next low-pressure area. The extinct migratory hordes of passenger pigeons were apparently similar in their effects on vegetation.

Many other examples of flocks occur in higher animals, especially insects and vertebrates. Most show little internal structure. The migratory hordes of armyworms that devastated midwestern corn fields in the United States before the days of synthetic insecticides, the swarms of male and female palolo worms in the ocean, the feeding swarms of sharks, and the dense schools of herring are all examples of flocks with little internal structure. The Austrian ethologist Konrad Lorenz calls them "anonymous flocks," because it matters little if individuals change places and bonds are seldom individualized. When the fish school turns, it is like an army platoon turning to the flank, for the former side fishes are now the leaders.

The mating choruses at frog ponds provide an example of an auditory lek. Some salamanders congregate to breed, generally by sight and by odour, in running streams. Snakes form winter dens in which there may be hundreds of individuals rolled up in balls.

Flocks, herds, and packs

Among birds, pigeons (Columbidae), starlings (Sturnidae), and various blackbirds (Icteridae) form dense flocks that wheel about in the sky or mill along the ground during foraging, the rearmost flying ahead to become briefly the leaders. Shorebirds and gulls gather on mud flats or elsewhere to feed. Such birds as the brown creepers (*Certhia familiaris*) reduce the winter cold by clumping together at night. Flocks of geese are made of many families of geese. Many birds gather into huge roosts, containing thousands or even millions of individuals in the case of the red-winged blackbirds (*Agelaius phoeniceus*) of North America. Tricoloured blackbirds (*A. tricolor*) of California are even more colonial than redwing blackbirds when nesting and by sheer force of numbers push into colonies of the more dominant redwings and displace them. The nesting colonies of queleas (*Quelea quelea*) in Africa contain millions of nests. There are many such colonies of birds, such as the phenomenal colonies of seabirds that are found on islands throughout the world.

Mammals often form herds or packs. Many herds are more structured than bird societies, simply because many

mammal groups are combined families plus males. Huge migratory herds of wildebeest and zebra wander the African plains. Each herd of zebra includes many familial harems that are held together by individual males. The herd of Scottish red deer is a matriarchal group led by an old female and composed of her extended family plus other extended families like hers.

Hunting mammals often have even more structured groups. A male and female wolf (*Canis lupus*) and their offspring form a hunting pack that may fuse with another pack or split apart. African hunting dogs (*Lycaon picta*)

George Porter—The National Audubon Society Collection/Photo Researchers

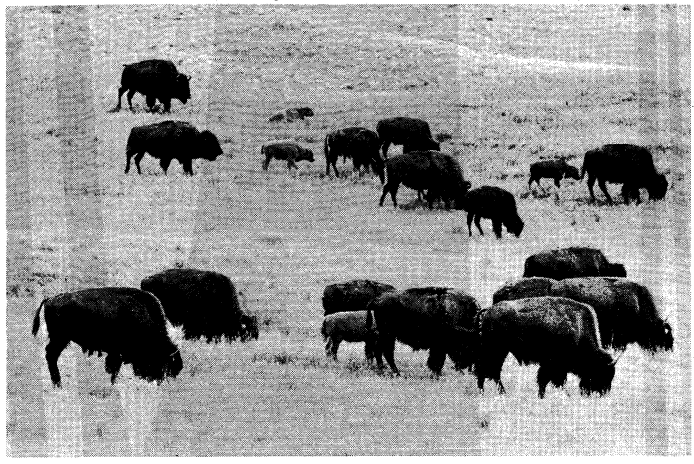


Figure 37: A herd of American bison (*Bison bison*).

and hyenas (*Hyaena* and *Crocuta*) form similarly flexible hunting packs, which are said to be even more effective than lion groups at running down prey.

Primate troops are often as complex as societies of hunting mammals or more so. Superimposed on their parental and sexual bonds is a group organization based on the occupation of a given area. The troop may also accept animals from outside. The society of baboons (Figure 36), as studied on the plains of Kenya, is very highly organized. Around the edge of the troop as it moves are the subadult males, watching carefully for predators and snatching bites of food when they can. Inside, the playful groups of juveniles stay close to the central hierarchy, a cooperating group of two or three big males that keep the subadult males at the periphery. The males of the central hierarchy are replaced, as they get old and toothless, by brash young males moving in from the periphery. With the central hierarchy march the females and their infants, protected both by the big males and by the peripheral younger males. The society is integrated by mutual grooming sessions, in which the big males get most of the grooming, and by domination by the big males.

INTERSPECIFIC ASSOCIATIONS

Individual animal interactions. Associations of animals often include more than one species. These groups may be called colonies-5 or colonies-0, since there is recent evidence that the nucleus and other parts of the cell were originally symbiotic viruses and bacteria. The most intimate form of interspecific association is that known as symbiosis, or mutualism, in which dissimilar organisms live together (see BIOSPHERE: *Biotic interactions*).

Multicellular animals often live on or beside other animals. Small fish live in the tentacles of some jellyfish, sea anemones, and even the Portuguese man-of-war, yet are able to evade or inactivate the stinging cells. Some hydroids and worms live on tubes of other worms or on the shells of other invertebrates. The tubes of worms and shrimp often harbour other worms or fish. At times, the animal will live only on one kind of shell and is found nowhere else. Paleontologists have found some evolutionary sequences in which such an animal first lived on rocks and shells of a wide variety of types, then developed larger forms that lived on one type of animal and probably got food from it.

Complex associations. Slave making is a kind of social relation that verges on parasitism. Certain kinds of ants raid colonies of other kinds of ants, carry off their young, and raise them as slaves. The slaves are perfectly socialized members of the colony and probably do not even realize that their social behaviour is misdirected. They exchange food and drugs with their captors as willingly as they would have with their own species, had they been reared by their own workers.

Associations in ant colonies

One of the most complex associations is that of animals around army and driver ants. In the huge colonies of army ants live dozens of other kinds of insects, millipedes, and mites. Some help clean debris below the nests; some have chemicals that allow them to fool the ant security guards and enter the colony. Mites, beetles, and others ride on the ants or march in their columns. Some may help the ants by cleaning them or by giving them chemicals they crave. Others are like wolves in the fold, eating the food of the ants or even their larvae. The swarms of ants are also waited upon by many kinds of flies and birds. The flies lay eggs on insects and spiders fleeing from the ants. The birds capture animals flushed by the ants. In the tropics of the Western Hemisphere, nearly 50 kinds of birds follow the army ants persistently and would probably die without them.

Some of these associations may not be social in any accepted sense of the word. Humans are not social with rats and fleas merely because they live with them. Some interspecific associations, however, are definitely social. These include the mixed flocks of antelope, zebra, and wildebeest on the African plains, for example, and mixed flocks of birds throughout the world.

One can travel for hours across the African plains or through a tropical forest scarcely seeing an animal and suddenly be surrounded by a herd or flock of many kinds of animals. Usually each animal eats a different kind or type of grass or fruit or insect, although sometimes there is overlap in the foods taken or ways of feeding. The flock moves along together, not spending much time at each concentrated food source. Often it includes parental groups (colonies-2 in the case of mothers carrying young, or colonies-3 after the young separate).

The fact that forest flocks are usually of several species rather than one probably reflects the fact that forests have more species of animals, and hence each species has less of a food supply and must not allow competitors of its own species about it, even its own offspring. In less complex habitats, there are usually very few species, and the animals can tolerate their own young even though these are competitors. They may even use their young—to detect predators (in the case of baboons) or to build a "city" (in the case of bees and ants).

Ecosystems

When a forest is destroyed and begins to grow back, the first animals that come in tend to be kinds that are solitary and very antagonistic or uncommunicative. Later, flocking animals become more common, although they still resist groups of outsiders. Finally, as the mature forest re-establishes itself, one finds mostly paired animals that do not keep their young with them. These paired animals tend to associate with pairs of other species to form mixed flocks. Eventually, every animal links itself with every other in the system, forming what ecologists call a complex "food web," "ecosystem," or "web of life." It is gradually being recognized that such a web is socially cooperative as well as socially competitive. The ecosystem eventually approaches a stage of "colony-6," or what the French biologist and philosopher Teilhard de Chardin called the noosphere.

Dynamics of social behaviour

COSTS AND GAINS

Social behaviour among humans is often regarded as an end in itself, the expression of a basic drive that has no necessary purpose. Biologists doubt that any animal has social tendencies without some adaptive advantage.

The costs. Social behaviour and communication not only take an animal's substance and energy; they impede feeding, drinking, and other inputs necessary for life. The

first cells that associated with other cells to form multicellular filaments lost the ability to absorb on the side by which they were attached. Perhaps the reason most multicellular filaments occur among animals that are attached to the ground or to some other surface is that such animals lose less proportionately than members of free-floating aggregations; attachment on one side to the ground already limits their input. Locomotion is impaired if animals must stay together. The single-celled ciliates could not readily have evolved into higher organisms, because dividing them into many joined cells would have slowed down these fast-moving predators. A speedy golden plover trying to stay with other shorebirds in a mixed flying group near shore constantly turns back to keep with them; it is impeded by its social tendency.

Social behaviour also attracts enemies. Groups of animals have epidemics, while solitary animals seldom do. Many disease-carrying parasites spread much more easily at times when animals are together. Some rabbit fleas are even adapted to the hormonal cycles of the rabbits, so that they reproduce at the times of year the rabbits are reproducing and hence are social.

Predators, like parasites, often have an easier time if animals are crowded together; the animals are often busy reacting to each other and the predator can sneak up without being observed. Their communicatory systems may even attract predators. Tuna prey specifically on fish in schools; a small hawk in tropical America (*Accipiter superciliosus*) mainly on mixed bird flocks.

Conflict and its avoidance

Social behaviour increases the number of interactions between animals and thus the chances of conflict. The conflicts may be solved by fighting, by patterns of dominance and submission (peck orders), or by mutual avoidance. Mutual fighting and mutual avoidance have the same result—a partitioning of resources for which the animals are competing.

The gains. Against these disadvantages of being social, it is possible to set a number of clear advantages. They fall into six broad categories, corresponding to the six possible kinds of animal behaviour. By social behaviour animals gain: (1) food and other resources, (2) reproductive advantages, and (3) shelter and space. They are enabled to avoid (4) physical and other small hazards, (5) competitors, and (6) predators or other large dangers. The first and third of these gains are reactions to desirable things of small (1) and medium to large size (3) respectively; the fourth and sixth are reactions to undesirable things of these sizes.

Food. The value of being social in getting food is obvious in the case of hunting bands. Cooperative hunting has been found among wolves and African hunting dogs, hyenas, lions, killer whales, porpoises, cormorants, white pelicans, pairs of eagles and of ravens, tuna when chasing small fish, army ants, primitive and modern men, and many other animals. Animals that hunt cooperatively can trap, chase, and tear apart prey that would otherwise be too fast, strong, or large for them. In African hunting dogs the chase is run by the leader of the pack, but the rest keep the antelope or other prey from dodging left or right and also help fall on it when the leader catches it. Flocks of wattled starlings (*Creatophora cinerea*) fly after African migratory locusts and surround one group after another, eating every trapped locust from each group. In army ants, the individuals are bound to each other by chemical "trail substances" so that no individual gets far from the group; when one finds prey, it grabs it and emits an "alarm" chemical that causes nearby ants to grab, bite, and sting so that the prey is overwhelmed within seconds. They then tear the prey, usually insects or other arthropods, limb from limb and carry it back to the nest.

Cooperative hunting

Interspecific groups of birds are sometimes food-getting societies. Drongos (*Dicrurus* species) of Africa flush much food, and other birds follow them to get it. Honey-guides (*Indicator* species) of Africa lead honey badgers or men to bee nests and eat wax after the mammals break open the nests for honey. Hawks have been known to follow railroad trains for the same reason, and hornbills and hawks follow monkeys. The birds, lizards, flies, and other animals that associate with army ants offer other examples of interspecific food-providing associations. One animal may steal food from another, as American widgeons (*Anas americana*) steal grass from redheads (*Aythya americana*).

In addition to hunting and flushing food cooperatively, animals sometimes lead others to food or teach them to use it. Parents, especially among mammals, often teach their young to hunt or lead them to food. Animals that must migrate or depend upon seasonally available resources often depend on others to show them what foods are good and where. Vultures and jackals flock to carcasses on the African plains. American robins (*Turdus migratorius*) in California have been observed learning to use certain berries after flocks of cedar waxwings (*Bombycilla cedrorum*) came through and started eating the berries. Tests with a tape recorder show that the recorded calls of some birds that follow army ants will attract unrelated kinds of birds that also follow ants. In the laboratory, some animals learn to push a lever for food by watching others get food that way and learn to avoid distasteful foods by watching others cough it up. In studies of Japanese monkeys (*Macaca fuscata*), the habit of washing potatoes before eating spread from the younger to older monkeys of a troupe. In Britain, a few titmice learned to open milk bottles and drink cream; the habit spread much too rapidly to be a genetic change. (Coverage of learned behaviour is found in LEARNING, ANIMAL.)

Mating behaviour

Reproduction. The reproductive advantages of social behaviour have mostly been discussed earlier. It was noted that sex is a way of combining desirable genes from different lines, genes that otherwise might slowly or never get together. In many lines of animals, parental behaviour is clearly useful in protecting or teaching the young. This normally requires the adult to have fewer young. The careful parent loses in time and energy and number of offspring but comes to prevail in evolution if it has more descendants than does a careless parent that lets its young die. The careless parent prevails if it can get more young out by caring for each one less; some parasites are careless parents because each of the young needs little care and a large number must be produced to get to an extremely distant host.

Shelter. Social behaviour is often used in habitat selection and shelter selection, even to the extent of making it possible for the animal to improve the environment it finds. Male birds that later will fight with each other over territorial boundaries gather first at areas where they hear another bird singing, rather than hunting for a more isolated (and probably unsuitable) place. Certain beetles that attack pines put out a scent that attracts other beetles; only as a result of concerted attack by all beetles can the protective pitch of the tree be reduced so that all may enter. Movement to a flock is a good way to find a patch of habitat or a shelter. It has been suggested that flocking increases the accuracy of migration, since the average direction taken by a flock is more correct than the individual directions taken by individual birds. Small flocks of European starlings returning to a California roost were less accurate in their direction than large flocks. Cooperative building of structures is well known in humans, prairie dogs, rats (whose tunnel systems rival the catacombs in complexity), beavers, certain weaver finches, wasps, bees, termites, and many others; symbiotic use of structures occurs in many animals.

Hazards. Social behaviour can also help animals avoid small hazards. This includes avoiding heat or cold and wet or dry situations as well as preening or grooming to keep off dirt, parasites, and other small environmental hazards. A goose cleaving the air for its companions at the front of a V-shaped flock, a parent bird brooding its young or sheltering it from the Sun, a group of creepers roosting together to help each other survive the cold winter night, and a group of baboons grooming each other to pick off ticks furnish other examples.

Competitors. Dangers from competition are avoided by agonistic behaviour. The five basic types of agonistic behaviour are aggressive display (threat), submissive display (appeasement), attack, avoidance, and fighting.

Agonistic behaviour

Social aggressive display is not common. Males of a troupe of howler monkeys all yell at a neighbouring troupe to make them keep their distance. Baboon males in the

"central hierarchy" cooperate to keep aggressive young males from winning, backing each other up with threats. Social attack occurs in some birds and mammals that keep group territories and may lead to fighting if the other group attacks or threatens.

Highly social submissive display and escape also are not common. A baboon troupe may retreat as another moves in at a water hole. But even when a single animal retreats from a competitor it is a social act. Territoriality is certainly a system in which an animal defends its right to be dominant in part of its home range. The basic feature of territoriality, however, is not aggression in a certain area but submission outside that area. The common idea that strong animals survive and the weak do not is true only in the short run, for in a few generations all reproducing animals are equally strong. Strong animals will begin to lose if they keep on chasing others. An animal that keeps too large a territory will spend more time chasing away intruders than it will in eating or reproducing, unless it can get others to help it. Bees get help by drugging the nonreproductive members of their colony. Most animals limit themselves so that the territory of the most dominant animal or group of animals never exceeds about twice the size of the least dominant animal or group of animals of that species. Most often the young animal has a small territory but defends a larger one as he gains experience, then gradually loses it as he reaches old age.

Grouping and flocking

Predators. The final reason for social behaviour, and one of the most important, is to avoid predators or other large dangers. Just as animals can sometimes overcome large prey by grouping to attack it, so they can sometimes overcome large predators by grouping to defend against them. Cooperative and spirited attacks upon predators occur in most animals that protect their young and are a regular phenomenon in gull and tern colonies, in baboon troupes, in bees and wasps, and many others. "Mobbing" is a similar phenomenon in which the attack is not carried all the way to the predator but so harasses it that it departs or at least is prevented from getting its prey. The massed effect of many mobbing birds is more intimidating to a predator than is mobbing by one or two birds.

Grouping also helps against predators because a predator is distracted by the "confusion effect" of so many shapes, sounds, or smells. Human hunters know that one cannot shoot a duck out of a flock by aiming at the flock; the shot is more likely to pass between the birds than if the hunter aims at one of them. Similarly, hawks have been seen to drive through a flock and miss every bird. Successful predators either dive to break up a flock and then grab a separate animal or pick off an outlying one at the start. Butterflies on tropical trails also swirl up in a confusion effect from a mud puddle. The phenomenon is caused by the difficulty the eye or other sense organ has in analyzing or following very complex motions that cross each other.

Another advantage of the group or flock is that many eyes can see a predator more quickly than can one pair of eyes. Ornithologists have found that social birds are nervous outside of a flock and must spend too much time watching to be able to forage effectively. Certain species that forage by peering in dense vegetation are especially in danger and must associate with other species that look about more actively in open foliage. The peering species often are good at yelling and perhaps help the other birds by scaring or disturbing predators. This suggests that a social organization may have many reasons for being.

DEVELOPMENT FACTORS

Instinct and learning

As noted above, behaviour changes somewhat in the course of evolution. Biologists commonly call the genetic determinants of behaviour in a line of organisms instinct. Every behaviour pattern, however, can be changed somewhat by the individual animal in the course of its experience. The old view that instinct and learning are two different types of behaviour is seldom accepted today, even though some kinds of behaviour certainly have little learning superimposed.

The real question is how social behaviour develops. It is possible to breed animals for aggressiveness or nonaggressiveness, and by further crosses to study the inheritance of behaviour. Mouse strains that show different degrees of aggressiveness are easy to develop. Biochemical imbalances also affect behaviour: in many animals an oversupply of male hormones causes aggressive and antisocial behaviour. Pituitary hormones, especially luteinizing hormone, have the same effect in other animals, such as starlings.

Stimulation of the brain or removal of part of it gives evidence of a structural basis for behaviour; stimulation of the hypothalamus produces many social behaviour patterns, such as sexual activity and aggression. There are even "pleasure centres" that the animal will stimulate on its own, if given a bar to press that sends a shock to its head. Sexual centres are one of the "pleasure centres" that rats are fond of stimulating.

Another approach is to isolate the animal and see if it still develops a particular behaviour. Young pigeons reared in cardboard tubes will fly soon after release, showing that practice is not necessary. Some young songbirds reared in isolation develop normal songs, and many develop normal calls. Many songbirds must listen to songs of their own species at a particular age, however, to learn them.

An animal may develop social behaviour while still in the egg or mother. Baby ducklings peep to the calling mother from the egg. An animal may develop social behaviour soon after it emerges or at some critical period later. A young duckling follows the first object it sees, be it a duck or a duckling or the hand of the experimenter. Young birds, ants, and some mammals "imprint" on the first object they see to such an extent that they may court it or show agonistic behaviour to it later. A mother goat given a lamb in exchange for her kid soon after birth will adopt the lamb and drive away her own kid when it is returned to her.

Later learning also influences social behaviour. Mice that experience defeat learn to run rather than fight; the opposite holds for mice that win. Most animals, however, start at the bottom of a peck order and take defeats in stride, later becoming the dominant animals if they manage to survive. It has been found that association with other young monkeys helps a monkey to behave properly in sexual activity later, although many learn to copulate properly without this opportunity.

THE EVOLUTION OF SOCIALITY

Social behaviour as an adaptation

The fact that bees and ants form complex societies, more complex in some ways than those of apes, shows that social behaviour occurs in small animals as well as in large ones, in animals with small brains or large ones, and in both major lines of evolution. If bacteria can be rather social and humans rather solitary, there is no reason to suppose sociality is more advanced in evolution than is solitary life.

Social behaviour is instead an adaptation to certain environmental opportunities. The evolution of sociality can be glimpsed in the line that leads from the earwig through wood-eating cockroaches to termites, or in the line from solitary bees to social ones. Communication systems also evolve, as may be seen in the line leading from dully coloured monogamous crows to brightly coloured birds of paradise and plain bowerbirds in New Guinea. In this system, the male bird of paradise is brightly coloured to attract the crowlike female. The bright male also attracts predators. The bowerbirds have lost the bright plumage; instead they make elaborate maypoles or bowers decorated with flowers to attract females. One bowerbird even paints the walls of his bower, using a mashed berry or a straw stained in berry juice. These bowerbirds have become safely coloured, for they have replaced bright plumage with bright objects.

The ecological maturity and regularity of a habitat seem to determine to some extent how social its inhabitants will be. Among African weaver finches, the forest-living ones are solitary and monogamous; birds of savannas and marshes flock and nest polygamously; and those of very dry habitats tend to be relatively solitary. The same phenomenon has been noted for cats; leopards of the forest and cheetahs of very open country tend to be less social than lions of open savanna areas. Antelope, deer, monkeys, and apes show similar differences. The general rule

is that, as an environment grows up from the level of bare ground to that of savanna and finally forest, the solitary animals are replaced by social ones and then by solitary ones again. In the forest, however, one-species societies decline in importance and societies of several species form. The same things happen as a marine community goes from bare rock to the complexity of a coral reef.

Societies of the same species, therefore, seem adapted for intermediate habitats that are in transition between bare ground and forest. It may be that the reason for this is that most intermediate habitats are unstable, likely to be limited in space or time. (E.O.W./Ed.)

BIBLIOGRAPHY

General works: KONRAD LORENZ, *Er redete mit dem Vieh, den Vögeln und den Fischen,* 6–8th ed. (1952; Eng. trans., *King Solomon's Ring: New Light on Animal Ways,* 1952; reduced photographic reprint, 1961), is a highly recommended popular treatment; and P.R. MARLER and W.J. HAMILTON, *Methods of Animal Behavior* (1966). Semipopular works include DESMOND MORRIS, *The Naked Ape* (1969) and *The Human Zoo* (1969); NIKOLAAS TINBERGEN, *The Herring Gull's World: A Study of the Social Behaviour of Birds* (1960), *Curious Naturalists* (1958, reprinted 1968), highly recommended to all interested in behaviour, and with the EDITORS OF LIFE, *Animal Behavior* (1965), an excellent résumé of ethology.

Nature and patterns of animal behaviour: IRENAUS EIBLEIBESFELDT, *Ethologie, die Biologie des Verhaltens* (1966; Eng. trans., *Ethology: The Biology of Behavior,* 1970); ANNE ROE and GEORGE G. SIMPSON (eds.), *Behavior and Evolution* (1958); KENNETH D. ROEDER, *Nerve Cells and Insect Behavior,* rev. ed. (1967); and CLAIRE H. SCHILLER (ed.), *Instinctive Behavior* (1957). More advanced are ROBERT A. HINDE, *Animal Behaviour: A Synthesis of Ethology and Comparative Psychology,* 2nd ed. (1970); P.H. KLOPFER, *Behavioral Aspects of Ecology* (1962), and (comp.), *Behavioral Ecology* (1970); KONRAD LORENZ, *Evolution and Modification of Behavior* (1965); WLADYSLAW SLUCKIN, *Imprinting and Early Learning* (1964); and E.L. BLISS (ed.), *Roots of Behavior: Genetics, Instinct and Socialization in Animal Behavior* (1962).

Unlearned behavioral reactions: V.G. DETHIER and ELIOT STELLAR, *Animal Behavior: Its Evolutionary and Neurological Basis,* 2nd ed. (1964); R.A. HINDE, *Animal Behaviour: A Synthesis of Ethology and Comparative Psychology,* 2nd ed. (1970); and W.H. THORPE, *Learning and Instinct in Animals* (1963).

Studies emphasizing stereotyped responses include J.D. CARTHY, *Animal Navigation: How Animals Find Their Way About* (1956); G.S. FRAENKEL and D.L. GANN, *The Orientation of Animals: Kinesis, Taxes, and Compass Reactions* (1940, reprinted 1961); and J. LOEB, *Forced Movements: Tropisms and Animal Conduct* (1918).

Wide-ranging surveys of instinctive behaviour include IRANAUS EIBL-EIBESFELDT, *Grundriss der vergleichenden Verhaltensforschung: Ethologie* (1967); PETER R. MARLER and WILLIAM J. HAMILTON, *Mechanisms of Animal Behaviour* (1966); WILLIAM H. THORPE, *Learning and Instinct in Animals,* 2nd ed. (1963); NIKOLAAS TINBERGEN, *The Study of Instinct* (1951). Papers and books devoted to special aspects of instinctive behaviour are T.H. BULLOCK, "The Origins of Patterned Nervous Discharge," *Behaviour,* 17:48–59 (1961); and in *Nervous and Hormonal Mechanisms of Integration* (1966); ROBERT A. HINDE (ed.), *Bird Vocalizations: Their Relation to Current Problems in Biology and Psychology* (1969) and *Non-Verbal Communication* (1972); KONRAD LORENZ, "The Innate Bases of Learning," in KARL H. PRIBRAM (ed.), *On the Biology of Learning* (1969); and CLAIRE H. SCHILLER (ed.), *Instinctive Behavior: The Development of a Modern Concept* (1957).

Biological periodicity is treated in the following works: F.A. BROWN, JR., *Biological Clocks* (1962), a brief, elementary account of the nature of, and problems related to, the phenomena of plant and animal periodicities; E. BUNNING, *The Physiological Clock,* rev. 2nd ed. (1967), particularly good for the earlier literature and classical views; J.L. CLOUDSLEY-THOMPSON, *Rhythmic Activity in Animal Physiology and Behaviour* (1961), best for its treatment of ecological significances; J.E. HARKER, *The Physiology of Diurnal Rhythms* (1964), emphasizing the periodicities associated with day–night; and B.M. SWEENEY, *Rhythmic Phenomena in Plants* (1969). In a class by itself is A. SOLLBERGER, *Biological Rhythm Research* (1965), a comprehensive, well-indexed, and documented reference work containing more than 2,500 literature citations, well-balanced in distribution over the subject. F.A. BROWN, JR., J.W. HASTINGS, and J.D. PALMER, *The Biological Clock: Two Views* (1970), is a succinct presentation of the two currently held alternative hypotheses, internal and external timing of the rhythms, together with the evidence upon which they are based. G.G. LUCE, *Biological Rhythms in Psychiatry and Medicine* (1970), is a readable, informative, extensively documented account of occurrence and significances of daily, lunar, and annual periodicities for man. The phenomenon of photoperiodism receives thorough coverage in STANLEY D. BECK, *Animal Photoperiodism* (1963), a concise general presentation of the subject on mammals, man, birds, and insects; *Insect Photoperiodism* (1968), an advanced book on all aspects of photoperiodism in insects, with an extensive bibliography of the original research; and BRIAN LOFTS, *Animal Photoperiodism* (1970), a popular work written for the general audience.

Basic behavioral activities of individuals: J.A.C. NICOL, *The Biology of Marine Animals* (1960), gives a good introduction to the classification of feeding patterns with typical examples. W.C. ALLEE *et al., Principles of Animal Ecology* (1949), a classical survey of the entire field, is still useful as an introduction to the relations of animals to their food environment and gives many examples. The best review available of the physiology of feeding behaviour is C.F. CODE (sect. ed.), "Control of Food and Water Intake," in *Handbook of Physiology,* sect. 6, vol. 1 (1967), which is largely though not entirely restricted to vertebrates. A methodologically important systems analysis of the behaviour of vertebrate and invertebrate selective feeders may be found in C.S. HOLLING, "The Functional Response of Predators to Prey Density and Its Role in Mimicry and Population Regulation," *Mem. Ent. Soc. Can. 45* (1965), and "The Functional Response of Invertebrate Predators to Prey Density," *ibid. 48* (1966). An introductory survey of the relations of insects to their food plants (and feeding behaviour of insects in general) is presented by V.G. DETHIER in P.T. HASKELL (ed.), *Insect Behaviour* (1966). More detailed and specialized material on this point is contained in J. DE WILDE and L.M. SCHOONHOVEN (eds.), *Insect and Host Plant* (1969).

R.B. CLARK, *Dynamics in Metazoan Evolution* (1964), discusses locomotor patterns of invertebrates with hydrostatic skeletons. J. GRAY, *Animal Locomotion* (1968), provides a synthesis of most aspects of invertebrate and vertebrate locomotion. H. HERTEL, *Struktur, Form Bewegung* (1963; *Structure, Form and Movement,* 1966), treats the mechanics of flight and undulatory swimming. E. MUYBRIDGE, *Animals in Motion* (1899, reprinted 1957), is a classic work on the subject. R.A.R. and B.J.K. TRICKER, *The Science of Movement* (1966), is a useful introduction to the physics of locomotion.

R.A. HINDE, *Animal Behaviour: A Synthesis of Ethology and Comparative Psychology,* 2nd ed. (1970), and S.P. GROSSMAN, *A Textbook of Physiological Psychology* (1967), both contain valuable material on several aspects of avoidance behaviour.

SILVIO GARATTINI and E.B. SIGG (eds.), *Aggressive Behaviour* (1969), discusses modern theories and experiments on aggressive behavioral responses of animals. KONRAD LORENZ, *Das sogenannte Böse* (1963; Eng. trans., *On Aggression,* 1966), is a highly readable account. S.A. BARNETT, *A Study in Behaviour* (1963), is a general book with particularly useful sections on aggression.

The following works cover all aspects of animal migratory behaviour: G.M. ALLEN, *Bats* (1939), a classic book on bat biology with a chapter on migration; F. BOURLIERE, *Vie et moeurs des mammifères* (1951; Eng. trans., *The Natural History of Mammals,* 3rd ed. rev., 1964), a modern review of mammal biology with a chapter on migration; J. DORST, *Les Migrations des oiseaux* (1956; Eng. trans., *The Migrations of Birds,* 1962), a modern review of all aspects of bird migration; D.R. GRIFFIN, *Bird Migration* (1964), a useful summary with a good chapter on orientation; F.R. HARDEN JONES, *Fish Migration* (1968), a standard book, with many classic references; G.V.T. MATTHEWS, *Bird Navigation* (1955), an original attempt to explain bird orientation; R.T. ORR, *Animals in Migration* (1970), a modern review of migration patterns among animals; E.J. SLIJPER, *Walvissen* (1958; Eng. trans., *Whales,* 1962), a modern review of whale biology, including a chapter on migration; A.L. THOMSON, *Bird Migration,* rev. ed. (1942); and C.B. WILLIAMS, *The Migration of Butterflies* (1930), two classic books, still useful; A. WOLFSON, "Ecologic and Physiologic Factors in the Regulation of Spring Migration and Reproductive Cycles in Birds," in A. GORBMAN, *Comparative Endocrinology* (1959), an accurate review of the physiology of bird migration, somewhat out of date but still useful.

Dormancy and related mechanism are treated in H.W. WOOLHOUSE (ed.), *Dormancy and Survival* (1969), a symposium incorporating a wide variety of review articles covering dormancy in a broad spectrum of organisms from bacteria through mammals; K.C. FISHER (ed.), *Mammalian Hibernation: Proceedings of the Third International Symposium on Natural Mammalian Hibernation* (1967), 25 papers dealing directly or indirectly with the topic of mammalian hibernation accompanied by over 1,300 bibliographical references relative to this topic; X.J. MUSACCHIA and J.F. SAUNDERS (eds.), *Depressed Metabolism* (1969), a variety of studies relative to hibernation, hypothermia, and thermic instability; W.V. MAYER, "Hibernation"

(1964), a popular pamphlet concerned with hibernation in both birds and mammals; J.P. HANNON and E. VIERECK, *Comparative Physiology of Temperature Regulation* (1962), deals with temperature regulation in both cold- and warm-blooded animals, including hypothermia and hibernation; C. KAYSER, *The Physiology of Natural Hibernation* (1961), an intensive look at hibernation in birds and mammals, including hypothermia and estivation, with emphasis on functional changes; and C.P. LYMAN and A.R. DAWE (eds.), *Mammalian Hibernation* (1960), hibernation and hypothermia in birds and mammals, including articles on such thermally unstable forms as bears.

All modes of behaviour related to reproduction in animals other than humans are dealt with in the references cited below: MARGARET BASTOCK, *Courtship: An Ethological Study* (1967), an excellent survey; DESMOND MORRIS, *Patterns of Reproductive Behaviour* (1970), a compilation of some classical papers, all by Morris; N. TINBERGEN and the EDITORS OF LIFE, *Animal Behaviour* (1965), a good introduction; ARI VAN TIENHOVEN, *Reproductive Physiology of Vertebrates* (1968), a good survey with an emphasis on the hormonal and neurophysiological aspects of reproductive behaviour, especially ch. 9, 11, 13, and 14; S.A. ASDELL, *Patterns of Mammalian Reproduction* (1964), a comprehensive survey stressing anatomical and physiological aspects; C.M. BREDER and D.E. ROSEN, *Modes of Reproduction in Fishes* (1966), the best modern survey of reproductive behaviour in fishes; E.O. WILSON, *The Insect Societies* (1971), the best general treatment of this group; and JOHN SPARKS, *Bird Behaviour* (1969), an excellent introduction to the reproductive behaviour of birds, with excellent illustration.

Behaviour of animals in groups: W.C. ALLEE, *The Social Life of Animals* (1938), a readable classic, emphasizing peck order and social facilitation at the expense of other aspects of social behaviour; ROBERT ARDREY, *The Territorial Imperative* (1966), a somewhat tendentious discussion of animal territories by a playwright; R.D. BARNES, *Invertebrate Zoology,* 2nd ed. (1968); J.T. BONNER, *Cells and Societies* (1955), a readable account of social life from the howler monkeys down to the cell; T.D. BROCK, *Biology of Microorganisms* (1970); J.H. CROOK (ed.), *Social Behaviour in Birds and Mammals* (1970), several excellent technical summaries of modern research in social behaviour, including a discussion of habitat and society; FRANK FRASER DARLING, *A Herd of Red Deer* (1937), one of the earliest field studies of a wild society, establishing that deer are matriarchal; IRVEN DEVORE (ed.), *Primate Behavior* (1965), the best collection of relatively nontechnical articles on the behaviour of free-living monkeys and apes; S.J. DIMOND, *The Social Behaviour of Animals* (1970), a discussion of recent experiments on learning of social behaviour in domestic and caged animals; JOHN F. EISENBERG, "The Social Organizations of Mammals," *Handb. Zool.* 10:1–92 (1965), a review of mammalian social behaviour that shows it derives mostly from maternal societies; PEGGY E. ELLIS (ed.), *Social Organization of Animal Communities* (1965), a useful set of rather technical articles, concentrating on social behaviour in insects; WILLIAM ETKIN (ed.), *Social Behaviour and Organization Among Vertebrates* (1964), a set of moderately technical articles on social behaviour; E.S.E. HAFEZ (ed.), *The Behaviour of Domestic Animals* (1969), a set of technical articles on animals such as cats, dogs, sheep, and goats, with an excellent discussion of physiology; S. MARK HENRY (ed.), *Symbiosis,* 2 vol. (1966–67), informative summaries of a few of the many symbioses known to occur, from viruses to man; DAVID LACK, *Ecological Adaptations for Breeding in Birds* (1968), a demonstration that the social behaviour of nesting birds depends on their habitats and their foraging; CHARLES DARWIN, *The Expression of the Emotions in Man and Animals* (1872, reprinted 1965), the first major attempt to trace the evolution of facial signals; WESLEY E. LANYON and W.N. TAVOLGA (eds.), *Animal Sounds and Communication* (1960), accompanied by a phono record; RENE GUY BUSNEL (ed.), *Acoustic Behaviour of Animals* (1963), a general survey; THOMAS A. SEBEOK (ed.), *Animal Communication* (1968); provides good coverage of various aspects of the subject; MARTIN LINDAUER, *Communication Among Social Bees* (1961), a fascinating account of evolution in and experiments on social bees; KONRAD LORENZ, *Das sogenannte Böse* (1963; Eng. trans., *On Aggression,* 1966); LYNN MARGULIS, *Origin of Eukaryotic Cells* (1970), a discussion of the origin of cells by symbiosis; HARRIET L. RHEINGOLD (ed.), *Maternal Behaviour in Mammals* (1963), a set of moderately technical articles on the mother–infant relationship in mammals; JOHN PAUL SCOTT and JOHN L. FULLER, *Genetics and the Social Behaviour of the Dog* (1965), a scientific analysis of heredity and learning that shows how they interact; EDWARD C. SIMMEL, RONALD A. HOPPE, and G. ALEXANDER MILTON (eds.), *Social Facilitation and Imitative Behaviour* (1968), scientific articles on imitative learning in animals and men; NIKO TINBERGEN, *Social Behaviour in Animals, with Special Reference to Vertebrates* (1953), a popular account by a founder of ethology, concentrating on birds, fish, and insects; E.O. WILSON, *The Insect Societies* (1971), one of the best surveys of the behaviour of social insects, but omits locusts; and V.C. WYNNE-EDWARDS, *Animal Dispersion in Relation to Social Behaviour* (1962), a polemic reviewing much of social behaviour to support the view that animals practice birth control by means of social behaviour—but control is usually by agonistic reactions.

The Development of Human Behaviour

Humanists and scientists alike have long speculated about the relative importance of endowment and experience, or of "nature" and "nurture," in the development of human behaviour. During the 19th century, developmental psychology emerged as a discipline devoted to the systematic investigation of these factors. The term "behaviour" may be understood to embrace both the expressed and potential capacity for activity in the physical, mental, and social spheres of life. The "development of human behaviour," in turn, encompasses three interrelated phenomena: the genesis, or origins, of behaviour; continuity, or constancy, of behaviour over time; and discontinuity, or change, in behaviour. The concerns of those who study the development of human behaviour comprise, first, the description of human capacities and activities over the entire life cycle; second, the comparison of human capacities and activities at different points in the life span, across different cultural and ecological contexts, and often among different species; and, third, the explanation of the interaction of the endogenous (internally arising) and exogenous (externally arising) forces that guide and influence the development of human capacities and activities. The breadth of this inquiry places scientists who study the development of human behaviour in a position to provide information about the natural history of behaviour, the growth of individual behaviour, the functions of behaviour, and the causes of behaviour—in short, what develops and when, how, and why. As a consequence, some scientists believe that all biological and social inquiry is circumscribed by the study of the development of human behaviour.

The present article treats both the study and the phenomena of the development of human behaviour. Because of the intimate relation between physical and behavioral aspects of human development, a brief discussion of physical changes is offered for each developmental stage. For further treatment of biological development, see GROWTH AND DEVELOPMENT, BIOLOGICAL. For further treatment of particular facets of behavioral development, see EMOTION AND MOTIVATION, HUMAN; LEARNING AND COGNITION, HUMAN; PERCEPTION, HUMAN; PERSONALITY; SEX AND SEXUALITY.

This article is divided into the following sections:

Theory and practice in research

THEORIES OF DEVELOPMENT

The concept of development is complex and is defined by several characteristics, or principles. One is that development must have an orderly, cumulative, and progressive character. Continuity and discontinuity in life reflect systematic, successive adaptation in organization over time. A second principle is that development is not a straightforward empirical concept; if it were, the direct inspection of data would indicate to any observer whether or not development had taken place. Scientists often disagree about development because the concept is actually a postulation. In fact, whether or not development has occurred in any given instance is generally determined by whether the features of the data fit an implicit or explicit concept of what development entails.

Debates about the character of development arise because different scientists recognize different criteria and, in turn, because they are committed to different philosophical beliefs regarding fundamental aspects of the nature of human life. As a consequence, in discussing orientation, theory, and data about the development of human behaviour, it is necessary to consider the different philosophical perspectives, or metatheories, that stand behind scientific thinking.

A metatheory is a theory about a theory; that is, a philosophical view about what constitutes a good and useful theory. Issues of metatheory arise in debates about what sorts of ideas ought to be included in a useful theory of behavioral development. For instance, should the theory stress ideas about genetic inheritance or about environmental influence, or should it stress the intimate relation between the two? Scientists differ on this question because adhering to different metatheories means espousing different views of the nature of life. In spite of the multitude of differences, though, the study of human behavioral development shows the influences of three distinct metatheories.

Three metatheories of behavioral development

Some scientists advocate the view that human beings do not differ qualitatively from other natural phenomena and that human beings are therefore controlled by the same forces that control nature. Since all natural phenomena are composed of the same units (*i.e.*, atoms and molecules), the mechanical laws of chemistry and physics that explain the actions of these basic units provide information about how human beings act as well. The basic model of this mechanistic metatheory is the machine, in which complex phenomena are ultimately reducible to the workings of elementary parts and their relations. Movement of the parts is initiated by an application of forces outside the unit and results in a chainlike sequence of events. Applied to the study of human development, the mechanistic model postulates human beings who are essentially passive and only reactive to outside forces. The individual is at rest until activity is caused by external forces (stimuli) that bring about change (responses). Complex human activities, such as the mental act of problem solving or even the feeling of emotions, can be measured quantitatively (at least in theory) as the action of a multitude of stimulus–response connections.

A contrasting organismic metatheory holds that mechanism is inapplicable to the study of human beings because atoms and molecules within humans fuse to create char-

acteristics that do not exist in isolated parts. A proper knowledge of human beings is lost if parts or elements are studied by themselves; hence, human development must be apprehended in a holistic rather than mechanistic manner. This organismic view applied to the study of human development yields an active model in which individuals are constantly active and in which qualitative change is evident in the individual's action on the environment.

A third metatheory postulates that human development arises from continuous interaction among different levels of organization, including inner physical and biological phenomena as well as cultural and historical events. The interaction of phenomena at all these levels contributes to human development. In this contextual view, all levels subject to analysis are reciprocally related and are constantly changing. For example, biological changes in the person influence psychological functioning, just as social mores influence psychological and biological characteristics.

The metatheories outlined above exert telling influences on how scientists address basic questions about behavioral development. They also drive deep divisions between scientists over several key issues in the character and sources of development. The purpose of specific developmental theories is to comprehensively explain the origins of and the constancy or change in human behaviour. Three broad perspectives have been articulated in response to these two salient issues. Each perspective is based on one of the metatheories outlined above, and each succeeds to a degree in explaining development, although no single perspective accounts satisfactorily for the development of human behaviour. (Reciprocally, development could never be understood by assessing any individual cause–effect relation.)

Two salient issues: origins and constancy or change

Two of these major perspectives were initiated in the 17th and 18th centuries out of epistemological debate, that is, debate regarding the origins and growth of knowledge. Nativists, notably René Descartes and Immanuel Kant, argued that human beings are created in the image of God and are born with a good nature and worldly knowledge. Empiricists, notably George Berkeley and John Locke, argued that human nature and knowledge are not innate but develop out of experience. Thus, one school contended that behaviour emerges, maintains itself, or changes because of endogenous forces, or forces arising from within the individual; such a view implies an organismic model of the world, according to which individuals are seen as dynamic and change is seen as qualitative. The second school contended that behaviour emerges, maintains itself, or changes because of exogenous forces, or those arising from outside the individual; in the mechanistic worldview implied here, all change merely comprises sequences of cause–effect reactions. A third school argues that development takes place through the interaction over time of endogenous and exogenous influences.

The endogenous view. An organismic view (in modern terms identified with genes, heredity, and the processes of maturation) locates the impetus for developmental genesis, constancy, and change within the individual. Development is seen to reflect an unfolding biological program regarded as largely fixed and universal. Among the features of behavioral development, proponents of this view distinguish species-typical tendencies, those that all human beings share by virtue of being human—such as the propensity to cry when distressed or to attend to novelty—from heritable tendencies, those that individuals possess by virtue of their particular genetic endowment—such as activity level or the disposition to distractibility. Of course, these two sets of tendencies articulate, or occur in systematic interrelation.

Two false beliefs are commonly held about endogenous influences in development. The first is that any behaviour with biological origins must be present at birth; to the contrary, behaviours associated with puberty emerge naturally more than a decade after birth. The second is that behaviours with biological origins are necessarily fixed. Rather, they can change by themselves or be influenced by experience; for example, the readiness to cry may vary naturally among babies, but the actual frequency of any baby's crying can be influenced by the parents' responses.

Fields of investigation

The study of endogenous factors in human behaviour falls within several fields of investigation. Behaviour genetics attempts to chart the influence of genetic programming on the development of behaviour throughout the life cycle. In this field, for example, consanguinity studies (studies of blood relatives) bear out the common observation that the more closely individuals are related the more similar they tend to be in certain physical and behavioral traits (*e.g.,* height and shyness, respectively); heritability quantifies the relative contribution of genetics to behavioral expression. The disciplines of ethology (the study of animal behaviour) and sociobiology place the development of human behaviour in its larger evolutionary framework. These disciplines often argue that human behaviour and its development are best understood in the context of their evolutionary history. For example, ethological comparisons reveal the values of such factors as relative physical immaturity and social play for cognitive growth in the young of different species.

Endogenous forces would appear to account most easily for those aspects of development that are sequenced, or time bound, occurring at particular periods in the life cycle. The development of walking is a good example; the sequence of crawling, toddling, and walking and the ages at which these developments occur are consistent enough among individuals to suggest a genetic timetable for their unfolding. Exogenous forces, however, might also explain such behaviour: environments could uniformly pressure children to start to walk and to follow the same sequence and roughly the same schedule in doing so. It is not difficult, therefore, to attribute motive in development incorrectly, especially in cases where only the behaviour can be observed and where that behaviour unfolds over long periods of time.

The exogenous view. The direct alternative to the endogenous view is a mechanistic one, which locates the forces for development in the individual's environment (*i.e.,* in experience and learning). Behavioral genesis, constancy, and change, according to this view, reflect the operation of a small set of simple, elegant, and powerful principles of learning and socialization, which include associative conditioning and observation. Thus, for example, spontaneous behaviours (*e.g.,* babbling in infants) that elicit positive responses (*e.g.,* parental smiles) tend to recur. Of course, development could not proceed were it governed solely by such principles of reinforcement; actual responses over time are never so consistently appropriate as the theory of learning requires, and development that depended on reward would be remarkably haphazard, slow, or limited. Imitation is a kind of omnibus learning that requires only observation, internalization, and reproduction of what is experienced and so is extremely efficient and ubiquitous.

The roles of experience

In the exogenous view, experience influences the development of human behaviour in three basic ways. The most dramatic form of influence occurs when experience induces otherwise undeveloped behaviours to emerge. If behaviours are partially developed, experience may serve to facilitate their further development. Experience may also be necessary simply to maintain partially or even fully developed behaviours; in the absence of experience, emerging or mature behaviours may be lost.

Doubtless, truths are to be found in each of these extreme positions—the one attributing the development of human behaviour to nature, the other to nurture—and there exist examples of genesis, continuity, and discontinuity in behaviour that support each. But during the past century of developmental study it has become clear that neither nature nor nurture alone wholly accounts for development; environment provides individuals with stimuli and experiences, and genetics contributes sensitivities, propensities, and capacities by which individuals avail themselves of those stimuli and experiences. Endogenous and exogenous factors together determine development, and both must be considered.

The interactionist view. Endogenous and exogenous factors combine to influence development according to three basic patterns: normative age-graded influences, normative history-graded influences, and nonnormative life-

Normative and nonnormative influences

event influences. Normative age-graded influences consist of biological and environmental determinants of individual development that are tied to chronological age. They are normative to the extent that their timing, duration, and clustering are similar for most people. Examples include maturation events, such as growth spurts and changes in endocrine system function, as well as socialization events, such as marriage and retirement. Normative history-graded influences consist of biological and environmental determinants of individual development that are correlated with historical time. They are normative to the extent that they are experienced by most people living in the same era. Examples include historical events such as periods of economic depression and even the noticeable change in modern times in the age at which puberty begins. Sociocultural developments include changes in sex-role expectations and child-rearing practices. Both age-graded and history-graded influences vary with time. Nonnormative life-event influences are not directly indexed by time since they occur more or less at random. Examples include illness, divorce, and the death of a spouse. In short, the life span is filled with events that are common to all people (*e.g.,* birth, puberty, and death), peculiar to certain groups of people (*e.g.,* religious or social ceremonies), and specific to individuals (*e.g.,* encountering a certain role model or having a particular illness). As a consequence, all human beings exhibit certain behavioral developments that are universal and others that are unique; moreover, each new generation develops in a world that shares some features with the world of its parents and some with the world of its children but also has features peculiar to its historical moment.

In essence, then, human beings develop as changing individuals in a changing world. They develop in response to events that occur at all times in life. At any one time, phenomena affecting the life course are present at all levels of being—the inner-biological, the individual-psychological, the social, and the historical. Since events at each plane affect events at every other, the forces at each of these planes interact to influence human development. People may in turn influence the environments in which they live as much as environments influence them. A person is both a product of biological, psychological, sociological, and historical forces and a producer of those same forces. This concept of the nature of human development evokes the idea of a dynamic interaction among the many dimensions of development across the life span.

Both specific and general theories of human development have been constructed around this notion of interaction.

Piaget, Freud, and interaction

Three important specific theories are those of Jean Piaget, Sigmund Freud, and the developmental contextualists. Both Freud and Piaget posited universal, biologically driven stages of development that colour the individual's changing knowledge and interpretation of the environment. (See below *Infant and child development.*)

Piaget's theory is concerned almost exclusively with the development of the mental life of the child. He supposed that at different points in the life cycle human beings interpret their encounters with the world in identifiably distinct ways and that at particular points, based on the interplay of maturation and feedback from the exercise of particular strategies of understanding, they become dissatisfied with immature and inaccurate interpretations and construct increasingly maturer and more accurate ones. In Piaget's structuralist view, equilibration—the process of adjusting and perfecting the balance between mental understanding and function in the world—is the central engine of change. For Piaget an individual pursues and takes advantage of experiences to achieve higher levels of understanding. A young child, for example, may at first draw a chimney top parallel to the angled roof of the house rather than parallel to the ground. Such a feature does not reflect the child's own experience but rather a naive assumption, which as a result of growing interactive experience is eventually modified to conform with reality.

Freud's theory is concerned almost exclusively with social and emotional development. He supposed that development consists of the resolution of confrontations between the evolving individual and the conflicting demands of the social world. According to Freud, individuals are biologically endowed with needs that motivate their behaviours and are modified and transformed through their social experiences. Although for Freud the emergence of psychosexual stages is determined primarily by maturation, experiences during different periods in development forever colour the individual's socio-emotional growth.

Developmental contextualists also argue that development proceeds out of interactions between the individual and the world; in addition, they stress interactions within the person (for example, among cognitive, emotional, and motivational processes) and interactions among biological and physical-environmental processes. Thus, this theory is probabilistic (*i.e.,* based on the assumption that certainty is impossible) with regard to what a person can know, as well as what can be known about that person—knowledge is always a function of a particular set of biological, psychological, social, ecological, and historical conditions.

Prevailing opinion among behavioral scientists posits an even more intricate transaction between nature and nurture, or endowment and environment. According to this view, individuals are born into the world with both cognitive and social propensities that help to shape their experiences, and in turn they are shaped by their experiences to the degree that the environment presents certain events and that the individuals themselves are plastic (responsive) to them. In other words, development is a transactional process in which the individual and the environment continuously affect each other through time. Attributes of the individual have meaning for development only by virtue of their interaction with a particular set of time-bound contextual conditions, and, reciprocally, development can only be understood by specifying relations of the context to specific developmental features of the individual. This position has two significant corollaries. One is that individuals help to produce their own development by provoking significant others in their lives, by processing information and events in the world idiosyncratically, and by actively selecting or shaping the context within which they act or interact. A second is that events in one period of life help to shape and to texture what transpires or how what transpires is understood in other periods of life that come before as well as after.

The theoretical orientations described above have implications for key issues in the study of human development, for what develops, and when, how, and why. In the mechanistic view, development is the quantitative accretion of identical elements; in the case of behaviour, for instance, these elements are stimulus–response connections. The number of such connections changes over time, and there is no a priori direction of change. Moreover, a stage in development merely summarizes the organization of such connections; there is nothing qualitatively distinct about the individual at one or another stage. In the organismic tradition, however, qualitative status and change in both structure and function are emphasized, and change constitutes a specified movement toward a predetermined end state. A stage in development denotes an organization of behaviour that is qualitatively different from those of prior or subsequent stages. Developmental contextualism emphasizes the bidirectional character of both qualitative and quantitative change in structure–function relations. The interactionist view admits of stages, born of an intrinsic interplay between nature and nurture.

The concept of developmental stages

TECHNIQUES OF RESEARCH

All scientific research rests on the pursuit of knowledge by observation and experimentation. To study the development of human behaviour, a scientist must decide, first, the means by which observations and experiments should be made and, second, the designs according to which observations and experiments should be scheduled.

Means of observation and experiment. Scientists observe genesis, continuity, and discontinuity in human behaviour in many different ways. Choice of technique is determined by many factors, including an interest in avoiding reactivity (influences a person experiences as a result of being observed). One useful way to study the development of human behaviour is to observe it natu-

ralistically, either by participant observation (in which the researcher actually becomes a part of the study setting for an extended period of time) or by structured observation (adopting a more detached and rigorous stance with regard to the setting). Either way, the observations must be systematic, and the behaviours investigated must be explicitly operationalized (*i.e.*, they must be defined so that they can be replicated independently). Naturalistic observation is reasonably faithful to everyday life and usually provides valuable information about behaviour. Its role in the study of the development of human behaviour has been important; indeed, modern developmental study was initiated by Charles Darwin's observations of his infant son William ("Doddy"). Limitations of naturalistic observation include the fact that it may be inefficient in that the behaviours of interest may occur only at infrequent or irregular intervals, that researchers may be unable to attend simultaneously to a variety of behaviours of interest (even with the aid of supplementary recording apparatus), that the descriptive data yielded by naturalistic observation often do not serve the purpose of explanation, and that uncontrolled as well as unknown factors may influence the behaviours being observed. Consider, for example, a study of the development of aggression in five-, six-, and seven-year-old children. A researcher may choose appropriate samples of children and then observe predetermined categories of aggressive behaviour while the children are at play in a schoolyard. It is possible, however, that after several days the researcher would find that he had too few observations of the behaviours of interest, that the observed field of action was too great in scope, that the activity itself was too furious, or that the observations yielded only descriptive information that aggressive acts occurred with certain frequencies in children at certain ages but did not explain why they occurred. Furthermore, factors peculiar to the situation but not observed by the researcher could selectively influence the expression of behaviours.

Two other approaches to the study of behavioral development are controlled and experimental observation. In controlled observation, the research situation is naturalistic but structured, and participants' behaviours are not directly manipulated. In experimental observation, the researcher exercises maximum control over the observations and performs direct, calculated manipulations of behaviours.

To illustrate controlled observation, suppose that the researcher is interested in the bases on which adolescents form heterosexual relationships. The researcher may place a group of male and female adolescents unacquainted with each other in a classroom and instruct them to form two-person study groups, ostensibly to examine a topic about which all have little background. Observers may then rate the physical attraction between couples. The situation has been controlled, and some characteristics of participants within the situation have been arranged. To illustrate experimental observation, suppose that the researcher is interested in the extent to which each of three types of instructional techniques influences learning. Since variable factors such as sex, age, social class, intelligence, race, religion, and type of school can also influence learning, these factors must be controlled in order to focus the study on the instructional techniques. Thus, the subjects of the study may be limited to 10-year-old, white, middle-class, Protestant boys of average intelligence, attending a public elementary school in the southern United States. The only factor varied would be the type of instruction to which these children are exposed so that only the precise effects of instructional technique on learning are determined. In this type of experimental observation, conditions are manipulated in such a manner that only those factors whose effects are to be ascertained actually vary, and this variation itself is also controlled, yielding particularly clear data on the phenomenon. Unfortunately, information yielded in such a study would be limited to the conditions of the study, and because of this limitation it may be far removed from everyday life. The effects of the instructional techniques on boys or girls of different racial compositions and social backgrounds, or who attend schools in different places, would be unknown. Moreover, the known effects of instructional technique even on the subjects of the study would be limited; in the real world the factors that the researcher would have controlled would vary and would, no doubt, influence his subjects along with the instructional mode.

Observational techniques are useful only when the behaviours in question are observable and are ethically open to scientific study. Alternative techniques may be required in some circumstances: when it is desirable to study large numbers of people over intervals shorter than those permitted by behavioral observation; when there is no overt behaviour to observe (as when feelings, attitudes, values, or recollections are of interest); when the behaviour may not readily or ethically be seen through observation (as in activities that may be harmful or embarrassing); or when the presence of a researcher may distort the behaviour (as in sexual interactions or voting). Questionnaires and interviews can circumvent these difficulties, although subjects' recollections or predictions regarding their behaviour must not be considered a true index of actual behaviour.

Most methods of studying behavioral development involve an element of compromise. For example, naturalistic observations are inefficient but retain real-life validity; experimental observations are more efficient but risk invalidity for the sake of control. Obviously, both approaches contribute to the scientist's goals of objectivity, reliability, efficiency, replicability, and representativeness. In order to increase both validity and accuracy, scientists often employ what are called converging operations, bringing several different research strategies to bear upon the same problem or question.

Designs of observation and experiment. The basic stages of behavioral development proceed according to a biological timetable measured by chronological age; historical context, however, strongly influences certain expressions of development, as many characteristic behaviours and attitudes of each succeeding generation may differ from those of the previous one. In designing investigations into particular aspects and patterns of behavioral development, scientists must take both of these influences into account. "Design" here refers to the selection of a particular array of individuals for observation or study. Of the many standard designs in developmental research, two are most popular. Each has utility, but each also has weaknesses that must be weighed against its strengths. In the longitudinal design, the scientist observes and compares the same individuals over time. This procedure provides the only means of assessing directly continuity and discontinuity in the individual's behaviour across stages of development, of interrelating development in different domains, and of evaluating the effects of early experience on later behaviour. Longitudinal study is expensive and time consuming, however, and many potential subjects are unwilling to participate in research over long periods of time, so that samples studied tend to be small; also, willing participants may not be representative of the population at large. Furthermore, repeated assessment risks altering the behaviour of participants. The most widely used developmental research schedule is the cross-sectional design, in which individuals of different age groups, or birth cohorts, are studied at the same time. Observations of different ages can be completed relatively quickly in this design, without practice effects (behavioral changes caused by the observation itself) or attrition, and this approach is relatively inexpensive. Cross-sectional design, however, does not permit many of the comparisons (*e.g.*, individual growth) allowed by longitudinal evaluation. Also, in a cross-sectional study it is difficult to control all of the variables that may affect behaviour differences among groups, since such differences may reflect real age changes or initial inequality among age groups. Therefore, whether and the extent to which age-group differences correspond to age changes within individuals over time is left open to question.

Longitudinal design

Cross-sectional design

Alone, neither of these two developmental research designs allows for adequate determination of the contributions of the separate factors of chronological age and birth cohort since each involves a confounding of potential components of change. When variables are confounded,

the influence of one behaviour cannot be separated from that of another since the two may be influencing behaviour at the same time. Thus, findings about development gained from longitudinal study may reflect age-related changes, determined endogenously, or exogenously determined changes related to the historical context of birth and development. Likewise, cross-sectional results confound chronological age and birth cohort, so that at any one time individuals of different cohorts may exhibit different behaviours caused by either their different chronological ages or by their having been born during different time periods.

Some limitations of conventional designs are transcended by combining longitudinal and cross-sectional features in sequential designs. In the so-called cross-sequential design a succession of two or more cross-sectional studies is conducted; in the so-called cohort-sequential design two or more longitudinal studies are conducted in succession. Sequential designs assess the relative contributions of age and cohort simultaneously, but they are difficult and often expensive to undertake.

Sequential designs

Thus, conventional research designs do not allow for unconfounded assessment of the contributions of age and cohort, and less conventional ones are difficult to implement. Because of their respective benefits and shortcomings, the various methods yield variously useful depictions of development. No research method is without limitations, and the general view is that all of the conventional designs of research can contribute to understanding the development of behaviour if they are employed with clear recognition of their limitations.

Development across the life cycle

Any attempt to answer central questions as to what develops, when, how, and why requires two assumptions. First, it is necessary to partition the life cycle; sometimes chronological age (a continuous variable) is an appropriate measure, but at other times some category of the life span is more apt (*e.g.*, infancy and childhood). Second, it is necessary to distinguish among the various domains in which development occurs (*e.g.*, sensation and perception). These two decisions are not as straightforward as they appear. For example, there is some reason to suspect that the concept of "adolescence" as a separate stage in the life cycle is historically relatively new, and there is also debate about whether perception differs from cognition. Nevertheless, certain biological events, such as birth or the onset of reproductive maturity at puberty, or social events, such as the acquisition of language or retirement from work, mark off distinctive and often universal phases of the life span. Likewise, development takes shape in separate physical, mental, and social spheres. With these categories in mind, therefore, it is possible to examine highlights of development in three domains—the physical, mental, and social—during five phases of the life cycle—prenatal life (physiological development only, for the purposes of this discussion), infancy and childhood, adolescence, adulthood, and old age. A common view depicts human development in terms of genesis before birth; growth and maturation during infancy, childhood, and adolescence; stability during adulthood, and inevitable degeneration and decline during old age. However convenient and reasonable such a breakdown may appear, closer examination of each stage reveals disparities and continuities that defy simple categorization.

Phases of development

PRENATAL PHYSIOLOGICAL DEVELOPMENT

In order to examine the development of behaviour throughout all phases of the life cycle, it is necessary to begin with the phase before birth, in which behaviour as it is considered in this article does not yet emerge but in which the "constitution" of the individual, to whatever degree such a concept is regarded as operational, may be said to take shape.

Three periods of prenatal development have been distinguished—the zygotic, the embryonic, and the fetal. The period of the zygote, or fertilized egg, lasts from the moment of conception through the time of implantation in the endometrium, or lining of the uterus. The hallmark of this period is cell multiplication. The succeeding period of the embryo lasts for approximately two months, during which time fundamental morphological and structural developments occur. At this stage, three layers of the organism are distinguishable—the ectoderm, presumptive of mature skin, sense organs, and central nervous system structures of brain and spinal cord; the mesoderm, presumptive of mature muscles, blood, and the circulatory system; and the endoderm, presumptive of mature digestive, respiratory, and other internal organs. At two months after conception the embryo is little more than a couple of centimetres long, but it is morphologically humanoid: all of the major organs are differentiated, although they are not fully developed or maturely proportioned. For example, arms and legs are recognizable, as are fingers and toes; the stomach produces digestive juices; the kidneys filter blood; and the heart beats. The following period, that of the fetus, lasts from two months after conception to the end of gestation and is a time of increasing function and change in bodily proportion. In the fetal period, fingers separate and grow nails, veins develop, muscles interconnect; the fetus has been observed to suck, blink, and even yawn. By four months, fetuses are felt to "quicken," or move in utero; by six months, the expansion and contraction of the lungs and the functioning of the nasolacrimal ducts indicate breathing and crying; and by seven months, fetuses may survive in the extrauterine environment.

It was once believed that a "preformed" human organism resided either in the mother's egg, waiting to be released, or in the head of the father's sperm, waiting for a medium in which to grow. It is now known that each parent endows the offspring with one complementary member of each of 23 pairs of chromosomes that form its genetic constitution, or genotype. The genes that make up chromosomes are themselves composed of chemical codes that guide the development of both structure and behaviour. Although genetic endowment is critically important in the life of the individual, the genotype does not necessarily predict the phenotype (the individual's observed characteristics). Different people with different genetic makeups sometimes look or behave similarly, just as "identical" twins, whose genetic makeups are the same, may look or behave differently. Genetic endowment and experience interact even at the earliest stages of the life cycle to shape the structure and behaviour of the individual.

Genotype and phenotype

Because the time before birth is a period of rapid and extensive development, many factors may influence the course and quality of prenatal growth. These factors include parental age (birth complications tend to increase with parental age), diet (protein is advantageous and malnutrition is perilous since nutrients constitute both the building materials of development and the fuel by which development is propelled), and ingestion of drugs, as well as exposure to disease and environmental toxins. The mechanism of action of such effects is not straightforward: At prescribed times, known commonly as sensitive periods, the organism becomes especially vulnerable to exogenous influences that may temporarily or permanently alter structures or behaviours. These sensitive periods usually occur between the time a particular structure or function emerges and the time it reaches its mature state. For example, the eyes and the visual system develop most rapidly during the second month of pregnancy, and maternal contraction of rubella (German measles) at this time presents a nearly 50 percent risk of causing infant cataracts. Developmental timing is so important that different toxins have similar (often serious) effects during the same phase of prenatal life, yet they may not affect development in identifiable ways at other stages. Further, the effect may be disproportionate to the dose, and it may be delayed.

"Sensitive periods"

Structures related to particular functions emerge at various times during prenatal development. Subcortical structures that control the behavioral state, including the hypothalamus and arousal system of the reticular formation, emerge first. Components of the limbic system and basal ganglia, which govern emotion, instinct, and posture, develop next. Finally, the cortex and its cortical association areas, concerned with awareness, attention,

memory, and thought, emerge last. Because of the nervous system's complexities and astonishing ability to regulate and integrate information, its development in prenatal life is especially remarkable. The adult human brain contains approximately 100,000,000,000 neurons; because virtually no new neurons are generated after birth, some 250,000 cells must be generated every minute before birth. In human beings, genetic programming ensures the cell abundance that is required to perceive, think, feel, and act. The individual cell comes into being, matures, develops sophisticated interconnections with other cells, and subsequently specializes in function. In short order individual neurons aggregate morphologically to initiate meaningful human brain functions. This predetermined specialization does not preclude the plasticity of individual cells or of cell masses since their functions may be modified if they are relocated (through transplantation, for example) before a sensitive period occurs; in addition, some cells may assume the functions of other cells that die, exhibiting further exchange of function. Neurological research suggests that the newborn brain is particularly open or plastic in these regards. Although the neonatal or newborn period is commonly considered a time of growth, expansion, and development, another essential neurological function in early life is the contraction, death, and elimination of initial cell overproduction. Overproduction is thought to underpin plasticity early in the life cycle and the restriction of function later.

Plasticity

Brain function is intimately linked to the development of the sensory systems that enable the organism to register, integrate, and interpret its world. The senses do not lie dormant until they are suddenly switched on at birth; evidence suggests that the sensory systems function well before birth. The different senses seem to achieve structural and functional maturity at different times prenatally—in the order of sensitivity to touch, position, sound, and light—reflecting a staggered developmental schedule that loosely follows the order in which these systems developed during the species' evolution.

Development of the senses

Although it is notoriously difficult to establish causal relations between the brain and behaviour, it would be surprising if behavioral development were not reflected in analogous changes in the structure and working of the human brain. Indeed, in large measure, the psychological accomplishments of infancy reflect impressive prenatal developments of the nervous system. In the space of approximately nine months a single fertilized egg develops into a complex, self-regulating organism, and in an additional nine months it develops further into a sentient child capable of some intelligent feelings, thoughts, and actions. In short, development of the central nervous system prepares the fetus for its new role as a child outside the womb.

INFANT AND CHILD DEVELOPMENT

By definition, infancy is the period of life between birth and the acquisition of language, approximately one to two years later. Despite its brevity, this phase of development has attracted a disproportionate amount of attention and interest, perhaps because the physical and behavioral changes that take place during this time are more dramatic than at any other period in the human life span. The average newborn weighs approximately $7\frac{1}{2}$ pounds (3.4 kilograms) and measures approximately 20 inches (51 centimetres); by the end of the first year of postnatal life the average infant weighs approximately 20 pounds, or nearly three times the birth weight, and measures approximately 30 inches in height, or half again the birth length. Change is equally remarkable in every sphere of development: in the shape and capacity of the body and its muscles, in the advancing complexity of the nervous system, in the growth of sensory and perceptual capacity, in the ability to make sense of and to negotiate the physical world, and in the formation of social affinities and the emergence of personal and social styles. At no other point in the life span do such major changes occur in so many aspects of development so quickly; nor at any other time are they so thoroughgoing. The pervasiveness, rapidity, and clarity of developmental change in infancy help to account for the universal fascination with this stage in the life cycle.

Physiological aspects. After approximately 266 days of gestation, an unknown factor causes the maternal pituitary gland to release a hormone (oxytocin) that in turn initiates muscular contractions of the uterus and the subsequent expulsion of the fetus. The duration of this process, called labour, is influenced principally by maternal age and parity (number of children previously born); it is approximately 17 hours on average for firstborn babies. Significant risks during this period to later psychological development include oxygen deprivation (anoxia) and overuse of anesthetics, both of which have been associated with a variety of adverse long-term sequelae, such as motor defects and impaired cognitive performance. At birth, infants are examined immediately to determine the need for medical intervention and to establish normal functioning. For all its drama, however, birth neither terminates nor initiates development; it is simply one stage in an ongoing developmental process.

The birth process

Newborn activity appears to be spontaneous and random; newborn babies move their eyes, mouths, hands, and feet constantly and without apparent purpose, and they seem to shift abruptly and unpredictably among states of sleep and alertness. Long-term observation, however, reveals considerable regularity in the pattern of cycles of many behaviours in the newborn. Involuntary activity is organized at greatly different rhythms. The rhythms of breathing and sucking are rapid, cycling regularly at frequencies of one or more per second; general bodily movements cycle at periodicities of a minute or two; and waking, quiet sleep, and active sleep are low-frequency phenomena that cycle at periods of one or more hours. At any given time, newborn activity reflects the simultaneous but independent cycles of several complex rhythms.

Periodicity of activity in newborns

In the first month, newborns are awake for about a third of the day. State of consciousness is an important consideration in infancy, since most of what infants learn about the world, about people, and about their own abilities can only be acquired during periods of quiet alertness and attentiveness. Although human neonates appear helpless, they are capable of a small number of integrated and organized (if limited) behaviours: These are reflexes, simple and unlearned stimulus–response sequences that are biologically meaningful and suggest an evolutionary and adaptive significance. For example, approach reflexes serve to maintain sustenance and include breathing, sucking, and swallowing. Avoidance reflexes, which ward off danger, include coughing, sneezing, blinking, and muscle withdrawal. The regularity of reflex function in human infants provides a means of assessing normal neurological development, since both the emergence and disappearance of many reflexes by the end of the first year indicate neurological integrity.

Physical and motor development in infancy and early childhood are impressive because they are so evident and because change is so rapid. Moreover, physical and motor development have both practical and theoretical implications for growth in other spheres of psychological functioning. Because physical growth is so easy to observe and quantify, it was one of the earliest subjects of study by scientists. Its study betrays several general principles, including directionality. Whether in terms of anatomy, complexity of function, or voluntary control, growth tends to proceed cephalocaudally, or from the head of the body to the feet, and proximodistally, or from the centre of the body outward. A second principle of growth is the independence of systems. In the first years of development, the nervous system achieves more than half of its adult status, physical characteristics of the body develop to less than a third of their eventual goal, and secondary sexual characteristics develop hardly at all. A third principle of physical growth and development is canalization, the narrowing or restricting of alternatives so that genetically endowed targets tend to be reached in preference to others. A fourth general principle combines the normal (Gaussian) distribution of many physical, biological, and psychological characteristics with the proviso of individual differences. The normal distribution shows species-typical characteristics. For example, very few adults are four or seven feet (1.2 or 2.1 metres) tall; many more are five and

Principles of physical growth

six feet tall, and most fall in between. These distributions of structure and function are important since they define norms as well as probable ranges for individual differences. For many normal developmental achievements, the age of acquisition has enormous import, but the true range of individual differences is extraordinary especially when considered in proportion to the child's age. For example, some children first walk at 10 months, others at 18 months; some children say their first word at nine months, others at 29 months. Developmental variation in physical growth, as in other spheres of life, may arise in many different ways. Parents particularly tend to be interested in knowing how individual children develop characteristic capacities; scientists, too, find questions concerning the origins and growth of individual differences especially interesting, but they tend more often to focus on general developmental capacities and trends. It is important to keep in mind that, statistically, the central tendency of a distribution does not necessarily best represent that distribution, nor is it necessarily optimal for members of the distribution. Moreover, general developmental trends need not necessarily represent how all or even any individuals develop.

Infant motor development

Like physical growth, motor development is dramatic in the first years of life. From the newborn, unable even to grasp an object or to roll over from the position in which originally placed, emerges the toddler, who is manually deft and able to locomote anywhere at will. Psychomotor development also suggests a program of maturation, although research shows that experiences can channel its course. Depending on the behaviour involved, maturation and experience vary in relative importance. For example, though experience facilitates the coordination and strengthening of muscles generally, experience seems to play a less significant role in the development of gross skills such as walking (in which limits are set largely by the rate of physical maturation) than in the development of finer skills such as skiing (in which earlier and better training seems to make a great difference).

Perception in infants

Although hardly mature, the senses function well at birth. Perception constitutes a necessary first step in experiencing and interpreting what the world has to offer, and for this reason philosophers and scientists have been attracted to the study of perception at or near the beginning of life. Determining how the senses function in infancy also offers a glimpse into the infant's perceptual world and suggests what aspects of the environment can influence that infant's development. No matter how early in life a perceptual capacity manifests itself, it can only manifest itself in terms of actual experience; potential is not observable until it is brought into use. Conversely, no matter how late in an individual's life an ability emerges, its emergence can never be attributed exclusively to experience. Research into the origins of perception has focused on the related goals of, first, determining whether or not particular perceptual capacities are present at the beginning of life and, second, tracing continuity or discontinuity in those capacities as they develop. Perceptual ability is a natural outgrowth of biological sensory function; however, experiences are also critical for normal perceptual growth and development. An experiment conducted with institutionalized newborns, for example, showed that introducing a visually interesting stabile, or fixed sculptural form, into the infants' otherwise bland environment at one month after birth nearly doubled their visual attentiveness and visually directed reaching. Studies of naturalistic mother–infant interaction also suggest that experience influences infants' competence in visual and tactual exploration. Furthermore, it is known that in the development of taste, the greater the variety of infantile experiences, the more open to new tastes the child will be.

Research shows the achievement of extraordinary perceptual sophistication over the first months and years of life. In vision, for instance, newborn babies attend selectively to parts of patterns in which there is information. During the first half year of life there is rapid development in acuity, from 20/800 vision (in Snellen notation) among two-week-olds to 20/70 vision in 5½-month-olds to 20/20 vision at five years. Research shows, too, that well before the end of the first year infants perceive form as form (for example, they perceive an object's shape to remain stable even through transformations of the object's image on the retina), perceive depth in space (showing sensitivity to a "visual cliff" and to an object looming at them), distinguish orientation (smiling at faces on the vertical axis more often than at upside-down or horizontally rotated ones), locate objects in space (reaching in ways that indicate good prediction of location), and display sensitivity to object movement (discriminating among different kinds of motion). Before they are six months of age, infants give evidence that they perceive the colour spectrum in a qualitative, relatively mature fashion, partitioned into categories of hue. In audition, newborns give evidence that they hear, and young infants have been found to display sensitivity to basic dimensions of sound as well as to elemental qualities of human speech. Indeed, young infants are especially attracted to speech as meaningful communication rather than mere noise or a composition of different sounds. Knowledge of the development of taste, smell, and touch in early life is more rudimentary than that of seeing and hearing; nevertheless, by their facial expressions when sweet, sour, and bitter substances are placed on their tongues, neonates give evidence that they discriminate among common tastes and rate them hedonically, or according to their relative pleasurableness. Moreover, by evidence of their facial expressions and attempts to approach or withdraw from diverse odours introduced by means of cotton swabs held beneath the nose, neonates give evidence that they discriminate among common odours and rate them hedonically. Indeed, breast-feeding babies only days old recognize maternal scents. Likewise, infants explore tactually with increasing efficiency. With age, mouthing of objects decreases while fingering and manipulation increase. Exploratory activities rapidly develop to match the characteristics of the object explored; *e.g.,* infants have been observed to respond to a change in an object's shape by rotating the object and to a change in an object's texture by increased fingering.

Integrated perception

Just as an adult at the ocean sees the waves, hears the surf, smells the salt air, and experiences these sensations as an integrated percept, so over the first year does the infant experience the world by means of senses working in concert. For example, manipulation allows tactual exploration and enhances visual investigation. Studies show that infants are extremely sensitive to multimodal information even when it stimulates different senses in distinctly different ways; thus, they can associate a particular face with a particular voice. In uncovering these facts about early perceptual life, investigators have eradicated the once staunchly held notion that infants and young children are perceptually deficient.

Cognition and language. Cognitive development, which takes place against this backdrop of remarkable neurological, physical, and perceptual change, entails the child's growing ability to make sense of the physical and the social environment, that is, to recognize specific objects and persons, to learn the laws that govern their behaviour, and to participate meaningfully in communicative dialogue. What learning, thinking, and language are and how they develop over the first years of life constitute central concerns of scientists interested in the development of human behaviour.

The principle of association

According to the principle of association, two events that occur in close proximity in space or in time tend to be remembered together so that the recurrence of one will bring the other to mind. Because it is a basic mechanism of learning, this principle has long been thought to subserve cognitive and emotional growth. Several types of learned association have been identified; among them are classical conditioning (in which a neutral stimulus assumes the properties of an evocative stimulus through association with it), instrumental conditioning (in which actions that are followed by reward tend to recur and actions followed by punishment tend to be inhibited), and exposure learning (as in the progressive decline in responding to an event that is available continuously or repeatedly). These types of learning are thought to involve the construction of a mental representation of an event

and ongoing comparison of available stimulation with that mental representation. Research clearly demonstrates the very early functioning of these different modes of learning. Newborns can associate a particular orientation of the mother's face with the sweet-tasting substance that is consistently paired with it; they alter their sucking patterns for the reward of mother's voice; and they pay diminishing attention to repeated stimuli. There is even evidence that these modes of learning operate among fetuses in utero.

All of these phenomena indicate that infants from the start of life can avail themselves of many experiences. The fact that human beings can learn so early in life, however, does not preclude development of the learning capacity. Indeed, even in the first years of life, growth in the ability to learn is impressive; for example, the rate of exposure learning shows steady and dramatic change even during the first postnatal year, perhaps reflecting the ongoing physiological maturation of the central nervous system. Factors other than age also influence early learning. These factors include the behavioral state (usually the attentiveness of the infant or child), the elements to be associated (it is by no means self-evident that classical conditioning can be established between any two events or that any behaviour can be strengthened through instrumental learning; rather, human beings are evolutionally "prepared" to learn some associations with greater facility), and the complexity of the learning task. When learning begins, more than associations are learned; knowledge is also acquired about the task of learning. It is widely believed that learning in infants leads to the development of effectance (self-confidence in the ability to control experience), to the motivation to learn more, and thereby to even greater discovery and mastery.

The development of memory

Learning requires memory; the effects of learning are minimal in the absence of any storage and aggregation of knowledge. In considering the growth of memory, scientists distinguish between two pairs of concepts: First, there is the storage, or encoding, of incoming information as opposed to the retrieval of that information later. Second, there is recognition, in which the memory of an experience is cued by repeating the experience, as opposed to recall, in which the memory is retrieved without any cues. Research demonstrates reasonable capacity for recognition even among newborns, as well as rapidly improved encoding and retrieval over the first years of life. For example, infants familiar with one face will later look at a novel face in preference to the familiar one, and toddlers observing the hiding of a toy will later successfully search for it. Again, age combines with conditions of learning to affect the robustness of memory. For example, training conducted over several intervals (distributed practice) typically results in better memory than training conducted all at once (massed practice), and evaluation of memory after a short delay shows more positive results than does evaluation after a longer delay. In general, encoding skills and short-term memory seem to improve rapidly over the first year, and retrieval skills and long-term memory seem to improve gradually over the entire course of childhood.

Two approaches dominate psychological evaluation of the development of thought. One tradition is rooted in the psychometric testing of intelligence, and the other is founded on the Piagetian theory of qualitative changes in mental structure. The two views converge in the beliefs that intelligence implies successful adaptation to the environment and that genetic inheritance and experience interact in contributing to the development of intelligence. Just as it would be disadvantageous for nature to fix an organism's intelligence in advance of its experiences in the world, it would be highly counterproductive to leave the development of intelligence wholly to experience, since individuals' experiences can vary so widely. Although intelligence may manifest itself in many ways, psychological tradition has fixed on verbal and logical skills to evaluate individual achievement.

The mental testing movement was founded at the beginning of the 20th century to study differences in adult perceptual functioning and to identify incipient mental deficiency in childhood. Two compelling questions related to the early development of intelligence soon arose—namely, how early in life intelligence might manifest itself and whether an individual's intelligence early in life was related in a meaningful way to that individual's intelligence in maturity. In general, traditional infant tests applied in the first years of life have shown poor reliability (consistency of test scores over time) as well as poor validity (relation of test scores to independent and objective criteria). Several rationales have been provided by way of explanation: Some have argued, for example, that early development proceeds by fits and starts so that unreliability and unpredictability are inherent to this stage of life; some have argued that early life is a period of such rapid growth that test items appropriate for one age may be inappropriate for another; and some have argued that early life is a period of such extreme susceptibility to experiences effecting rapid change that children affected by different experiences may be inappropriately tested with the same items. By contrast with results of traditional infant tests, new measures of attention in early life that attempt to assess the child's efficiency in acquiring information from the environment exhibit reasonable reliability and validity. Individual differences among infants and young children of given ages, as well as different levels of competence across ages for comparable cognitive tasks, have long been attested, but early psychometric tests repeatedly showed that only after four to five years of age did children's scores begin to stabilize in relation to their performance in maturity. Recent work suggests a degree of predictability in cognitive growth from an earlier period in the life cycle.

Problems of testing infant intelligence

Contemporary research supports the theory of an endowed intelligence that is open to postnatal experience. Studies of both the origins and the maintenance of intelligence from infancy into adulthood show the roles of both endogenous and exogenous factors. On the one hand, identical twins, who share 100 percent of their heredity and most of their environment, are more alike in the pattern and development of their intelligence than are fraternal twins, who also share a good deal of their environment but only half their heredity on the average; likewise, parent and child intelligence test scores are strongly associated. On the other hand, parental intelligence is thought to influence parental behaviour toward children, which in turn is presumed to influence child mental performance. So, for example, institutionalized children and children from lower social classes tend not to score as highly on standardized intelligence tests as do children from intact families or upper social strata, owing presumably to the disparity in formative experiences. The demonstrable success of cognitive interventions with deprived children supports these findings and goes far to discount alternative explanations that the apparent influence of family structure and social class on test scores reflects inherent differences in cognitive ability among the comparison groups. Indeed, research has progressed significantly toward pinpointing the types of experience that affect cognitive development, including the amount, variety, appropriateness, and contingency of stimulation, as well as the interactive quality between child and adult. Research shows, too, how the efficacy of these different types of experiences varies across infancy and childhood. Recognizing the role that experience plays in the development of intelligence is especially important, since, in everyday life, infants and young children normally provoke adults to act in different ways with them, and individual as well as age differences clearly affect the types of learning experiences to which children are exposed.

Complexities of family and social influence

There can never be empirical certainty about which factors contribute to the development of cognition. Environmental and genetic bases for characteristics such as intelligence in children reared by their biological parents cannot be separated because these two sets of forces are ineluctably intertwined.

Based on extensive observations and informal tests of his own three children early in the 20th century, Jean Piaget proposed what he called a genetic epistemology, according to which infants and young children constantly reconstruct their own means of knowing in ways that join their evolutionary history with their particular experiences in the world. In doing so, Piaget rejected both the

Piaget's "genetic epistemology"

nativism and the empiricism then prevailing and introduced a new set of concepts. In contrast to nativists, for whom the basic elements of knowledge are innate ideas, and to empiricists, for whom the basic element is sensation, Piaget introduced the notion of schemes, roughly defined as the various faculties of the central nervous system (including sight, hearing, symbolic thought, and language) by which information about the environment is processed. In Piaget's view, intelligence is a function of adaptation employing the complementary techniques of extracting meaning from data by organizing them according to existing schemes (assimilation) and modifying existing schemes in order to assimilate new information more efficiently (accommodation). Assimilation, for example, predominates during pretend play, in which reality is interpreted as the child wishes, and accommodation predominates during imitation, in which the child mimics new realities as closely as possible.

In Piaget's view, human beings throughout their lives actively adapt their own mental structures, which unfold according to a genetic timetable, in order to maintain equilibrium with the perceived environment. Two examples are the decline in childhood egocentrism and the rise of the concept of objects as independent entities; both are based on the child's transactions with the physical and social environment. Piaget proposed that the development of understanding in the child proceeds through a series of four universal stages, fixed biologically and manifesting distinct dimensions of mental activity: (1) the sensorimotor stage (first two years of life), in which the child experiences his physical self, through reflexes and external stimuli, as an entity in an environment; (2) the preoperational stage (age two to six or seven), in which the child learns to manipulate the environment by means of symbolic thought and language; (3) the concrete operational stage (age seven to 11 or 12), in which the beginnings of logic appear in the form of classifications of ideas and an understanding of time and number; (4) the formal operational stage (age 12 to adulthood), characterized by the manipulation of abstract concepts and the mastery of reasoning. Piaget's theory has been influential in educational policy by suggesting that there are constraints on how significantly the environment can influence cognitive growth if a child has not reached the appropriate stage to assimilate certain experiences. It has influenced linguistic theories by suggesting that aspects of both language acquisition and social attachment attend the cognitive apprehension of the permanence of objects and the development of mental representation. It has also provoked deeper inquiry into processes related to cognition—such as children's play, the development of which has been shown to be orderly, sequential, based on action, and increasingly oriented toward the environment.

Piaget's influence

Piaget's work has an important extension into the study of moral growth in childhood. Piaget hypothesized that all people pass through two phases of moral reasoning. In the first phase an objective and concrete morality is based on the constraints imposed by the powerful (*e.g.,* adults) on the nonpowerful (*e.g.,* children). In the second phase the individual develops a subjective morality based on an abstract understanding of the contracts implicit in cooperative and autonomous relationships. Piaget's stress on the universal ordering of moral growth represents an approach that is quite distinct from relativistic, response-centred approaches. As such, it has stimulated considerable interest among developmental researchers. In particular, the American psychologist Lawrence Kohlberg proposed several stages of development in order to account for the many observable qualitative changes in a person's moral reasoning. Kohlberg's theory of moral development, like Piaget's, is based on the idea that to focus only on the response in a moral situation is to ignore important differences in people's moral reasoning at different points in their lives. These differences in reasoning may, in fact, give different meanings to identical responses evoked at various developmental stages.

In order to investigate the reasoning underlying moral responses, Kohlberg devised a moral development interview based on a series of stories, each presenting imaginary moral dilemmas. On the basis of reasons people give in answering questions about dilemmas in the interview, Kohlberg classified them into three stages of moral reasoning. At the first stage, preconventional moral reasoning, a child uses external and physical events and objects (such as pleasure or pain) as the source for decisions about moral rightness or wrongness. At the second stage, conventional moral reasoning, the child acts as others expect, in accordance with the established order of society. At the last stage, postconventional moral reasoning, moral judgments are made according to the view that there are arbitrary, subjective elements in social rules and that rules and institutions of society are not absolute but relative. Such postconventional reasoning is related to formal operational thinking and thus begins with adolescence (see below *Adolescent development: Cognition*). Kohlberg's stages of moral reasoning follow the social perspective of the person, which moves toward increasingly greater scope (*i.e.,* including more people and their institutions) and greater abstraction (*i.e.,* from reasoning about physical events such as pain or pleasure to reasoning about values, rights, and implicit contracts). Transition from one stage to another is characterized by gradual shifts in the most frequent type of reasoning; thus, at any given point in life a person may function at more than one stage at the same time. Moreover, different people pass through the stages of moral reasoning at different rates. Finally, different people are likely to reach different levels of moral thinking in their lives, raising the possibility that some people may never reach the third, most abstract, stage.

Kohlberg's stages of moral reasoning

Language occupies the central place among early accomplishments in perceptual, cognitive, and social development. In their fourth month, infants indiscriminately respond to social communication; by their 24th month, toddlers comprehend the meaning of prepositions. It is significant that the English words *infant* and *baby* both have their origins in language-related concepts, the former deriving from Latin words meaning "nonspeaker," and the latter sharing a Middle English root with "babble." Many people believe that children enter personhood only when they begin to communicate with others through language.

Language, which has a complex and intricate structure, poses a formidable task for those who try to acquire it, whether they are mature adults or immature children; yet, paradoxically, babies meet the challenge of language learning seemingly with great facility. Language has complex productive (speech utterance) and receptive (comprehension) aspects and operates simultaneously on at least three planes. First, sounds that are to be linguistically meaningful must be produced and perceived as separate from noise. Second, meaning must be grasped in order to say and to understand how symbols are related to the real world as well as to imaginary referents. Third, to ensure meaningful communication, rules must be mastered for encoding and decoding word order.

The mystery of language acquisition has intrigued people throughout history. St. Augustine, for example, wrote that children learn language by imitating, and Constantine the Great supposed that infants could not articulate words because their teeth were not yet firmly rooted. King James I of England, who sponsored a translation of the Bible, sought to identify the original language of Adam and Eve. Toward this end he conceived of a unique experiment, proposing to place two infants on an otherwise uninhabited island in the care of a deaf-mute nurse; the King reasoned that if the two children spontaneously developed speech, theirs would be man's "natural language."

Theories of language acquisition

Some psycholinguists have argued that language learning proceeds on the basis of the child's experiences, whereas others have asserted that language acquisition could only develop on a foundation of biological endowment. This disagreement between empiricist associationism and nativism in so fathomless a domain as language learning has been called "the debate between the impossible and the miraculous." Research shows, however, that infants are surprisingly well prepared in both motivation and competencies to establish communication.

In order to discover how language might be acquired, psycholinguists have listened to and questioned children

as well as consulted their parents as reporters. Observation is the method that dominates this field of study, but the technique is limited, as shown by the exasperation parents often feel at their young child's failure to say with a stranger what the child readily says at home. Standardized assessment tests for the first few years are not yet well developed, and parents themselves (whose diaries have provided much of the earliest detailed information about language development) are often subject to bias.

Dimensions of language

Despite these methodological difficulties, there has been considerable progress in understanding language development. In sound perception, research reveals that, either natively or on the basis of little experience, very young infants divide the speech stream, map correspondences between the eye and the ear, and recognize particular speech. In vocalization, newborn infants sound cries that are distinguishable as signals of pain or hunger; later, infants follow a nearly universal course from babbling through single-word utterances to grammatical speech. Children's earliest sensitivities to sound and their earliest vocal expressions give evidence of a strong biological influence. Yet, even in the first year both the production and the perception of speech sounds are increasingly shaped by the linguistic environment, reflecting the child's exquisite sensitivity and susceptibility to specific experiences.

The acquisition of grammar has sparked some of the longest standing debates in developmental inquiry. Some researchers propose that children learn the rules of language through reinforcement or through imitation; others maintain that, given the enormous requirements of language learning, such a view is hopelessly simplistic and hardly practicable, and they theorize that human beings have innate dispositions toward basic language structure. Again, evidence supports both views. Children everywhere seem to move through a universal sequence of syntactic acquisition, lending support to a biological theory, although children learn the grammar of whatever their local language community happens to be. It is often hypothesized that the human brain has specially developed to meet linguistic requirements, but it is equally clear that adults engage with infants in motor and nonverbal activities that mirror language in structure and that pave the way for the child's comprehending and producing grammatically correct utterances.

Semantics and reference—properly mapping sounds to their meaning—constitutes another aspect of language study. Although the question of reference is not so much plagued by the nature–nurture debate—for the acquisition of meaning cannot readily be attributed to biology—its study is affected by disputes that range from defining exactly what is meant by a "word" to pinpointing the process of matching referents. Whatever the view may be on these issues, it is known that by three years of age children are learning at least two new words per day in an odyssey that has brought them from infancy to a working vocabulary of 3,000 words.

In language, as in other spheres of behavioral development, quantitative as well as qualitative individual differences abound. For example, some 13-month-olds comprehend only a dozen or so words, whereas others comprehend as many as 100; some do not speak at all, whereas others have a productive vocabulary of nearly 50 words. Further, some babies possess vocabularies organized mainly around names of objects, whereas others tend to converse in social formulas principally for the purpose of communicating their feelings and desires. Studies show that heredity and experience both contribute to these striking individual differences. For example, infants' competence in communicating tends to be associated with the competence of their biological (as opposed to their adoptive) parents, yet certain parental activities with infants are also known to assist the development of communication.

The development of language follows a clear and probably universal ontogeny that may, at least initially, be affected by variations in experience. Within limits, normal children seem to follow the same path to basic linguistic proficiency and to travel at more or less the same rate, largely (though not wholly) independent of their general intelligence, the language community in which they are reared, and the amount of early help they receive. Yet nearly all specific manifestations of language learning seem to be prompted and transformed by the child's specific experiences. Critical factors in language development appear to be the amount and variety of verbal stimulation as well as adult sensitivity to infant language levels, especially in dynamic interplay with infants' feedback. Toddlers obviously learn only the particular language to which they are exposed, and differences in experience result in corresponding differences in language competency. To ease this task, parents and teachers readily adjust the language they typically direct to infants. Thus, the processes involved in language acquisition appear to be both rigid and flexible. They assure certain outcomes even in the face of an unstable environment, and at the same time they encourage flexibility in response to a variegated environment.

The foregoing discussion analyzes language acquisition in the child by separating language into sound, syntax, and symbol both in perception and in production. Although each plane entails a separate kind of learning, these components are not independent; rather, they mesh in astonishing complexity to produce what is arguably an entirely new organism.

Personality and social development. Assessing and interpreting the earliest expressions of personality and sociability in human beings have long challenged those who study the development of behaviour. Personality can be thought to include the emotions, their characteristic pattern of expression in individual temperament, and the ways in which experiences and development in growing up enlarge and modify basic temperament and emotions. There is no generally accepted definition even of basic emotions, such as joy, anger, fear, and affection; nonetheless the emotions are recognized as playing a significant role in regulating internal psychological processes as well as in influencing interpersonal interaction—primarily through vocal, facial, or gestural expression. So, for example, the experience of joy not only energizes an individual, but its expression also encourages others around the person to remain near and to keep up a pleasurable interaction. Watching face and gesture and listening to voice, some theorists have argued that the earliest expressions of emotion are learned through experience, presumably through parent–infant interactions. By contrast, others have suggested that emotions are not necessarily outcomes or products of experiences or thought, that emotions are expressed and comprehended universally, and that emotions possess survival value as important means of communication with others of the species; on these bases, these theorists have concluded that basic emotions reflect innate capacities for which no social experience is required.

Earliest expressions of emotion

Developmental studies show that infants both express and interpret many distinct and presumably core emotions in the first years of life, contrary to previous views that they can exhibit only undifferentiated general excitement. Quite frequently in their everyday interactions, for example, newborns clearly respond to pleasant stimuli with facial expressions that may reasonably be interpreted as positive and to unpleasant stimuli with negative ones; this is not to imply that experience plays no role in shaping the expression of emotions, since different cultures obviously socialize different emotional displays in infants and young children. The very young are also recognized to discriminate among expressions of emotion conveyed in face, voice, and gesture. Although the ability to discriminate does not necessarily imply understanding of emotional meaning, research shows that children in the first and second years of life use the emotional expressions of others to regulate their own behaviour, especially with regard to decision making. So, for example, young children are more likely to approach and explore a toy when an adult they "consult" signals joy, as opposed to disgust or fear, and they will even physically position themselves so as to be able to read that adult's expressed emotions more clearly.

The concept of temperament

Most emotions remain available throughout life, but individuals differ in the characteristic patterns of emotions they display, that is, in their temperament. Temperament has many definitions; commonly, it is conceived of as

a constitutionally based source of characteristic socio-emotional elements in the individual, such as affect, attention, and motor activity. Although individual patterns of these components of personality are often thought of as enduring and valid in different situations, they, too, are certainly subject to modification according to the circumstances of life. Presumably personality constitutes the elaboration and transformation of basic temperamental dispositions in development through experience, social interaction, and cognition. Its constitutional base does not mean that temperament is genetically fixed, and models of temperamental variation acknowledge reciprocal influences of environment and organism acting through time so that aspects of temperament are identifiable early in life but may alter with age or environmental change. A convergence of methodologies that combines observing infants and children with interviewing parents has shown moderate agreement in the evaluation of infant and childhood temperament. Research on newborn variation (especially in twins) has strongly supported the appropriateness of a behaviour-genetic model of the origins of temperament. The study of childhood temperament is particularly attractive to clinicians since, if temperament is constitutional and enduring, understanding infant temperament promises to provide clues to understanding mature personality and perhaps to solving parts of the puzzle of pathology in personality. Moreover, individual differences in infant and childhood temperament can be expected to provoke different styles of caretaking and to shape, for good or ill, the nature of children's social interactions.

Infancy and childhood witness the dawning of social awareness, when children come to understand that their behaviour can affect the behaviour of others in consistent and predictable ways (intentionality and effectance) and that others can be counted on to respond when signaled (trust). These basic social characteristics are often born of turn taking (reciprocity) in interpersonal interaction. In their encounters first with parents and later with significant others, infants and young children appear to interact adaptively. Perhaps the central accomplishment in personality development in the first years of life is the establishment of specific and enduring emotional bonds, or attachment. The objects of attachment seem to be the persons who respond most consistently, predictably, and appropriately to the baby's signals, primarily the mother but also the father and eventually others. Interestingly, mothers and fathers have been observed to behave differently with their infants and young children: Mothers hold, comfort, and calm in predictable and rhythmic ways, whereas fathers play and excite in unpredictable and less rhythmic ways. Children quickly learn to expect different patterns of interaction from these two figures.

Attachment

Research suggests that varied social stimulation facilitates the growth of interactional skills in infants and young children. The development of social styles seems to depend on the nature of early interactions, on temperament, and on adult behavioral patterns; moreover, transactions among these factors have ramifications for children's social and cognitive development, including their propensity to explore the environment, the stability of their gender identity, and the facility with which they interact socially with peers. Toddlers who are more secure than others in their attachment to parents approach cognitive challenges such as problem solving more enthusiastically and persistently, and they behave in more socially competent and independent ways in preschool.

Perhaps the preeminent theorist of social and personality development is Sigmund Freud, who, in the first decades of the 20th century, established the modern practice of psychoanalysis. Freud saw a parallel between the transformation of energy in the physical world and events that occur in mental life. Human life, he hypothesized, is governed by a mental energy that he called libido. Libido cannot be created or destroyed; humans are born with a finite amount of it, and instead of its being transformed into another type of energy, its area of localization within the body successively changes over the course of development. Freud postulated a universal, biological progression in which libido is centred in particular body zones for determined lengths of time. Each stage in this progression is, in Freud's terms, a stage of psychosexual development. Where the libido is concentrated determines which stimulation is required to gratify it. Experience can facilitate or hinder the development of certain feeling states, but it is nature that determines where and when libido moves. Freud specified five stages of psychosexual development.

Freud's stages of development

The first psychosexual stage is the oral stage. During the first year of life, gratifying stimulation centres on the mouth region, where libido is concentrated, and pleasure is obtained by sucking and biting. If the baby's attempts at oral gratification are frustrated often enough, some of the libido may become fixated (arrested) at this stage of development; that is, not all the libido will be free to progress to the next stage. Freud believed that adult emotional problems result from fixation, and in the case of oral fixation such problems involve attempts to obtain the gratification missed at the oral stage. As an adult the individual may, for example, overeat. Fixation of libido is a potential problem at any stage.

The anal stage lasts from about age one to age three and occurs when libido centres in the rectal area. Children at this stage obtain pleasure through exercising their sphincter muscles, expelling or withholding bowel movements. The anal stage corresponds to the period when children in many Western societies are toilet trained, and frustration resulting from severe toilet training can result in anal fixation, in which adults are either "loose" (messy and wasteful) or "uptight" (holding back everything, including their feelings).

In the phallic stage, which according to Freud spans years three to five, the libido in both sexes moves to the genital area. For boys, gratification is obtained through manipulation and stimulation of the genitals. Mothers are most likely to provide such stimulation, and boys therefore conceive an incestuous sexual desire for their mothers. Boys recognize that their fathers stand in the way of such desires, and this arouses negative feelings toward them. Freud labeled this emotional reaction the Oedipus complex, after the legendary Greek king who mistakenly killed his father and married his mother. According to Freud's theory, when boys realize that their fathers are rivals, they come to fear that their fathers will punish them by castration. As a result, boys experience castration anxiety. The power of this anxiety forces boys to give up their desires for their mothers and in turn identify with their fathers; that is, boys come to model themselves after their fathers. Although presumably mothers also provide most genital stimulation for girls, girls (for reasons not perfectly clear even to Freud himself) fall in love not with their mothers but with their fathers. Then, analogous to what occurs with boys, girls desire to possess their fathers incestuously but realize that their mothers stand in the way. At this point the similarity with male development ends. Girls are afraid that their mothers will punish them for their incestuous desires, but, although it is possible that girls first fear punishment in the form of castration, their awareness of feminine genital structure causes them to realize that, in a sense, they already have been punished. That is, girls perceive that they do not have a penis but only an inferior (to Freud) clitoris. Hence, girls experience what Freud called penis envy. Penis envy impels girls to resolve the Oedipal conflict by relinquishing their incestuous love for their fathers and identifying with their mothers.

According to Freudian theory, the form of adolescent and adult development is determined in early childhood. The first five years, comprising the first three psychosexual stages, are especially critical for functioning in later life. After the end of the phallic stage (at about five years of age) and until puberty begins (at about 12 years of age), Freud said that the libido is latent. It is not localized in any bodily zone; thus, during the latent stage no zones of gratification emerge or exist. At puberty the libido reemerges, again in the genital area, but this time in a mature adult form of the phallic stage. If the person's psychosexual development has not been too severely restricted during the first five years of life, adult sexuality can now be directed to heterosexual union and reproduction.

Although Freud made great contributions to psycholog-

ical theory—particularly in his concept of unconscious urges—later critics pointed out that actual Freudian research is practically impossible. Such elegant concepts as the theory of psychosexual stages cannot be verified through scientific experimentation and empirical observation. But Freud's concentration on development in early childhood influenced even those schools of thought that rejected his theories.

The study of postnatal behavioral development prior to adolescence has tended to lay stronger emphasis on infancy than on childhood, not because later development is unimportant but because it is widely held by different schools of thought that the most significant developments in behaviour take place early in the life cycle. Psychoanalysis argues for the enduring quality of early experiences; learning theory argues that early experiences are prominent because they are primary, have no competing propensities, and are thus easy to establish and long lasting; ethology proposes that sensitive periods of susceptibility to environmental influence predominate in infant life; and Piaget's theory of genetic epistemology also suggests that advanced cognitive and social capacities build on simple developments that take place very early in life. Infancy and early childhood are thus premiere phases of the life cycle whose characteristics—given, developing, or acquired—are generally fundamental to later experiences that build on or modify them.

ADOLESCENT DEVELOPMENT

Adolescence may be defined as that period within the life span when most of a person's characteristics are changing from what is typically considered childlike to what is typically considered adultlike. Changes in the body are the most readily observed; but other, less definitive attributes such as thoughts, behaviour, and social relations also change radically during this period. The rate of such changes varies with the individual as well as with the particular characteristic.

Physiological aspects. The physical and physiological changes of adolescence do not proceed uniformly; however, a general sequence for these changes applies to most people. It is useful to speak of phases of bodily changes in adolescence in order to draw important distinctions among various degrees and types of change. Bodily changes affect height, weight, fat and muscle distribution, glandular secretions, and sexual characteristics. When some of these changes have begun, but most are yet to occur, the person is said to be in the prepubescent phase. When most of those bodily changes that will eventually take place have been initiated, the person is in the pubescent phase. Finally, when most of those bodily changes have already occurred the person is in the postpubescent phase; this period ends when all bodily changes associated with adolescence are completed.

Phases of adolescence

The bodily changes of adolescence relate to both primary and secondary sexual characteristics. Primary sexual characteristics are present at birth and comprise the external and internal genitalia (*e.g.,* the penis and testes in males and the vagina and ovaries in females). Secondary sexual characteristics are those that emerge during the prepubescent through postpubescent phases (*e.g.,* breasts in females and pigmented facial hair in males).

Several important bodily changes occur specifically within each of the three periods that characterize adolescent physical maturation. The period of prepubescence begins with the first indication of sexual maturation. It ends with the initial appearance of pubic hair. In males, there is a continuing enlargement of the testicles, an enlargement and reddening of the scrotal sac, and an increase in the length and circumference of the penis. These changes all involve primary sexual characteristics. Insofar as secondary sexual characteristics are concerned, there is no true pubic hair at this stage, although down may be present. In females, prepubescent changes typically begin an average of two years earlier than with males. The first phenomena of female development in this period are the enlargement of the ovaries and the ripening of the ova. In contrast with those of males, these changes in primary sexual characteristics are not outwardly observable. However, changes involving secondary sex characteristics can be seen (*e.g.,* the rounding of the hips and the first phase of breast development). The latter begins with an elevation of the areola surrounding the nipple, which produces a small cone-like growth called the breast bud. As with the male, there is no true pubic hair, although down may be present.

The onset of pubescence in both sexes occurs with the appearance of pubic hair, and this period ends when pubic hair development is complete. The peak velocity (speed) of growth in height and weight also occurs during this phase. This so-called growth spurt occurs about two years earlier in females than in males. Another key change of pubescence in females is menarche, or the onset of menstruation, which occurs about 18 months after the maximum height increase of the growth spurt and typically is not accompanied initially by ovulation. In pubescence the primary sexual characteristics continue the development initiated in prepubescence. In females the vulva and clitoris enlarge; in males the testes continue to enlarge, the scrotum grows and becomes pigmented, and the penis becomes longer and increases in circumference. In regard to secondary sexual characteristics, in females there is increased breast development, with the breast buds enlarging to form the primary breast; in males, the voice deepens and pigmented axillary and facial hair appear, usually about two years after the emergence of pubic hair.

The "growth spurt"

The phase of postpubescence starts when pubic hair growth is complete, there is a deceleration of growth in height, changes in the primary and secondary sexual characteristics are essentially complete, and the person is fertile. Some changes in primary and secondary sexual characteristics occur in this phase. For instance, in males, it is during this period that the beard begins to grow; in females, there may be further breast development.

Although, as noted, the ordering of these bodily changes is fairly uniform among individuals, there is considerable variation in the rate of change. Some adolescents mature more rapidly and others more slowly than most of their peers. Of course, there are also youths who pass through the periods of bodily change at the average rate. Variations in the rate of bodily change in adolescence often affect psychological and social development. Early-maturing adolescent boys are typically better adjusted than late maturers and have more favourable interactions with peers and adults. These advantages of early maturation and disadvantages of late maturation tend to continue through the middle adult years for males. For females, however, early maturation is associated with more psychosocial disadvantages than is late maturation. Maturing at an average rate seems to be most advantageous for females. However, the relations between female maturation rates and personality and social functioning in later life have not been determined.

Variations in rate of change

Bodily changes among adolescents can also differ according to sociocultural and historical influences. The age of menarche, for example, varies among countries and even among different cultures within one country. Moreover, there has been a historical trend downward in the average age of menarche, translating into a decrease of several months per decade from about 1840 to the present. This phenomenon is generally ascribed to the improved health and nutrition of children and adolescents.

Cognition. The dramatic physical and physiological changes characteristic of adolescence have an equally dramatic impact on cognitive and social functioning. Adolescents think about their "new" bodies and their "new" selves in qualitatively new ways. In contrast with sensorimotor and more limited spatiotemporal modes of thinking, which according to Piaget characterize infancy and childhood, beginning at about puberty the formal operational, or hypothetico-deductive, mode of thought emerges, characterized by reasoning and abstraction. In the formal operational stage, adolescents begin to discriminate between their thoughts about reality and reality itself and come to recognize that their assumptions have an element of arbitrariness and may not actually represent the true nature of experience. Thus, adolescent thinking becomes somewhat experimental in the scientific sense, employing hypotheses to test new ideas against outward reality.

Formal operational thought

In forming hypotheses about the world, adolescent cognition can be seen to grow along with formal, scientific, logical thinking. Consider, for example, a problem of combinatorial thought: An adolescent is presented with five jars, each containing a colourless liquid. Combining the liquids from three particular jars will produce a colour, whereas using the liquid from either of the two remaining jars will not produce a colour. The adolescent is told that a colour can be produced but is not shown which combination produces this effect. Children at the concrete operational stage typically try to solve this problem by combining liquids two at a time, but after combining all pairs, or possibly trying to mix all five liquids together, their search for the workable combination usually stops. An adolescent at the formal operational stage, on the other hand, will explore all possible solutions, systematically testing all possible combinations of two and three liquids until a colour is produced. As another example, consider adolescent thinking in respect to certain types of verbal problems—for instance, as represented by the question "If Jane is taller than Doris and shorter than Francine, who is the shortest of the three?" Concrete operational children may be able to solve an analogous problem (*e.g.*, one using sticks of various lengths, with the sticks actually present). Abstract verbal problems, however, are usually not solved until the capacity for formal operations has emerged.

Formal operational thought does not seem to be a stage characterizing all adolescents. Studies of older adolescents and adults in Western cultures show that not all individuals attain formal operations. In turn, in some non-Western groups there is a failure ever to attain formal operations. Some researchers have attributed these differences to the differences between rural and urban societies and the different kinds of schooling offered by each. There is, however, little evidence for socioeconomic or educational differences being associated with the achievement of formal operational thought.

Formal operational thinking also has limitations, predicated in part on the fact that adolescents often think about their own thinking. Just as the infant is preoccupied with his physical self in a world of new stimuli, so the adolescent may be preoccupied with his own thinking in a world of new ideas. Such preoccupation often leads to a kind of egocentrism, which can manifest itself in two ways: First, the individual may presume that his or her own concerns, values, and preoccupations are equally important to everyone else; second, the urgency of this new thinking may paradoxically give rise to an overestimation of one's uniqueness, often resulting in feelings of alienation or of being misunderstood. Although the formal operational stage is the last stage of cognitive development in Piaget's theory, the egocentrism of this stage diminishes over the course of the person's life, largely as a consequence of interactions with peers and elders and—most importantly—with the assumption of adult roles and responsibilities.

The social context. The adolescent's social context is broader and more complex than that of the infant and the child. The most notable social phenomenon of adolescence is the emergence of the marked importance of peer groups. The adolescent comes to rely heavily on the peer group for support, security, and guidance during a time when such things are urgently needed and since perhaps only others experiencing the same transition can be relied upon to understand what that experience is. Contrary to cultural stereotype, however, the family is quite influential for adolescents. Indeed, no social institution has as great an influence throughout development as does the family. Most studies indicate that most adolescents have few, if any, serious disagreements with parents. In fact, in choosing their peers adolescents typically gravitate toward those who exhibit attitudes and values consistent with those maintained by the parents and ultimately adopted by the adolescents themselves. For instance, while peers influence adolescents in regard to such issues as educational aspirations and performance, in most cases there is convergence between family and peer influences. While it is the case that adolescents and parents have somewhat different attitudes about issues of contemporary social concern (*e.g.*, politics, drug use, and sexuality), most of these differences reflect contrasts in attitude intensity rather than attitude direction. That is, rather than adolescents' and parents' standing on opposite sides of a particular issue, most generational differences simply involve different levels of support for the same position. In sum, there is virtually no evidence supporting the cultural stereotype of adolescence as a period of storm and stress. Most adolescents continue their close and supportive relationships with their parents, and their relationships with peers tend to support parental ideals rather than run against them.

Influence of the family

Personality. The dramatic changes that characterize puberty present the adolescent with serious psychosocial challenges. A person who has lived for 12 years has developed a certain sense of self as well as of self-capacity. In adolescence, however, this knowledge of self is challenged. As has been discussed, the rather sudden bodily changes in this period are accompanied by equally dramatic changes in thoughts and feelings. Thus, all the assumptions adolescents held about the self in earlier stages may no longer be relevant to the new individuals they find themselves to be. Because a coherent sense of self is necessary for functioning productively in society, adolescents ask a crucial psychosocial question: Who am I?

At precisely the time that adolescents feel unsure about who they are, society begins to ask them related questions. For instance, adolescents are expected to make the first steps toward career objectives. Society asks adolescents, then, what roles they will play as adults, that is, what socially prescribed set of behaviours they will choose to adopt. Thus a key aspect of this adolescent dilemma is that of finding a role, which is generally taken to be the outward expression of identity. The emotional upheaval provoked by this mandate is called the identity crisis. In order to resolve this crisis and achieve a sense of identity, it is necessary to synthesize psychological development and societal directives. The adolescent must find an orientation to life that not only fulfills the attributes of the self but at the same time is consistent with what society expects of a person; that is, a role cannot be something that is self-destructive (*e.g.*, sustained fasting) or socially disapproved (*e.g.*, criminal behaviour). In the search for an identity the adolescent must discover what he or she believes in and what his or her attitudes and ideals are. For commitment to a role entails, to a greater or lesser degree, commitment to a set of values.

The identity crisis

If the adolescent fails to resolve the identity crisis by the time of entry into adulthood, he or she will feel a sense of role confusion or identity diffusion. Some young adults waver between roles in a kind of prolonged "moratorium," or period of avoiding commitment. Others seem to avoid the crisis altogether and settle easily on an available, socially approved identity. Still others resolve their crises by adopting an available but socially disapproved role or ideology. This latter option is called negative identity formation and is often associated with delinquent behaviour. Resolution of the adolescent identity crisis has a profound influence on development during later adulthood.

All societies traditionally prescribe stereotyped roles to each sex. These roles have adaptive significance; that is, they allow society to maintain and perpetuate itself. From this reasoning, it follows that differences in sex-role behaviour, at least initially, arose from the different tasks males and females performed for survival—especially those tasks centred around reproduction. Differing biologies exert differing pressures on psychosocial development; however, these pressures do not occur independently of the demands of cultural and historical milieus. The biological basis of one's psychosocial functioning is believed to relate to adaptive orientations for survival. Many differences exist between males and females, but the nature of individual differences between the sexes is dependent on interactions among biological, psychological, sociocultural, and historical influences.

DEVELOPMENT IN ADULTHOOD AND OLD AGE

In sheer number of years, the periods labeled adulthood and aging constitute the major portion of the human life span. Historically, however, these periods were seen as less significant and interesting developmentally than infancy,

childhood, and adolescence. Adulthood was viewed as a time of continuity, a period when what had been developed earlier was utilized. Aging was viewed as a time of decline, a period when what had been developed earlier was lost. Contemporary opinion is that adulthood and aging are just as significant and interesting as are earlier periods of the life cycle. Adulthood and aging are characterized by both growth and decline.

Central nervous system processing. There is relatively clear evidence that, with advancing age, individuals show a tendency toward decreasing speed of response. This is a gradual change occurring across the entire life span that shows up in a variety of so-called speeded tasks (those in which errors would be unlikely if the individual had an unlimited amount of time to complete the tasks). For example, reaction time tests (which measure the time elapsing between the appearance of a signal and the beginning of a responding movement) are usually viewed as a measure of central nervous system processing. Mean speed of response on such tasks increases with age until the late teens, remains constant until the mid-20s, and then declines steadily throughout the remainder of the age range.

Considerable evidence has accumulated to link changes in brain electrical activity to the slowing of behaviour. The electroencephalogram (EEG) provides a record of the brain's electrical activity. The normal human EEG displays continuous rhythmic activity in the form of wavelike patterns varying in frequency and amplitude. The dominant rhythm is the alpha wave, which reaches its maximum frequency in adolescence and begins to slow gradually after young adulthood. This slowing may be related to disease processes (particularly vascular disease) as well as to basic aging processes. The older adult's central nervous system appears to be in a state of under-arousal in comparison to that of the younger adult.

Cognition. Decline in the rate of central nervous system processing does not necessarily imply a similar change in learning, memory, or other intellectual functions. However, considerable evidence indicates that the learning ability of young adults is superior to that of older adults and that the faster the pace of the task the more difference age makes. Older learners benefit more from slower pacing of tasks than do younger learners. When allowed to regulate the pace of the task themselves, older learners often show an improvement in performance, whereas this is not necessarily the case with younger learners.

Memory functions

However, in regard to memory, as opposed to the learning or acquisition of information, research suggests that there are relatively few age-related differences within the primary-memory system, the temporary maintenance system for conscious processing of information. Age is a significant factor, however, in the functioning of the secondary-memory system. Secondary memory depends on the elaboration and organization of information in terms of its semantic content or meaning. Compared to younger adults, older adults appear to be deficient in these processes. Generally, they do not spontaneously use organizational strategies as extensively as do younger adults, or, if they use them, they do so less effectively. However, when various organizational strategies are built into the situation, the performance of older adults improves significantly.

Most studies examining memory for general knowledge have found that older adults retrieve such information as well as or better than do younger adults. Within some contexts, older adults appear to integrate and retain the meaning of sets of sentences and texts as well as do younger adults. For example, when young, middle-aged, and elderly adults are asked questions covering such topics as famous people, news events, history, geography, the Bible, literature, sports, mythology, and general information (*e.g.,* "What was the former name of Muhammad Ali?"; "What is the capital of Kampuchea?"), their answers typically show no evidence of age differences in the retrieval of knowledge. In fact, elderly people may actually answer more questions correctly than do younger groups. Older adults also appear to have accurate knowledge about their own memory processes—knowledge that has been labeled metamemory. For example, research has found no age differences regarding subjects' assessments of the relative reliability of visual and verbal memory, regarding the use of memory strategies (*e.g.,* reminder notes), or regarding memory monitoring (*e.g.,* prediction of the number of items that would be recalled following various memory tasks).

Fluid and crystallized intelligence

Psychometric approaches to cognition suggest that intelligence is characterized by two distinct properties. Fluid intelligence, measured by tests that minimize the role of cultural knowledge, reflects the degree to which the individual has developed unique qualities of thinking through incidental learning. Crystallized intelligence, measured by tests that maximize the role of cultural knowledge, reflects the degree to which the individual has been acculturated through intentional learning. Fluid intelligence shows a steady decline from adolescence through middle age. Across the same age range, however, a steady increase occurs in crystallized intelligence. When measures of both properties are taken, few age-related differences appear.

Finally, there are age-related differences on several measures of cognitive functioning. In general, older adults perform more poorly than do younger adults on tasks requiring both concrete operational thought and formal operational thought.

Some features of cognitive functioning—speed of response, secondary memory, and fluid intelligence—seem to decline with age. Others—contextual memory and crystallized intelligence—increase. Aging does not inevitably precipitate a decline in cognitive functioning. Indeed, there is growing evidence that older persons can largely modify their intellectual performance, documenting a clear lifelong capacity for cognitive change in human beings.

Personality and social development. Several theories of personality development stress that adulthood and aging are periods of qualitative change, of discontinuity, and of transformations of earlier life patterns. These changes are believed to arise in relation to the demands of the person's changing biological status and social context—the family, the workplace, and society in general. Thus, personality development is both an individual and a social phenomenon.

Erikson's theory of crises

In the view of the German-born U.S. psychoanalyst Erik Erikson, certain psychosocial demands, or crises, confront the individual at distinct intervals throughout life. The young adult, for instance, is expected to enter into an institution—*i.e.,* marriage and family—that will perpetuate the society. The degree to which the basic need for intimacy on all levels—physical, emotional, and others—is met in such a relationship determines in most individuals the conception of the self as belonging or as isolated. In middle adulthood the crisis develops between the sense of generativity and the sense of stagnation. In this stage the individual is expected to play the role of a contributing, generative member of society. Generativity can take the form of providing the goods and services by which society functions or of producing, rearing, and socializing future members of society. The inability to develop a productive self-conception results in a feeling of stagnation. In maturity, according to Erikson, a crisis arises with regard to the sense of ego integrity versus the sense of despair. In this stage, individuals realize that they are reaching the end of life. If they have successfully progressed through the previous stages of development, they can face old age with satisfaction in the feeling that a full and complete life has been led. Individuals for whom this integrity of life is lacking often feel a sense of despair over "wasted" opportunity.

The American psychologist Daniel J. Levinson also divides adult life into qualitatively distinct periods. Confining his study to men, Levinson identified five eras within their lives that are not stages of biological, psychological, or social development but that together constitute a life-cycle structure. The eras are: (1) preadulthood (birth to age 22); (2) early adulthood (age 17 to 45); (3) middle adulthood (age 40 to 64); (4) late adulthood (age 60 to 85); and (5) late late adulthood (age 80 and over). Each of these eras is in turn made up of a series of developmental periods and transitions. For example, in early adulthood a first major transition, ordinarily beginning at

age 17 to 18 and extending until age 22 to 23, represents a developmental link between preadulthood and early adulthood. The young man in this early adult transition faces two major tasks. The first task is to modify relationships with his family and with other persons, groups, and institutions significant to his preadult world. The second task is to take a preliminary step into the adult world. This requires making initial explorations and choices for adult living. Major life events within this transition may include graduating from high school, moving out of the family home, seeking gainful employment, or attending college. Entering the adult world begins in the early 20s and extends until the late 20s. The focus of this period is on exploration and provisional commitment to adult roles and responsibilities. The young man faces two antithetical tasks. On the one hand, he must explore alternate possibilities for adult living, keeping options open and avoiding strong commitments. On the other hand, he must create a stable life structure, becoming responsible and "making something" of himself. Crucial life events during this period include occupational choice, first job, marriage, and the birth of children.

The age range of 28 to 33 years represents a transition between the period of entering the adult world and the next period of settling down. This transition provides the young man with an opportunity to adjust and enrich the provisional adult life structure that he created earlier. For most men, however, a moderate to severe crisis is common; divorce and occupational change are frequent during this time. A settling down period then follows, beginning in the early 30s and extending until about age 40. This period, during which the man's task is to become a full-fledged adult, emphasizes stability and security. The individual makes deeper commitments to his occupation, family, or whatever enterprises are significant to him. In addition, he generally concentrates on "making it." This involves long-range planning toward specific goals with a timetable for their achievement. Most men fix on a key life event, such as a promotion or new job, as representative of ultimate affirmation or evaluation by society. During the last years of the settling down period, there is a distinctive phase designated as "becoming one's own man," ordinarily occurring at age 36 to 40. The man's major task during this phase is to achieve greater independence and authority by striving for the goals of his various enterprises.

The mid-life transition spans four to six years, reaching a peak in the early 40s. It forms a developmental link between early adulthood and middle adulthood and, being part of both eras, represents a beginning and ending, a meeting of past and future. A task of the mid-life transition is to work on and partially resolve this discrepancy between what is and what might be. The transition may be relatively smooth, but it is more likely to involve considerable turmoil. The period of entering middle adulthood begins at about age 45 and extends until about age 50. Sometimes the start of this new life structure is marked by a significant life event, such as a change in job or occupation, a divorce or love affair, or a move to a new community. In other instances, the changes are more subtle.

Evidence of constancy in adult personality

Research evidence does not unequivocally support the discontinuous, stagelike changes in adult personality proposed by theorists such as Erikson and Levinson. In fact, several major studies of personality development during the adult and aged years present evidence for both change and constancy. For instance, studies of healthy adults between the ages of 40 and 80 residing in the Kansas City area during the 1950s found evidence for both continuity and change of adult personality. On the one hand, personality structure was stable; four personality types—integrated, defended, passive-dependent, and unintegrated—emerged among respondents regardless of age. Similarly, characteristics dealing with the socio-adaptational aspects of personality (*e.g.*, goal-directed behaviour, coping styles, and life satisfaction) were not age-related. It seems, therefore, that the ways in which healthy adults interact with the environment may be stable even though the roles they adopt alter with age. On the other hand, individual styles of coping with the inner world of experience showed marked age differences. For example, 40-year-olds felt in charge of their environment, viewed the self as a source of energy, and were positive about risk taking, whereas 60-year-olds saw the environment as threatening and even dangerous and viewed the self as passive and accommodating.

Other studies of personality development from birth through early adulthood have also found evidence for constancy and change. For example, behavioral dispositions of the early school years, including passive withdrawal from stressful situations, dependency on family, arousal of anger, involvement in intellectual mastery, sexual behaviour and sex-role identification, and anxiety over social interaction, have been found to carry over into adulthood. The degree of stability in these behaviours exhibited from childhood to adulthood seems to be closely related to cultural expectations of appropriate sex-role behaviour. If a pattern of childhood behaviour is consistent with sex-role expectations, it will more likely remain stable over time.

Other studies, concentrating on the development from adolescence into adulthood among people born in California in the late 1920s and early 1930s, found different personality types for males and females that showed substantial stability over time. Personality characteristics reflecting socialization and self-presentation, for example, tended to remain stable. On the other hand, two major types of personality characteristics, those reflecting information processing and those reflecting interpersonal relations, tended to change. There were differences between the sexes both in what was constant and in what changed across life. For example, life-style patterns among the parents of the California children were more continuous between young adulthood and old age for fathers than they were for mothers. Fathers who in early adulthood were unwell and disengaged from their families showed these same characteristics in late adulthood. In regard to personality, however, there was more apparent continuity between young adulthood and old age among mothers than among fathers. For example, mothers who were group-centred showed a more psychologically healthy personality and had a more satisfying life in later years than in earlier years, whereas non-group-centred mothers were happy and healthy at age 30 but lost their health and physical stamina by age 70.

In general, older adults tend to engage in greater introspection and self-reflection than do younger adults, showing a general movement from the outer world toward the inner world. They tend to withdraw emotional investments, be less assertive, and avoid challenges. Adulthood and old age involve both constancy and change. These periods of life are continuations of the past as well as new phases in their own right.

CONCLUSIONS

This article treats as separate various substantive spheres of human development—physical, perceptual, cognitive, linguistic, personality, and social—as it does various temporal phases of development—prenatal life, infancy, childhood, adolescence, adulthood, and old age. However, human beings are coherent wholes, and behavioral development is unified, so that development in any one arena of life at any one time is ineluctably interrelated with development in other arenas at the same and at other times in patterns of mutual influence. In the life course, genetic endowment and biology interact with cultural context and experience to shape the development of human behaviour. Each of these powerful sources of influence on development has distinctive characteristics, and it is their transaction over time as well as the degree of congruence between the two that influence outcome. Our elucidation of the processes that underpin human growth is central to understanding normal as well as abnormal development. Study of the development of human behaviour is unwieldy; life does not submit to elegant scientific analysis or to precise prediction. Therefore, developmental study takes as its goals the general description and explanation of origins, of constancy, and of change in perceiving, thinking, feeling, and behaving. Any such undertaking requires constant reconsideration in light of new data and new insights.

BIBLIOGRAPHY. Still the most popular and practical account of physical and (to some degree) psychological development is BENJAMIN SPOCK and MICHAEL B. ROTHENBERG, *Baby and Child Care,* rev. ed. (1985). Authoritative texts include THOMAS M. ACHENBACH, *Developmental Psychopathology,* 2nd ed. (1985); JAMES E. BIRREN (ed.), *The Handbooks of Aging,* 2nd ed., 3 vol. (1985); MARC H. BORNSTEIN and WILLIAM KESSEN, *Psychological Development from Infancy: Image to Intention* (1979); MARC H. BORNSTEIN and MICHAEL E. LAMB (eds.), *Developmental Psychology: An Advanced Textbook* (1984); URIE BRONFENBRENNER, *The Ecology of Human Development: Experiments by Nature and Design* (1979); ERIK H. ERIKSON, *Childhood and Society,* 2nd ed. (1964, reprinted 1985); SIGMUND FREUD, *An Outline of Psycho-analysis,* rev. ed. (1970; originally published in German, 1940); E. MAVIS HETHERINGTON and ROSS D. PARKE, *Child Psychology: A Contemporary Viewpoint,* 3rd ed. (1986); WILLIAM KESSEN, *The Child* (1965); MICHAEL E. LAMB and MARC H. BORNSTEIN, *Development in Infancy: An Introduction* (1986); RICHARD M. LERNER, *Concepts and Theories of Human Development,* 2nd ed. (1986); RICHARD M. LERNER and DAVID F. HULTSCH, *Human Development: A Life-Span Perspective* (1983); DANIEL J. LEVINSON *et al., The Seasons of a Man's Life* (1978); PAUL H. MUSSEN (ed.), *Handbook of Child Psychology,* 4th ed. (1983); JEAN PIAGET, *The Origins of Intelligence in Children* (1952, U.K. title, *The Origin of Intelligence in the Child,* 1953; originally published in French, 2nd ed., 1948); LAWRENCE KOHLBERG, *Essays on Moral Development* (1981–), with two volumes appearing to 1986; ROBERT PLOMIN, *Genetics, Development, and Psychology* (1986); and EDWARD O. WILSON, *Sociobiology: The New Synthesis* (1975, reprinted 1979). Journals include *International Journal of Aging & Human Development* (8 no. annually); *Child Development* (quarterly); *Developmental Psychology* (bimonthly); *Developmental Review* (quarterly); *Human Development* (quarterly); and *Journal of Experimental Child Psychology* (bimonthly).

(M.H.Bo./Ri.M.L.)

Beirut

Beirut (Arabic Bayrūt; French Beyrouth) is the capital, largest city, and main port of the Republic of Lebanon. It occupies a metropolitan area of approximately 26 square miles (67 square kilometres) on the Mediterranean coast, at the foot of Mt. Lebanon. It comprises two hills, al-Ashrafiyah (East Beirut) and al-Muṣayṭibah (West Beirut), that protrude into the sea as a roughly triangular peninsula; in the immediate hinterland lies a narrow coastal plain (as-Sāḥil) that extends from the mouth of the Nahr al-Kalb (Dog River) in the north to that of the Nahr ad-Dāmūr (Damur River) in the south.

The article is divided into the following sections:

Physical and human geography

THE LANDSCAPE

The city plan. Under the Ottoman *vilâyet* administration and the French Mandate, the growth of Beirut was planned, but after independence in 1943 it was as haphazard as it was rapid. It is estimated that the population of the city increased tenfold between the early 1930s and early 1970s and the city's area grew to three times the size it had been in 1900. By the 1950s, few traces of the old city were left, and most of those were destroyed in the 1974–76 Lebanese civil war and its aftermath.

The neighbourhoods. Street plans and block arrangements in the city and its suburbs are not consistent or uniform. In most quarters, modern high-rise buildings, walk-up apartments, slum tenements, modern villas, and traditional two-story houses with red-tiled roofs—all in varying states of repair—stand side by side. After 1974, countless houses and apartments, particularly in West Beirut, were forcibly occupied by refugees from rural areas, especially from the Shīʿī areas of South Lebanon.

Civil war destruction

The downtown area of central Beirut (the old city) was destroyed during the civil war and remained in ruins, a vacant belt between East and West Beirut that could not be reconstructed because of sporadic fighting there between rival factions. As a result, all business moved out of the area to establish new premises in the Christian and Muslim sides of the city. While few areas of Beirut were purely residential before 1974, none was by 1980.

THE PEOPLE

Ethnic and religious groups

The resident population of Beirut is more or less evenly divided between Muslims and Christians. The overwhelming majority in both religious groups is ethnically Arab and includes Palestinian refugees, Syrian residents, and others. The most important ethnic minority is the Christian Armenians; there is also a Kurdish ethnic minority among the Muslims. East Beirut is almost solidly Christian, West Beirut is predominantly Muslim, and a number of mixed neighbourhoods (notably in the district of Ra's Bayrūt) are cosmopolitan in character. The Jewish community is concentrated in the neighbourhood of Wādī Abū Jamīl. The larger Christian communities are the Maronites and the Greek Orthodox; the Christian minorities, apart from the Armenians, include Greek Catholics, Protestants, Roman Catholics, and others. Originally, the Sunnīs were the dominant Muslim community, but Shīʿī Muslims began moving into the city in increasing numbers from the 1960s. Small numbers of Druzes live in parts of West Beirut.

ECONOMIC AND POLITICAL CONDITIONS

Between 1952 and 1975, Beirut was the hub of economic, social, intellectual, and cultural life in the Arab Middle East. In an area dominated by authoritarian or militarist regimes, the Lebanese capital was generally regarded as a haven of liberalism, though a precarious one. With its seaport and airport—coupled with Lebanon's free economic and foreign exchange system, solid gold-backed currency, banking-secrecy law, and favourable interest rates—Beirut became an established banking centre for Arab wealth, much of which was invested in construction, commercial enterprise, and industry (mostly the manufacture of textiles and shoes, food processing, and printing). Foreign banking and business firms found in Beirut an ideal base for their operations in the Arab Middle East. The "free zone" of the Beirut port was a leading entrepôt for the region. A skilled professional class provided varied sophisticated services for a pan-Arab clientele. Beirut was also a centre for tourism. The large number of daily and weekly newspapers, journals, and other periodicals, which were normally uncensored, kept the Arab world informed about regional and world developments and provided a full array of editorial opinion. Beirut's schools, colleges, and universities—the American University of Beirut, Université Saint Joseph, Université Líbanaise (Lebanese University), and Beirut Arab University—attracted students from many Arab countries. An underlying lack of consistency and organization, however, and an undercurrent of social and political unrest never escaped notice.

Political terrorism

Beirut became a prominent centre for Palestinian resistance organizations after the Arab–Israeli War of 1967,

and became the headquarters of the movement after the Palestine Liberation Organization (PLO) in Jordan was crushed in 1970. The Lebanese government failed in repeated attempts to bring the Palestinian movement in Beirut and the rest of Lebanon under control. Arab nationalist and leftist political parties established armed militias for themselves, frequently in association with the Palestinian resistance movement. When the PLO was trapped by Israeli troops in West Beirut in 1982, the organization was removed from Lebanon by multinational forces. Sectarian violence continued after the Israeli withdrawal, destroying the established order of the city.

Beirut's destruction

Beirut is no longer the hub of the Arab Middle East. The incessant fighting that began with the civil war, the Israeli bombing of West Beirut, and the chronic shelling that continued after the Israeli withdrawal ate away at the city's infrastructure. The Green Line established in the 1970s to separate the Christian and Muslim factions in East and West Beirut, respectively, became a dangerous barricade permanently dividing the city. Businesses and residents alike left the city as hopes for a cease-fire waned, and even basic services such as water and electricity came to be only sporadically available.

History

THE EARLY PERIOD

The antiquity of Beirut is indicated by its name, derived from the Canaanite name of Be'erōt (Wells), referring to the underground water table that is still tapped by the local inhabitants for general use. Although the city is mentioned in Egyptian records of the 2nd millennium BC, it did not gain prominence until it was granted the status of a Roman colony, the Colonia Julia Augusta Felix Berytus, in 14 BC. The original town was located in the valley between the hills of al-Ashrafīyah and al-Muṣayṭibah. Its suburbs were also fashionable residential areas under the Romans, who constructed an aqueduct to augment the city's water supply. Between the 3rd and 6th centuries AD, Beirut was famous for its school of law. The Roman city was destroyed by a succession of earthquakes, culminating in the earthquake and tidal wave of AD 551. When the Muslim conquerors occupied Beirut in 635, it was still mostly in ruins.

Arab and Christian rule. Beirut was reconstructed on a small scale by the Muslims and reemerged as a small, walled garrison town administered from Baalbek as part of the *jund* (Muslim province) of Damascus. Until the 9th or 10th century, it remained commercially of no significance and was notable mainly for the careers of two eminent local jurists, al-Awzāʿī (died 774) and al-Makḥūl (died 933).

The Crusaders

A return of maritime commerce to the Mediterranean in the 10th century revived the importance of the town, particularly after Syria passed under the rule of the Fāṭimid caliphs of Egypt in 977. In 1110 Beirut was conquered by the military forces of the First Crusade and was organized, along with its coastal suburbs, as a fief of the Latin Kingdom of Jerusalem.

As a crusader outpost, Beirut conducted a flourishing trade with Genoa and other Italian cities; strategically, however, its position was precarious because it was subject to raids by the Druze tribesmen of the mountain hinterland. Saladin reconquered Beirut from the crusaders in 1187, but his successors lost it to them again 10 years later. It was the Mamlūks who finally drove the crusaders out of the town in 1291. Under Mamlūk rule, Beirut became the principal port of call in Syria for the spice merchants from Venice.

Ottoman rule. Beirut, along with the rest of Syria, passed under Ottoman rule in 1516, shortly after the Portuguese had rounded the African continent (1498) to divert the spice trade of the East away from Syria and Egypt. The commercial importance of Beirut declined in consequence. By the 17th century, however, the city had reemerged as an exporter of Lebanese silk to Europe, mainly to Italy and France. Beirut at the time was technically part of the Ottoman province (*eyalet*) of Damascus, and after 1660 of Sidon. Between 1598 and 1633, however, and again between 1749 and 1774, it fell under the control of the Maʿn and Shihāb emirs (feudal suzerains and fiscal agents) of the Druze and Maronite mountain hinterland. From the mid-17th to the late 18th century, Maronite notables from the mountains served as French consuls in Beirut, wielding considerable local influence. During the Russo-Turkish War of 1768–74, the town suffered heavy bombardment by the Russians. Subsequently it was wrested from the Shihāb emirs by the Ottomans, and it soon shrank into a village of about 6,000.

The growth of modern Beirut was a result of the Industrial Revolution in Europe. Factory-produced goods of the Western world began to invade the markets of Ottoman Syria, and Beirut, starting virtually from nought, stood only to profit from the modern industrial world. The occupation of Syria by the Egyptians (1832–40) under Muḥammad ʿAlī Pasha provided the needed stimulus for the town to enter on its new period of commercial growth. A brief setback came with the end of the Egyptian occupation; by 1848, however, the town had begun to outgrow its walls, and its population had increased to about 15,000. Civil wars in the mountains, culminating in a massacre of Christians by Druzes in 1860, further swelled Beirut's population, as Christian refugees arrived in large numbers. Meanwhile, the pacification of the mountains under an autonomous government guaranteed by the great powers (1861–1914) stabilized the relationship between the town and its hinterland, where traditional silk production was mechanized by French and local industrial concerns. In 1886 Beirut was made the capital of a separate province (*vilâyet*) comprising the whole of coastal Syria, including Palestine. By the turn of the century, it was a city of about 120,000.

Egyptian occupation

Meanwhile, Protestant missionaries from Great Britain, the United States, and Germany and Roman Catholic missionaries mainly from France became active in Beirut, particularly in education. In 1866 the American Protestant Mission established the Syrian Protestant College, which later became the American University of Beirut. In 1881 French Jesuit missionaries established the Université Saint Joseph. Printing presses, introduced earlier by Protestant and Roman Catholic missionaries, stimulated the growth of the city's publishing industry, mainly in Arabic but also in French and English. By 1900 Beirut was in the vanguard of Arabic journalism. A class of intellectuals sought to revive the Arabic cultural heritage and become the first spokesmen of a new Arab nationalism.

MODERN BEIRUT

Beirut was occupied by the Allies at the end of World War I, and the city was established by the French mandatory authorities in 1920 as the capital of the State of Greater Lebanon, which in 1926 became the Lebanese Republic. The Muslims of Beirut resented the inclusion of the city in a Christian-dominated Lebanon and declared loyalty to a broader pan-Arabism than most Christians would support. The resultant conflict became endemic. The accelerated economic growth of Beirut under the French Mandate (1920–43) and after produced rapid growth of the city's population and the rise of social tensions. These tensions were increased by the influx of thousands of Palestinian refugees after 1948. The political and social tensions in Beirut and elsewhere in Lebanon, coupled with Christian–Muslim tensions, flared into open hostilities in 1958, and even more violently in 1974–76. The conflict continued to simmer, with sporadic eruptions of violence, and Beirut became a divided city.

Muslim–Christian tensions

West Beirut was largely destroyed by heavy fighting between Israeli forces and members of the Palestine Liberation Organization (PLO) in 1982, when Israel launched a full-scale attack on PLO bases operating in the city. Israeli troops surrounded West Beirut, where most PLO guerrilla bases were located, and a series of negotiations brought about the evacuation of PLO troops and leaders from Lebanon to other Arab nations.

Divisive sectarian loyalties only increased after the Israeli withdrawal. Neither the continued Syrian military presence nor the formation of a coalition government could defuse the violence. The shelling persisted, and much of the population fled. (K.S.S./Ed.)

BIBLIOGRAPHY. SAID CHEHABE ED-DINE, *Géographie humaine de Beyrouth* (1960), though slightly outdated, is the only study that surveys the human geography of Beirut and accounts for its growth patterns. C.W. CHURCHILL, *The City of Beirut* (1954), is a descriptive, socioeconomic survey that provides detailed information on household composition, education, mobility, occupations, housing, income saving, and expenditure. FUAD I. KHURI, *From Village to Suburb* (1975), is a detailed account of the social life and organization of two peripheral villages engulfed by the Beirut metropolis. The BEIRUT EXECUTIVE BOARD OF MAJOR PROJECTS, *Comprehensive Plan Studies for the City of Beirut* (1968), a preliminary but comprehensive survey report, provides detailed information on the physical features, land use patterns, utilities, population, and economic characteristics of the city. See also S. KHALAF and P. KONGSTAD, *Hamra of Beirut: A Case of Rapid Urbanization* (1973), an empirical study that explores the ecological transformation and the social structure of one of Beirut's urban communities; and HARVEY PORTER, *The History of Beirut* (1912), a brief but instructive historical sketch of the city from the earliest times to the beginning of the 20th century.

(S.G.K.)

Belgian Literature

The literature of Belgium falls into two main divisions by language: Flemish (the Netherlandic language as spoken in Belgium, and equivalent to Dutch in The Netherlands) and French. To this may be added a third, Walloon, a literature written in local dialects of French and Latin origin that are spoken in the provinces of Hainaut, Liège, Namur, Luxembourg, and the south of Brabant. (These provinces together were known as Wallonia.) This article provides a brief historical account of the development of each of these three literary traditions.

The article is divided into the following sections:

Flemish

EARLY LITERATURE

Relationship with Dutch literature. When considering the literature of Flemish-speaking (Dutch-speaking) Belgium it must be remembered that the Belgian territories were united with the Netherlands politically, economically, and culturally until 1579, when, as a result of the Reformation, the northern (Reformed) provinces seceded from the Roman Catholic south. Thus until the early 17th century the literature of Flanders and Holland must be considered as a whole (see DUTCH LITERATURE). It was in Flanders that the literature of the medieval Low Countries flowered most profusely. It was, moreover, in Flanders and Brabant that learning showed new vigour under the influence of the Renaissance and the Reformation. In literature inspired by the Reformation the tone was set by the glowing satiric verse of the Catholic Anna Bijns and the polemical satire on the Catholic Church, *Biencorf der H. Roomsche Kercke* (1569; "The Beehive of the Roman Catholic Church"), of the Calvinist Philips van Marnix van Sint Aldegonde. The Renaissance in the Netherlands began with Lucas de Heere, Karel van Mander, and Jan Baptista van der Noot, all of whom, significantly, had fled from the south for religious reasons.

Decline. Many left the south as a result of the religious and political troubles before 1579, and the literary revival in Flanders and Brabant was interrupted. Whereas Holland was approaching its golden age, in the south a decline set in. But Justus de Harduwijn, a lyrical poet in the Classical style of the French Pléiade; Richard Verstegen, a polemicist and writer of prose characters; Adriaen Poirters, a popular moralist; the dramatists Willem Ogier and Cornelis de Bie, and, especially, Michiel de Swaen, the last important poet and playwright of the Baroque period, who was deeply inspired by his religion, compare favourably with most writers of their time. The decline was most noticeable in the early 18th century, when the aristocracy and intellectual elite came increasingly under French influence.

Revival. Before the end of the 18th century, however, Willem Verhoeven and Jan Baptist Verlooy had started a reaction against this French influence. Like contemporary historical and scientific writers they reverted to the work of the 16th-century Humanists but neglected the medieval masterpieces. Revival was helped by the *rederijker*s (rhetoricians), who continued, more or less successfully, to use Flemish, not French. Karel Broeckaert wrote dialogues modelled on Joseph Addison's *Spectator* essays in a spirit of rational liberalism, creating a literary figure, "Gysken," the ironic representative of the ancien régime; he also wrote the first Flemish prose story, *Jellen en Mietje* (1811). The poet Pieter Joost de Borchgrave embodied the transition from Classicism to Romanticism, and Jan Baptist Hofman, a prolific playwright, introduced middle-class sentimental tragedy, or *drame bourgeois*.

19TH-CENTURY LITERARY TRENDS

The Romantic movement. Romanticism made its influence felt in the 19th century and was linked to a revival of nationalist consciousness. The older generation—Jan Frans Willems, Jan Baptist David, Philip Blommaert, and Ferdinand Snellaert—while remaining rationalists, rediscovered the rich medieval inheritance. To their group belonged two important poets of the new age, Karel Lodewijk Ledeganck and Prudens van Duyse. The younger generation was more spontaneously Romantic, as was illustrated by the work of Hendrik Conscience, creator of the Flemish novel. Theodoor van Rijswijck and Johan Alfried de Laet freed poetry from Classical concepts and forms, and the ultra-Romantic stories of Eugeen Zetternam and Pieter Frans van Kerckhoven denounced social evils.

Realism and other post-Romantic trends. Led by a Realist, Domien Sleeckx, a reaction against Romanticism set in about 1860. Writing became characterized by acute observation, description of local scenery, humour, and, not infrequently, a basic pessimism, as could be seen in novels such as Anton Bergmann's *Ernest Staes* (1874) and Virginie Loveling's *Een dure eed* (1892; "A Dear Oath"). The poets Johan Michiel Dautzenberg, Jan van Beers, and Rosalie Loveling, together with the first important Flemish critic, Max Rooses, also reflected a new Realistic approach.

The reaction of Flemish Realism

Parallel to this Realism was a remarkable revival in poetry in West Flanders, headed by a Roman Catholic priest, Guido Gezelle, undoubtedly the greatest Flemish poet of the 19th century. Albrecht Rodenbach wrote militant songs, thoughtful lyrics, monumental epics, and a verse tragedy, *Gudrun* (1882).

The review *Van Nu en Straks* (1893–1901; "Of Now and Later"), which was to make Flemish literature of European importance, was influenced more by Gezelle and Rodenbach than by the Dutch generation of the 1880s. Led by Pol de Mont, an already complex modern poet, the writers of the 1880s had, however, widened horizons and, by emphasizing individualism and "art for art's sake," prepared the ground for their successors.

THE 20TH CENTURY

The turn of the century. The writers grouped around *Van Nu en Straks* helped to bring about a true revival of Flemish culture. Though they held a wide variety of opinions, they all strove for an art that would comprehend all human activity, and in which individual feelings would be given universal significance. In his masterly essays and his symbolic novel *De wandelende Jood* (1906; "The Wandering Jew"), their leader, August Vermeylen, advocated a rationalism infused with idealism. Prosper van Langendonck, on the other hand, interpreted the incurable suffering of the *poète maudit.* In 1898 Emmanuel de Bom published *Wrakken* ("Wrecks"), the first modern Flemish psychological novel, and *Starkadd,* an early Wagnerian drama by Alfred Hegenscheidt, was produced.

The poetry and prose of Karel van de Woestijne formed a symbolic autobiography of a typical fin de siècle personality, the sophisticated, world-weary sensualist striving for spiritual detachment. His work, a passionate confession of human frailty, represents one of the great achievements of European Symbolism.

Stijn Streuvels, a master of prose, made the West Flemish rural landscape his microcosm, a visionary world in which man is dwarfed by nature (*De Vlaschaard,* 1907; "The Flax Field"). The polished work of Herman Teirlinck, novelist, dramatist, and essayist, was characterized by imagination, sensuality, and a sonorous vocabulary. In the stylistically refined stories of F.V. Toussaint van Boelaere there were often tragic undertones.

As well as that resulting from the influence of *Van Nu en Straks,* other important tendencies developed. Naturalism found a first representative in Reimond Stijns. It reached its height in the robust tales and pithy plays of Cyriel Buysse and in the regional novel, as exemplified by the evocations of Bruges by Maurits Sabbe and the vivid treatment of Antwerp life by Lode Baekelmans. The pictorial verse of Omer Karel de Laey and the vigorous rhythms of René de Clercq have permanent value. The poems, plays, and essays of Cyriel Verschaeve are passionate expressions of his Augustinian outlook.

The group associated with the review *De Boomgaard* (1909–11; "The Orchard"), which included André de Ridder and Paul Gustave van Hecke, strove to be more cosmopolitan than *Van Nu en Straks* and defended a more dilettante attitude to culture. The elegiac poet Jan van Nijlen had affinities with this group.

After World War I. During World War I there was a new flowering of the picturesque regional tale: *Pallieter* (1916), by Felix Timmermans, and the roguish *De witte* (1920; *Whitey*), by Ernest Claes, became known outside Flanders. From the poetry of August van Cauwelaert and the prose of Franz de Backer it could be seen that the generation that fought in the war stressed life rather than literature. But a trend first revealed during the German occupation found its most direct outlet in revolutionary Expressionism, and the review *Ruimte* (1920–21; "Space") published its manifesto: ethics must take priority over aesthetics, and the art of the community over that of the individual. Expressionism was most apparent in lyrical poetry and drama. Wies Moens' poetry reflected this humanitarian trend, whereas Gaston Burssens remained less pathetic and more playful. The outstanding lyricist of the movement was Paul van Ostaijen, who expressed faith in humanity in *Het sienjaal* (1918; "The Signal") but soon went through a crisis of Dadaism, adopted rhythmic typography (*Bezette stad,* 1921), and wrote pure poetry concentrating on word and sound, grotesque verse and prose, and penetrating essays on art and poetry. The review *'t Fonteintje* (1921–24; "The Little Fountain"), whose editors included Richard Minne and Maurice Roelants, reacted against Expressionism. Expressionism also gave new life to drama. In the 1920s the Flemish Popular Theatre became one of the foremost avant-garde theatres in Europe. The standard of drama was raised by Herman Teirlinck and Anton van de Velde.

Flemish Expressionism

By 1930 the tide of Expressionism had run out and the novel came into its own. The regional novel was replaced by the psychological novel, introduced by Roelants with *Komen en gaan* (1927; "Coming and Going"), and was raised to great stylistic heights by Maurice Gilliams (*Elias,* 1936) who was also a subtle poet and essayist. Gerard Walschap wrote of man's social, religious, and moral problems, and Lode Zielens wrote about the lives of the poor.

The focal point of the novel was man. The skeptical Raymond Brulez and the disenchanted humanist Marnix Gijsen were both more or less detached observers of human weaknesses.

In Willem Elsschot's short but superb novels, such as *Lijmen* (1924; *Soft Soap*) and *Kaas* (1933; "Cheese"), caustic irony and an astringent style mask the underlying compassion. The new tone was set by the "personalistic" poets of the *Vormen* (1936–40; "Forms") group, of whom Pieter Geert Buckinx is representative.

After World War II. The major writers of World War II and the postwar period were novelists. The range of subjects and styles in the novel was remarkable. The problem novels by Paul Lebeau and Gaston Duribreux, the "magic-realist" work of Johan Daisne and Hubert Lampo, the Social Realism of Piet van Aken (*Het begeren,* 1952; "Desire") and Louis-Paul Boon (*De kapellekensbaan,* 1953; *Chapel Road*), the Existentialism of Jan Walravens (*Negatief,* 1958; "Negative"), and Hugo Claus's experimental novels are but a sample. Boon, Walravens, and Claus belonged to a review group called *Tijd en Mens* (1949–55; "Time and Man"), which was marked by postwar chaos, rebellion, and Experimentalism. Boon and Claus eventually became recognized as the outstanding postwar novelists.

In the 1960s the experimental trend in the novel led to new prose either based on stream-of-consciousness association (Hugo Raes, Ivo Michiels, and Paul de Wispelaere) or consisting of introverted "texts" dwelling largely on the act of writing itself (Willy Roggeman and Daniel Robberechts). Nevertheless, the tradition proved to be fertile; *e.g.,* in the satiric and allegorical novels by Ward Ruyslinck and in Jef Geeraerts' violent colonial novels. Walter van den Broeck later emerged as an accomplished writer, employing a mixture of autobiography and social history.

In postwar Flemish poetry the impact of Experimentalism made itself felt in the work of Albert Bontridder and Hugo Claus, whose *Oostakkerse gedichten* (1955; "Oostakker Poems") has remained a milestone. The playful Paul Snoek and the sombre Hugues Pernath continued the experimental line. In the 1970s a Neorealist reaction set in (Herman de Coninck, Roland Jooris), followed by a Neoromantic revival (Eddy van Vliet). The poetry of Freddy de Vree, on the other hand, is more intellectual.

Postwar drama, at first still dominated by Teirlinck, saw new talent emerging in Jozef van Hoeck (*Voorloping vonnis,* 1957; "Provisional Verdict"), in the literary plays of Herwig Hensen, and in the political theatre of Tone Brulin, but especially in the many original plays and adaptations of Claus, such as *Suiker* (1958; "Sugar") and *Vrijdag* (1969; *Friday*). Van den Broeck later made his mark with socially committed and naturalistic work.

French

BEGINNINGS

In the history of French literature, that by Belgian writers in French forms an important chapter. Even before Belgium achieved independence in 1830, many outstanding works were written in French by writers of Flemish

origin. They were responsible for some of the medieval chansons de geste. In the Middle Ages too, didactic, religious, and lyrical poetry, plays, and chronicles began to be written. The names of Jean Le Bel, Jean Froissart, Georges Chastellain, and Philippe de Commynes indicate the wealth of early historiography by Flemish writers, while Jean Lemaire de Belges was one of the great late medieval poets and rhetoricians (*rhétoriqueurs*).

The death of Margaret of Austria (1530) was followed by a period of literary sterility, prolonged until the end of the ancien régime by unstable economic conditions, the indifference to native culture of successive foreign governments, and the strong influence of 17th- and 18th-century French literature. Only a few writers are remembered, and notable among them is Charles Joseph, prince de Ligne.

Between the end of the 18th century and 1880 attempts were made to create an original, native literature. The poet André van Hasselt is still remembered. More important were the novelist Charles de Coster, whose *Légende . . . d'Ulenspiegel* (1867; *The Glorious Adventures of Tyl Ulenspiegel*) has attained the status of an epic of Flanders, and the influential essayist Octave Pirmez.

THE "JEUNE BELGIQUE" MOVEMENT

Impetus for the long-awaited literary renaissance came from Max Waller, founder in 1881 of an influential review, *La Jeune Belgique* ("Young Belgium"), which indicated a possible national literary consciousness; essentially, however, the review was the vehicle of expression of individual writers dedicated to the idea of art for art's sake.

Of novelists early associated with the movement, Camille Lemonnier, Eugène Demolder, and Georges Eekhoud were the most influential. A later *Jeune Belgique* novelist was Georges Rodenbach, celebrator of silence and spirituality, whose *Bruges-la-morte* (1892; "Bruges the Dead City") was the epitome of decadent fiction.

Stimulated by the *Jeune Belgique* movement was a group of poets much concerned with style and language. Among them were Grégoire Le Roy, a gifted lyrical Symbolist poet; Charles van Lerberghe, who explored the potential of Symbolist verse; and Albert Mockel, founder of an influential Symbolist review, *La Wallonie.*

They were overshadowed, however, by three poets of international stature: Émile Verhaeren, Maurice Maeterlinck, and Max Elskamp. Verhaeren extolled humanity's struggle toward social justice; Maeterlinck, creator of the Symbolist poetic drama, illuminated life's inner meaning; and Elskamp treated themes from folklore and legend.

Outstanding dramatists were Maeterlinck (*L'Oiseau bleu,* performed 1908; *The Blue Bird*) and Verhaeren. Edmond Picard, a playwright, novelist, and critic, propounded Socialism, Symbolism, and Impressionism. The period also saw a beginning, in the work of Godefroid Kurth, of modern historiography and criticism. An outstanding historian was Henri Pirenne. Art and literary criticism flourished, and an atmosphere for a flowering of scholarship culminated in the foundation (1920) of the Belgian Académie Royale de Langue et de Littérature Françaises.

THE MODERN PERIOD

Between World Wars I and II. A new generation of Belgians who wrote in French arose between World Wars I and II. The novels of André Baillon showed keen yet compassionate observation of life. Jean Tousseul was concerned with human suffering, whereas Charles Plisnier wrote powerfully analytical novels (*Mariages,* 1936; *Nothing to Chance*). Constant Malva was a radical socialist writer. Jean de Bosschère displayed a multifaceted talent. Other talented novelists were Marie Gevers and the visionary Franz Hellens. Paul Nougé and Marcel Lecomte were important figures in Belgian Surrealism, and Jean Ray was a pioneer of fantastic literature in Belgium. Somewhat later, Georges Simenon imbued the detective story with exceptional psychological penetration.

French-language novels

The poetry of this period was characterized by increased stylistic experiment and the development of a neoclassical poetry of great fluency. Henri Michaux, influenced by Surrealism, revolutionized the poetic language. Marcel Thiry and Edmond Vandercammen, whose lyrical style harmonized traditional and modern, and Géo Norge were chief exponents of an experimental use of words. Clément Pansaers was influenced by Dada. Another group was headed by Odilon-Jean Périer, an original poet of unusual clarity who was the leading light of his generation.

In drama Fernand Crommelynck wrote savage farces laced with poetry. Michel de Ghelderode astonished audiences by a love of anachronistic character and situation and puppetlike characters, and Herman Closson reinterpreted historical events and characters.

Developments after World War II. After World War II the novel became less regional than formerly, often set in foreign surroundings; generally it subordinated action to detailed psychological analysis of characters' reactions in somewhat contrived situations. Albert Ayguesparse, Hubert Juin, and Maud Frère were all ultimately concerned with the theme of man. Breadth of subject matter and meticulous style characterized the work of Suzanne Lilar and of Françoise Mallet-Joris, who won instant success with *Le Rempart des béguines* (1951; *Into the Labyrinth*). Gaston Compère and Dominique Rolin displayed both technical prowess and emotional intensity. Notable younger novelists included René Kalisky, Pierre Mertens, Conrad Detrez, and Marcel Moreau.

Postwar poetry often reverted to regular metrical forms in poems on the eternal themes of love and death. Christian Dotremont, Jean-Pierre Verheggen, Jacques Izoard, and Werner Lambersy investigated the frontiers of poetic language. Outstanding dramatists were Georges Sion, who wrote comedy, religious and historical plays, and translations from Shakespeare; Charles Bertin, whose *Prétendants* (1947), a modern version of the Ulysses story, was played abroad as *Love in a Labyrinth;* and Jean Mogin, whose plays, successful in both Paris and Brussels, explored "being" and "nonbeing." Later dramatists of note were Kalisky, Jean Sigrid, and Paul Willems.

Literary periodicals, which were usually organs for political or literary groups of writers, were influential, especially *Le Journal des Poètes,* which in 1952 founded a biennial international conference of poets and critics, held at Knokke–Het Zoute and later at Liège.

Walloon

EARLY WRITINGS

The origins of dialect literature in Wallonia are obscure. From the 9th to the 11th century Latin held sway in the abbeys, the only intellectual centres of the period. With the exception of the *Cantilène de Sainte Eulalie* (*c.* 900), the first vernacular writings date only from the middle of the 12th century. They are chiefly anonymous tracts, among which the *Poème moral,* consisting of nearly 4,000 alexandrines, stands out. During the next three centuries Walloon literature is marked by the importance of its local chronicles and certain aspects of its religious drama.

At the beginning of the 17th century, Wallonia—particularly the district of Liège—became conscious of the literary possibilities of dialect, and from then on the number of writings increased. An "Ode" in the Liège dialect appeared in 1620, and *pasquèyes* (*paskeyes, paskeilles*), poems describing local life and history, enjoyed a vogue.

THE 18TH AND 19TH CENTURIES

Use of the patois broadened in the 18th century. The success of comic opera at Liège resulted in several noteworthy librettos. *Li Voyadjue di Tchaudfontaine* (1757; "The Journey to Chaudfontaine"), *Li Lidjwès egagî* ("The Enlisted Liégeois"), and *Les Hypocondres* ("The Hypochondriacs") resulted in the formation of the Théâtre Liégeois. In lyric poetry the cramignon (a type of song for dancing) and the *Noëls* (Christmas carols and dialogue) adopted a genuine realism.

The number of Walloon poets and other dialect writers increased during the 19th century. Charles-Nicolas Simonon wrote the moving stanzas of "Li Côpareye" (the name of the clock of the cathedral of Saint-Lambert), François Bailleux his charming "Mareye," and the first great Walloon lyric poet, Nicolas Defrêcheux, his famous "Leyiz-m'plorer" (1854; "Let Me Weep"). The establish-

The Société Liégeoise de Littérature Wallonne

ment at Liège, in 1856, of the Société Liégeoise de Littérature Wallonne had considerable influence on both language and literature. The number of poems, songs, plays, and even translations into Walloon of such authors as La Fontaine, Ovid, and Horace increased.

Other parts of Belgium, apart from prolific Liège, still remained active centres of dialect writing. In the 19th century, Namur could boast especially of Charles Wérotte and Nicolas Bosret, poet of the touching song "Bia Bouquet." The works of Jean-Baptiste Descamps and others originated in Hainaut. Walloon Brabant was the home of a truculent Abbé Michel Renard.

By the end of the 19th century many writers working in Walloon dialects chose a rather doctrinaire Realism to depict workaday existence and remained somewhat hidebound by social conventions. Poets included Joseph Vrindts and, above all, Henri Simon, who sang of working peasantry. Successful playwrights included André Delchef and Édouard Remouchamps, whose vaudeville comedy in verse, *Tâtî l'pèriquî* (performed 1885; "Tati the Hairdresser"), married observation and technical dexterity.

THE 20TH CENTURY

Walloon literature explored new paths in the 20th century. Dialect studies were undertaken by an army of scholars, and the literary possibilities of the dialect were extended as a result of a standardizing of rules of spelling and grammar, as well as of attempts by Émile Lempereur and some other writers to renew the sources of inspiration. Alongside several veteran authors, such as the talented prose writer Joseph Calozet of Namur, the younger generations sought to achieve a strict unity of thought and technique. Among poets the following were to be noted: Franz Dewandelaer, Charles Geerts, Willy Bal, Henri Collette, Émile Gilliard, Jean Guillaume, Marcel Hicter, Albert Maquet, Georges Smal, and Jenny d'Invérno. Widely praised storytellers and novelists were Léon Mahy, Dieudonné Boverie, and Léon Maret, among many others. The dramatists included François Roland, Jules Evrard, Georges Charles, Charles-Henri Derache, François Masset, and J. Rathmès.

The work of dialect writers continued to be assisted by the Société de Littérature Wallonne, with its associations and publishing centres at Liège, Namur, Mons, La Louvière, Nivelles, and Brussels.

BIBLIOGRAPHY

General works: Among works useful to the general reader are JETHRO BITHELL, *Contemporary Belgian Literature* (1915); PAUL HAMÉLIUS, *Introduction à la littérature française et flamande de Belgique* (1921); SUZANNE LILAR, *The Belgian Theatre Since 1890,* 3rd ed. (1962); JAMES A. RUSSELL, *Romance and Realism* (1959); VERNON MALLINSON, *Modern Belgian Literature 1830–1960* (1966).

Flemish: JAN A. GORIS, *Belgian Letters,* 3rd ed. (1950); R.F. LISSENS, *De Vlaamse letterkunde van 1780 tot heden,* 4th ed. (1967, reissued 1974); THEODOOR WEEVERS, *Poetry of the Netherlands in Its European Context: 1170–1930* (1960); JEAN WEISGERBER, *Aspecten van de Vlaamse reman: 1927–1960,* 2nd ed. (1968); and REINDER P. MEIJER, *Literature of the Low Countries,* 2nd ed. (1978). Anthologies include JETHRO BITHELL (ed.), *Contemporary Flemish Poetry* (1917); JAN GRESHOFF (ed.), *Harvest of the Lowlands* (1945); and WILLIAM J. SMITH and JAMES S. HOLMES (eds.), *Dutch Interiors* (1984).

French: GEORGES DOUTREPONT, *Histoire illustrée de la littérature française en Belgique* (1939); CAMILLE HANLET, *Les Écrivains belges contemporains de langue française: 1800–1946,* 2 vol. (1946); ANDREW J. MATHEWS, *La Wallonie, 1886–1892: The Symbolist Movement in Belgium* (1947); GUSTAVE CHARLIER and JOSEPH HANSE (eds.), *Histoire illustrée des lettres françaises de Belgique* (1958); ADRIEN JANS (ed.), *Lettres vivantes: deux générations d'écrivains français en Belgique: 1945–75* (1975); JEAN MUNO, *La Littérature belge d'expression française,* 2nd ed., ed. by ROBERT BURNIAUX and ROBERT FRICKX (1980); and M. QUAGHEBEUR and ALBERTE SPINETTE (eds.), *Alphabet des lettres belges de langue française* (1982).

Walloon: M. WILMOTTE, *Le Wallon: histoire et littérature dès origines à la fin du XVIII^e siècle* (1893); LUCIEN and PAUL MARÉCHAL (eds.), *Anthologie des poètes wallons namurois* (1930); RITA LEJEUNE DEHOUSSE, *Histoire sommaire de la littérature wallonne* (1942); MAURICE PIRON (ed.), *Les Lettres wallonnes contemporaines* (1944), and *Poètes wallons d'aujourd'hui* (1961); ÉMILE LEMPEREUR, *Essai de catalogue d'une bibliothèque de littérature et de folklore wallons (1890–1947)* (1948); MAURICE DELBOUILLE (ed.), *Petite anthologie liégeoise: choix de textes wallons (XVII^e–XX^e siècle)* (1950); YANN LOVELOCK, *The Colour of the Weather* (1980), an introduction to Belgian dialect poetry; and RITA LEJEUNE and JACQUES STIENNON, *La Wallonie: le pays et les hommes: lettres, arts, culture,* 4 vol. (1977–81).

(R.F.Li./J.-E.-M.-G.D./T.J.H./V.N./J.P.Mo.)

Berlin

The capital and chief urban centre of Germany, Berlin lies at the heart of the North German Plain, athwart an east–west commercial and geographic axis that helped make it the capital of the kingdom of Prussia and then of a unified Germany. Berlin's former glory ended in 1945, but the city survived the destruction of World War II. It was rebuilt and came to show amazing economic growth. Germany's division after the war put Berlin entirely within the territory of the German Democratic Republic (East Germany). The city itself echoed the national partition—East Berlin being designated the capital of East Germany and West Berlin being a *Land,* or state (though not constitutionally incorporated as such), of the Federal Republic of Germany (West Germany), from the rest of which it was physically isolated. This unique status, later enforced and symbolized by the concrete barrier erected in 1961 and known as the Berlin Wall, made Berlin a continuous focus of crisis and confrontation between the Eastern and Western powers for some 40 years. The sudden downfall of the East German communist regime—and the accompanying opening of the wall—in late 1989 unexpectedly raised the prospect for Berlin's reinstatement as the all-German capital. That status was restored in 1990 under terms of the reunification treaty, and subsequently Berlin was designated a *Land* (state), one of the 16 constituting unified Germany. In addition to their political significance, these developments heralded the city's return to its historic position of prominence in European culture and commerce.

This article is divided into the following sections:

Physical and human geography

THE LANDSCAPE

The city site. Berlin is situated about 112 miles (180 kilometres) south of the Baltic Sea, 118 miles north of Czechoslovakia, and 110 miles east of the border that partitioned Germany from 1949 to 1990. It lies in the wide glacial valley of the Spree River, which runs through the centre of the city. The mean elevation of Berlin is 115 feet (35 metres) above sea level. The highest point near the centre of Berlin is the peak of the Kreuzberg, a hill that rises 216 feet above sea level northwest of the Tempelhof Airport.

Measuring about 23 miles from north to south and 28 miles from east to west, Berlin is by far the largest German city. It is built mainly on sandy glacial soil amid an extensive belt of forest-rimmed lakes, formed from the waters of the Dahme River to the southeast and the Havel to the west; indeed a third of the Greater Berlin area is still covered by sandy pine and birch woods, lakes, and beaches. "Devil's Mountain" (Teufelsberg), one of several hills constructed from the rubble left by World War II bombing, rises to 394 feet and has been turned into a winter sports area for skiing and sledding.

"Devil's Mountain"

Climate. Berlin lies where the influence of the Atlantic Ocean fades and where the climate of the continental plain begins. This climatic crossroads gives the city what the inhabitants call *Berliner Luft,* a type of air colder than but not as damp as that of western Germany and fresher, despite increasing industrial smog, because of quick replenishment from across lake-studded flatlands. The political events of 1989 and 1990 made possible the development of citywide pollution-control programs. Berlin's mean annual temperature is about 48° F (9° C), and mean temperatures range from 32° F (0° C) in winter to 65° F (18° C) in summer. The average precipitation is 23 inches (598 millimetres). About one-fifth to one-fourth of the total falls as snow.

The city layout. The original twin towns of Berlin and Kölln developed from the 13th century on an island of the Spree River (the site of Kölln) and a small portion of land on the north bank of the river facing the island (the site of Berlin). An independent Hanseatic city in the Middle Ages, Berlin, still a small town, became the capital of the electors of Brandenburg from the 15th century onward. The narrow streets and alleys of the medieval town remained largely intact until they were destroyed during the massive aerial bombing of World War II.

From the late 17th and early 18th centuries, when the electors of Brandenburg (also kings of Prussia from 1701) developed into powerful figures on the European political stage, the central quarter expanded and was embellished with broad avenues, handsome squares, and grandiose stone buildings. The central area acquired broad north–south avenues, such as Wilhelmstrasse and Friedrichstrasse, and also its characteristic east–west road axis. Supplementing this main axis are several great exit roads that now serve as major traffic arteries. The developing suburbs of the later 19th century grew up around these arteries and their subsidiary streets. Where destruction during World War II was massive, there has been large-scale construction of modern high-rise apartment and office buildings. Although the only major park near the centre of Berlin is the Tiergarten, which lies just west of the Brandenburg Gate, Berlin has always been a surprisingly green city, with luxuriant trees softening the effect of the stone apartment blocks in most streets. Water is even more prevalent. With the Spree River running through the city's centre, the broad belt of lakes spreading out east and west, canals running through much of the city, and several artificial ports, Berlin is very much a city on water.

Until the political eruption of 1989, the most notorious and pervasive feature of Berlin's topography was the Berlin Wall, originally erected by the East German government on Aug. 13, 1961, to stop free movement between East Berlin (and indeed East Germany) and West Berlin. The boundary between East and West Berlin and the boundary between West Berlin and East Germany, for a combined length of 103 miles, were closed until 1989 by a solid ring of barriers, consisting mostly of wall. There were several heavily guarded crossing points between West and East Berlin, the most famous of which was Checkpoint Charlie on Friedrichstrasse.

The city divided

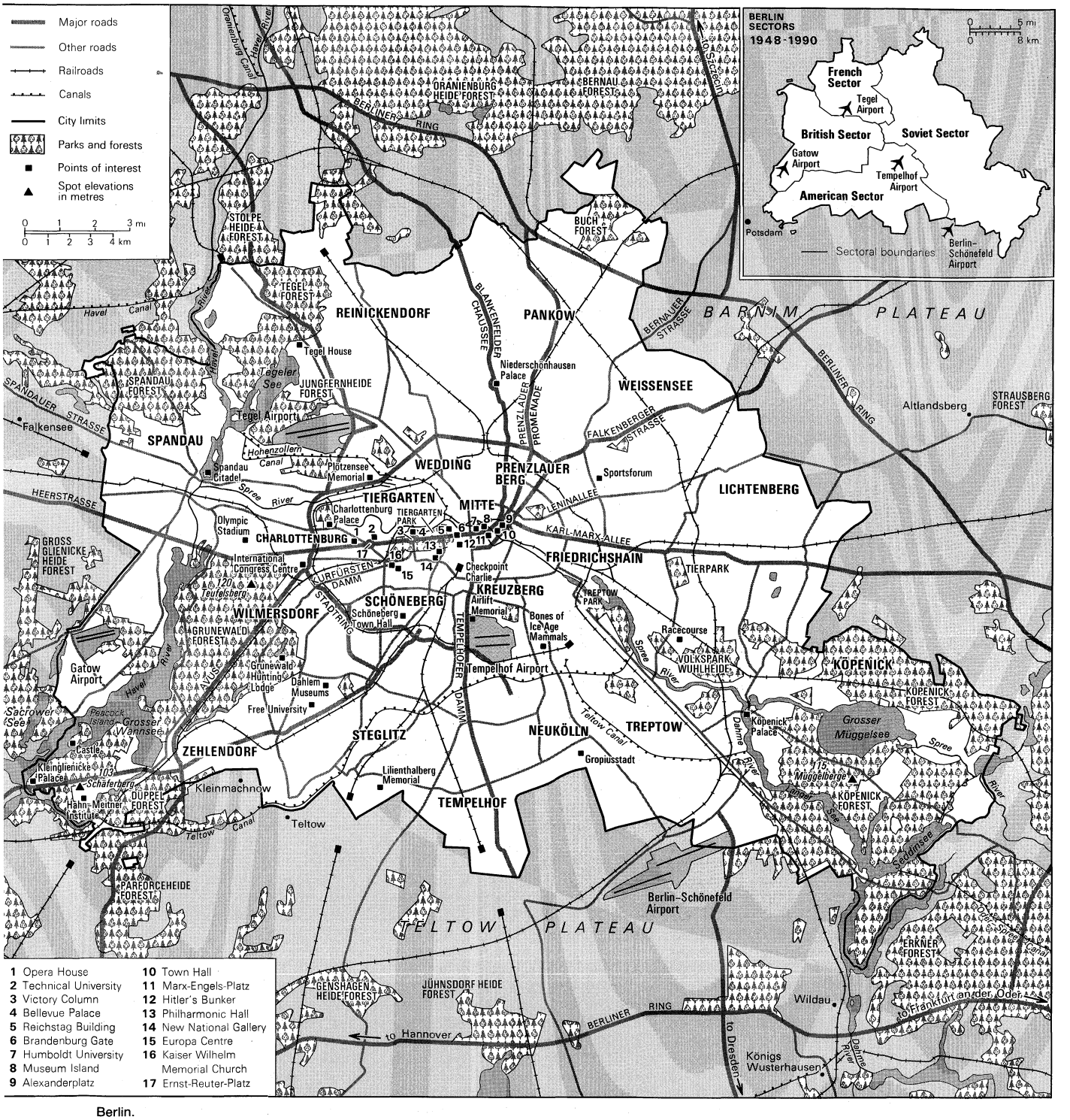
The landscape immediately next to the wall became one of the distinctive features of a divided Berlin. In East Berlin, a closed zone, guarded by watchtowers and armed patrols, flanked the wall, with the area closest to the wall constituting the so-called "death strip." On its western side, the wall was accessible from the start, and many visitors, Germans and non-Germans alike, endowed it with graffiti, ranging in quality from crude to artistically significant. Over the years, the original wall structure of steel-girded concrete was replaced incrementally (up to three times in some sections) with prefabricated concrete slabs set between steel pilings. After the opening of the wall on Nov. 9, 1989, new crossing points, such as that at Potsdamer Platz, were created by simply lifting out whole sections with cranes. Enterprising individuals from both sides who broke through with hammers and chisels—or who salvaged fragments for the souvenir market—received the nickname *Mauerspechte,* or "wall woodpeckers." Alongside the remnants of the wall lie open spaces and tracks of abandoned streetcar lines. In some places buildings had immediately adjoined the wall, and in the early days of division some people died attempting to jump to freedom from their upper floors. Today, crosses mark the places where these and other would-be refugees, numbering at least 80, lost their lives.

In terms of urban planning, the effects of the political and physical division of Berlin were many and pervasive, partly because the walled boundary created, in effect, an urban frontier immediately west of what had been the city's central administrative quarter, Berlin Mitte, which became part of East Berlin. West Berlin was thus forced to develop a new central area of its own around the Kurfürstendamm and the nearby Zoo railway station in the former suburb of Charlottenburg, separated from Berlin Mitte by the broad expanse of the Tiergarten. The area had been a distinctive commercial and entertainment district since the late 19th century, but rebuilding following extensive damage from World War II gave it a decidedly modern character.

Throughout the city, an effort to blend the modern with the traditional is evident. The most striking example in the western part is the Kaiser Wilhelm Memorial Church (Kaiser-Wilhelm-Gedächtniskirche), which incorporates the bell tower of the original 19th-century structure (ruined in World War II) into a dramatic glass and concrete church built in 1961. A landmark of more conventional historic preservation in the west is the restored Reichstag building. The decision to restore the former parliament house in the 1970s at a cost of 100 million deutsche marks was a controversial one—the building had been torched in the early days of Hitler's chancellorship (a key event in his assumption of dictatorial powers) and heavily bombarded during the final Soviet offensive in April 1945. Following the events of late 1989, the future role of the Reichstag in the new Berlin became a focus of national speculation. In early 1990, work was completed to make the building's plenary session chamber suitable for parliamentary use. Other buildings of note in the western part of the city include the Philharmonic Hall (Philharmonie; 1963) and the New National Gallery of modern art (Neue Nationalgalerie); the gallery was the last creation of the architect Ludwig Mies van der Rohe, who first worked in Berlin before World War I. The Hall for Chamber Music (Kammermusiksaal), a companion facility to the Philharmonic Hall, opened in 1987. The Charlottenburg Palace, dating from the late 17th century, is perhaps the city's most outstanding example of Baroque design.

Buildings in the east

The eastern part of the city has its own architectural symbol and war-memorial church—St. Nicholas Church (Nikolaikirche), dating from about 1200. Only the red-brick shell of Berlin's oldest building remained standing after a bombing attack during World War II, but restoration was completed in 1987, the 750th anniversary of Berlin's founding. The church, capped by two steeples brought from West Germany, serves as the centrepiece



Berlin.

of the old city enclave, the St. Nicholas Quarter (Nikolaiviertel), which includes replicas of townhouses from three centuries.

Two structures erected by the communist state dominate central Berlin—a 1,197-foot (365-metre) television tower (the Fernsehturm) and the Palace of the Republic (Palast der Republik), both adjacent to the Alexanderplatz. The tower, completed in 1969 to mark the 20th anniversary of the founding of East Germany, commands the Berlin skyline and has a revolving restaurant at the 800-foot level. The Palace of the Republic was opened in 1976 as the new seat of the East German parliament (Volkskammer), occupying the site of the former palace of the Prussian and German kings and kaisers. The decision to raze rather than to restore that badly damaged residence met with wide public disapproval. Also on the Alexanderplatz, which has prospects of once more becoming a crossroads of a united Berlin, rises the 39-story hotel Stadt Berlin, one of the city's tallest buildings.

In the same general area stand Berlin's oldest surviving church, St. Mary's Church (Marienkirche), and the Museum Island, on which are located the Old (Altes) and New (Neues) museums, the National Gallery (Nationalgalerie), the Bode Museum, and Pergamon Museum with its famous Greek altar of Zeus. Also in this area are the Town Hall, built of red brick; the State Council Building; and the rebuilt St. Hedwig's Cathedral, which dates from 1747 and which was the first Roman Catholic church to be built in Berlin after the Reformation.

The cultural district centred on Unter den Linden, the

broad avenue leading from near the Alexanderplatz to the Brandenburg Gate, also reflects the old and new. At its eastern end stands the Berlin cathedral (Berliner Dom), restoration of which did not begin until the late 1970s. For its entire length the avenue features modern hotels and shops and landmarks including the restored Arsenal (Zeughaus), New Guardhouse (Neue Wache), Berlin Palace (formerly the Crown Prince's Palace), Princesses' Palace (Prinzessinnenpalais), Opera House, National Library, Kaiser Wilhelm Palace, and Humboldt University. The Brandenburg Gate's sculptured chariot with four horses was restored in 1958; damage incurred during the celebrations marking New Year's Eve 1989 necessitated another restoration, to be completed by 1991, the 200th anniversary of the gate's construction.

South of Unter den Linden is the old Gendarme Market, renamed Academy Square, one of the finest architectural centres in Berlin, where restoration of the German cathedral and the Schauspielhaus, the former royal playhouse, was completed in time for the city's 750th anniversary. By 1990, restoration of the French cathedral, the dome of which matches the German one, also was nearing completion. Wilhelmstrasse, which runs north–south and which retains its name in the western part of the city, was once the site of Prussian and Reich government buildings. Removal of the wall west of the street exposed the remains of Hitler's bunker and the nearby Potsdamer Platz, once the city's busiest traffic hub. Before its sudden collapse, the East German government had bulldozed the bunker area and begun erecting a series of apartment buildings. Souvenir hunters and others reopened the underground complex, which has again become a focus of historical examination. Potsdamer Platz, a storied place in the memories of older Berliners, was the site where the wall was breached on Nov. 12, 1989. Since then, city planners, architects, and commercial enterprises have proposed various schemes for the area's revitalization.

Restored cathedrals

THE PEOPLE

Although the two sectors divided by the wall were approximately equal in area, the population of East Berlin numbered less than two-thirds that of West Berlin. Because the average age of West Berliners was higher than that of other West Germans, West Berlin encouraged the immigration of younger West German workers. With the end of partition, new or restored patterns of population growth quickly emerged. Some from the west sought cheaper housing in the east. Property values soared throughout the city. Many international firms sought Berlin locations. Some projections estimated that Berlin's total population would reach or exceed 5 million by the year 2000, as compared to a total of 3.4 million in mid-1990.

THE ECONOMY

Industry and trade. To a large extent, traditional economic activities, greatly reduced by World War II, have been revived throughout Greater Berlin. These include the production of textiles, metals, clothing, porcelain and china, bicycles, and machinery. Electronics became a principal postwar industry in both city halves. The cigarette and confectionery manufacturing that developed in West Berlin continues to flourish.

During the years of partition, West Berlin was a typical Western city, with plentiful consumer goods and luxury items. East Berlin authorities, on the other hand, tried to maintain morale through provision of cheaper basic elements in the cost of living, although salaries and pensions were lower than in the West. Central to the creation of a unified German state was the introduction of the deutsche mark, the West German currency, as the only legitimate monetary unit in the whole of the country on July 1, 1990.

Transportation. In urban transportation, the bus is now the mainstay, although East Berlin maintained streetcar service as well. Modern rapid-transit systems brought some postwar unity to the sphere of transport. The S-Bahn (Stadtbahn), an elevated railway system started in 1871, serves the whole city and metropolitan area. Through trains are few, with only one major station in the western part of the city. Construction of the subway, or Untergrundbahn (called the U-Bahn), began in 1897, and by World War II the city had one of the finest systems in Europe.

Rail systems

Air traffic has played an important role since 1945, particularly in the West in 1948, at the time of the Soviet blockade of the Western sectors. Tempelhof, the main field of the airlift, lost its traditional role as the centre of Berlin's air traffic during the 1970s. The airfield remained in use under the control of the U.S. Air Force, but outlying Tegel, created as an auxiliary airfield by the Western Allies during the Soviet blockade, became West Berlin's primary field, handling all civilian traffic. Berlin-Schönefeld in East Berlin was developed to handle the largest of modern aircraft. During the partition, only planes of the United States, Great Britain, and France were able to use the air corridors to the West. German reunification was expected to bring a general revision of Berlin's commercial air traffic pattern. Planners studied proposals to build a still larger third major airport outside the city.

The Reichsautobahn (National Expressway) in Berlin is part of a national superhighway network inaugurated before World War II. The system is linked with the Berliner Ring, a circle of autobahns around the city, putting Berlin in the centre of access spokes. Even before the events of November 1989, both Germanys had cooperated in developing road and rail traffic to and from Berlin. An autobahn connecting West Berlin with Hamburg was financed largely by West Germany.

ADMINISTRATION AND SOCIAL CONDITIONS

Government. In the East and West separately, district government after World War II continued much as before, with a chief burgomaster, or mayor, a city assembly, or parliament, and district mayors and councils, although a trend toward centralization ran strongly in such matters as citywide integration of primary and secondary education.

The parliamentary and municipal elections held throughout East Germany in the spring of 1990—the first free elections in this part of Germany since Hitler's rise to power in 1933—initiated changes that were intended to reunify both Germany and Berlin politically as well as practically. The central East German communist government, whose presence had long overshadowed East Berlin's city council, was quickly turned out of power by the freely elected delegates of the People's Chamber (Volkskammer). The Social Democratic Party, which had originated in what became East Germany but was subsequently ousted or absorbed by the communist regime, made the strongest showing in East Berlin. The mayor of West Berlin, also a Social Democrat, soon thereafter called for a joint city administration to seat itself in Berlin's traditional Town Hall in the eastern sector.

Joint city administration

In most respects, West Berlin was considered the 11th *Land* of West Germany. The city's delegates to the Bundestag in Bonn, however, lacked full voting powers, because the *Land* Berlin remained under occupation status, not constitutionally a part of West Germany. Formal reunification ended four-power jurisdiction in Berlin, and the city then became one of Germany's 16 *Länder*.

Overt postwar remilitarization by Germans was apparent only in East Berlin. The West German Bundeswehr (Federal Defense Force) was barred from West Berlin, and, although West Berliners could volunteer for military service, they were not eligible for the draft. For this reason, there was a steady movement of young West German males to West Berlin. East Germany openly drafted East Berliners, but by early 1990 this whole system had broken down.

In matters of justice, East Berlin was fully integrated within the overall East German court system. In West Berlin the Allies forbade the West German Constitutional Court from exercising jurisdiction. In practical law, however, West German justice and legislation applied in West Berlin under the federal constitution just as the East German system applied in East Berlin.

Health. Far-reaching health insurance is available throughout the city, which forms Germany's largest centre of medical activity. In the eastern part the Charité, which was founded as a royal hospital in 1710, was among those institutions hard hit by a shortage of medical professionals following the massive emigration in 1989. Authorities in

both sectors hoped that the opening of the wall would contribute to the stabilization and improvement of health services throughout the east. In the western part a modern Klinikum, or teaching hospital, has introduced new methods to medical practice.

Education. Berlin has traditionally played a leading role in German education. Communist ideology formed the basis of education in East Berlin until the fall of the regime made necessary an overhaul of both curriculum and the teacher-selection process. East Berlin had an earlier start in rebuilding its university system through physical control of the Frederick William University, now Humboldt University. Because of communist hegemony, noncommunist academics left East Berlin in 1948 and founded the Free University in West Berlin. From its inception, the Free University drew political activists from all over Germany. At first, many students participated in daring operations to bring refugees out of the East. By 1965 a new left had emerged whose militancy was carried into the streets, leading to clashes with the police. After a decline in student activism in subsequent decades, the events of 1989 and the movement for reunification brought a resurgence of political participation on West Berlin campuses. In East Berlin student dissidence had been limited until reform groups, often led by academics, emerged throughout the city in the late 1980s, notably under the protection of Protestant churches. It was these groups who first protested the falsified results of a municipal election in May 1989, triggering a wave of similar demonstrations elsewhere in East Germany. By November 1989, these efforts had coalesced into a centralized movement, which shortly before the opening of the wall brought more than one million demonstrators into the centre of East Berlin on a single afternoon.

Students and politics

With more than 100,000 students already enrolled in the three major universities in the west, the coming together of both the city halves and their respective educational systems posed new questions regarding acceptance of academic credentials and the management and accommodation of an east–west flow of students and teachers.

Several noteworthy libraries operate in the city. Two libraries of acclaim in the west are the American Memorial Library, built with U.S. aid after the Berlin blockade, and the Art Library, a state museum founded in 1867. The National Library in the east is also a major cultural and educational centre.

CULTURAL LIFE

When Berlin was a provincial capital, it only rarely rivaled cities such as London and Paris as a cultural magnet and, because of the regionalism of German life, seldom monopolized talented individuals as did other national capitals. From the 18th century, however, its cultural contribution became distinctive, and, if its 19th-century title "Athen an der Spree" ("Athens on the Spree") seems exaggerated, the contribution of Berliners to the arts and sciences, nevertheless, has been considerable. By 1750 the Prussian State Opera on Unter den Linden was rated among the finest opera houses in Europe, and the city's link with musical excellence was firmly established. Berlin never rivaled Vienna as a centre for German composers, although Felix Mendelssohn, a scion of one of the city's distinguished Jewish families, lived much of his life in Berlin.

Literary renaissance

The renaissance of German literature, dating from the late 18th century, found at least one of its homes in Berlin. Among the finest 19th-century writers associated with Berlin was Theodor Fontane, who wrote for the city's newspapers the *Kreuzzeitung* and *Vossiche Zeitung* and who perfected the German realistic novel. Other noted 19th-century writers who flourished in Berlin were the playwright Heinrich von Kleist and E.T.A. Hoffmann, who is best known for his fantastic short stories.

From the 18th century the Prussian state was served by a line of distinguished architects. Among these were Andreas Schlüter, who initiated the late German Baroque style; Georg Venzeslaus von Knobelsdorff, who built Sanssouci Palace outside Berlin for Frederick the Great; and Karl Friedrich Schinkel, who gave the centre of Berlin its characteristic Neoclassical grandeur. The court painter Antoine Pesne and the sculptor Christian Daniel Rauch, among lesser talents, lived in Berlin.

From the founding of the Frederick William (now Humboldt) University in 1810, Berlin became one of the foremost centres of German intellectual life. The city once rivaled Leipzig as a centre for German publishing, but its publishers' row was almost wiped out by wartime bombing. In the 19th century Berlin was also the centre for German newspaper publishing, and it still has more daily newspapers than most large cities.

The Weimar period

Berlin's role as a city of the imagination, of myth and symbol, reached its zenith not during the years of imperial splendour but during the era that followed, the period of the troubled Weimar Republic in the 1920s, when Berlin developed an extraordinary reputation for cultural brilliance and intellectual ferment. For many foreigners Berlin's more garish image of the period was epitomized by the British novelist Christopher Isherwood in such works as *Mr Norris Changes Trains* and *Goodbye to Berlin.* The clubs, cabarets, and other amusement enterprises that made Berlin notorious in the 1920s continued into the postwar period. The image of Weimar cultural brilliance was to be succeeded by another image of Berlin: the city of smoking and charred ruins when the Nazi regime crashed into defeat. Some of the loss brought by the Nazi interregnum could never be entirely regained—for instance, the cultural impact of the decimated Jewish community. Berlin has made strenuous efforts, however, to salvage and rebuild its cultural life.

The new Opera House (Deutsche Oper Berlin) was opened in West Berlin in 1961, and it quickly established a position as one of the leading opera houses of the Western world. The Opera House in East Berlin, burned down in 1843 and destroyed again by World War II bombs, was rebuilt in 1951; it is home to the long-established Deutsche Staatsoper (German National Opera). East Berlin's Comic Opera also gained fame. Classical music in general still finds a distinguished home in Berlin. Foremost among many notable musical ensembles is the world-famous Berlin Philharmonic Orchestra, founded in 1882 and reaching new heights in the postwar period under the leadership of the conductor Herbert von Karajan.

In the theatre of East Berlin, international acclaim was won by the Berliner Ensemble, founded by the playwright Bertolt Brecht after he returned to the city in 1948 following wartime exile. In the western part are three state theatres—the Schiller-Theater; its associated theatre workshop, the Schiller-Theater-Werkstatt; and the Schlosspark-Theater—as well as numerous privately operated theatres. In the early 20th century Berlin became the German centre of film production. From the 1960s a notable revival of filmmaking began in West Berlin. The Berlin Film Festival, founded in 1951, became one of the most important in the world.

Berlin is famous for its many excellent museums (see above *The city layout*). Because the prewar museum sites and parts of the old collections were located in what became East Berlin, a magnificent new museum complex, collectively called the Dahlem Museums, was built in the western district of Dahlem. The Egyptian Museum, in the western part, is also noted for its outstanding collection, which includes the celebrated bust of Queen Nefertiti. In the Reichstag building there is a permanent exhibition illustrating German history. The transfer of art works among formerly estranged institutions began in 1990 even before unification was formalized.

History

THE EARLY PERIOD

Founding

Origins. The name Berlin appears for the first time in recorded history in 1244, seven years after that of its sister town, Kölln, with which it later merged. Both were founded at about the beginning of the 13th century. In 1987, both East and West Berlin celebrated the city's 750th anniversary. Whatever the date of foundation, it is certain that the two towns were established for geographic and mercantile reasons, as they commanded a natural east–west trade route over the Spree River.

The way for their founding was opened by a Germanic resurgence in the area, which had been abandoned to the Slavs by the original Germanic tribes as they migrated westward. The Slavs were subdued by Albert I the Bear, a Saxon who crossed the Elbe River from the west. His successors took the title margrave of the mark (border territory) of Brandenburg. Berlin still retains as its symbol a defiant black bear standing on its hind legs.

The settlements of Spandau and Köpenick, now metropolitan districts, preceded the establishment of Berlin-Kölln. The Ascanians, followers of Albert I the Bear, established their fortress in 1160 at Spandau in the north where the Spree flows into the Havel River; by 1232 the fortress had earned the privileges of a town. Berlin-Kölln emerged between Spandau to the northwest and Köpenick to the southeast. By 1250 Berlin-Kölln dominated the mark of Brandenburg east to the Oder River, where a fort had been built in 1214, and in the 14th century it joined the Hanseatic League of northern German towns.

The Hohenzollerns. In 1411 the mark of Brandenburg came under the governorship of the Nürnberg feudal baron Frederick VI, beginning Berlin's association with the Hohenzollerns, who from the end of the 15th century as prince electors of Brandenburg established Berlin-Kölln as their capital and permanent residence.

The Thirty Years' War of 1618–48 laid a heavy financial burden on the city, and the population was reduced from 12,000 to 7,500. When Frederick William the Great Elector assumed power in 1640, he embarked on a building program, including fortifications that enabled him to expel Swedish invaders. His rule also marked the beginning of the development of canals, which by 1669 provided a direct link between Breslau (now Wrocław, Pol.) in the east and Hamburg and the open sea in the west. In 1709 the framework of Greater Berlin was laid when Berlin-Kölln and the newer towns of Friedrichswerder, Dorotheenstadt, and Friedrichstadt were put under a single magistrate. The population grew from 12,000 in 1670 to 61,000 in 1712, including 6,000 French Huguenot refugees.

The role of Frederick the Great

During the first half of the 18th century, Berlin expanded in all directions. Frederick II the Great adorned the city with a number of new buildings, and, soon after his death, the Brandenburg Gate was completed (1791). In 1810 the scholar Wilhelm von Humboldt founded what became East Berlin's Humboldt University. It early attracted such outstanding thinkers as the philosopher Georg Wilhelm Friedrich Hegel and the father of communism, Karl Marx. Berlin had its first popular uprising in 1830 when tailors' apprentices took to the streets over working conditions. The Revolution of 1848 produced 200 dead in a clash between soldiers and citizenry. By this time the city's population had risen to 415,000, from about 100,000 a century before. With the opening of the Berlin–Potsdam line in 1838, Berlin became the centre of an expanding rail network.

The period of railway growth was also that of Otto von Bismarck, whose successful military ventures as prime minister of Prussia paved the way for the creation of a united Germany. The Second Reich came into being when the king of Prussia was crowned Kaiser (Emperor) William I in 1871, at which time the population of Berlin, his capital, was 826,000.

THE 20TH CENTURY

The republic and Hitler. Three times in the 20th century, the date of November 9 has marked dramatic events in the history of Germany and Berlin. On that date in 1918, Berlin became the capital of the first German republic. Exactly 20 years later, Nazi storm troopers (Sturmabteilung) vandalized Jewish synagogues, shops, and other properties in the night of violence known as Kristallnacht (Night of Broken Glass). And on Nov. 9, 1989, East German authorities opened the barricade that had divided the city for 28 years.

The period 1918–33, in Berlin as elsewhere in Germany, was one of great disorder, runaway inflation, mass unemployment, and the rise to power of Adolf Hitler. Political and economic chaos spurred his ascendancy. By 1932 the unemployed in Berlin alone totaled 636,000, and on Jan. 31, 1933, Hitler became chancellor, his storm troopers marching through the Brandenburg Gate with massed flags and torches. He took absolute power in February and March of that year.

In 1936 Berlin was the scene of the most spectacular of modern Olympic Games, held in a specially built 100,000-seat stadium complex. Two years later, Kristallnacht, so called after the shattered windows of Jewish-owned buildings, was a landmark event of another kind. This rampage initiated a wave of persecution that in Berlin reduced a Jewish population of 170,000 to 5,000 by 1945. In May 1990, drawn by the momentous changes sweeping central and eastern Europe, the World Jewish Congress met in Germany for the first time in the organization's history, holding sessions in both East and West Berlin. Already in late 1989, both halves of the city had announced that they would vigorously pursue a bid to bring the Olympics back to Berlin in the year 2000.

Impact of World War II

Allied aerial bombing during World War II cost Berlin an estimated 52,000 dead. Another 100,000 civilians died in the battle for Berlin launched by the Soviet Army on April 16, 1945. Berlin's residential districts, factories, military facilities, streets, and cultural buildings were pounded into the sandy plain (one-sixth of all the wartime rubble in Germany lay in Berlin). The war left Berlin with a population that was 70 percent female, and in the aftermath of bombardment, the *Trümmerfrauen* ("rubble women") cleared the refuse. On April 30, 1945, Hitler committed suicide in his bunker below the Chancellery. On May 8 Soviet officials staged an elaborate German surrender ceremony in Berlin, rivaling that held by the Western Allies the day before in Reims, Fr. This divided ritual began the drama of the postwar city, the focal point of an East–West political struggle.

Berlin divided. On Oct. 1, 1920, the creation of a metropolitan Berlin took place through the fusing of 7 districts, 59 country communities, and 27 landed estates into a single association. Twenty resultant districts became integral parts of metropolitan Berlin but still remained largely autonomous. At the end of World War II the Soviet Union took eight of Berlin's districts as its sector of occupation: Berlin Mitte, Prenzlauer Berg, Friedrichshain, Treptow, Köpenick, Lichtenberg, Weissensee, and Pankow (Lichtenberg and Weissensee later being re-formed into the three districts of Hohenscheinhausen, Marzahn, and Hellersdorf). What was called the New West End, developed after old Berlin had outgrown its space, became West Berlin. The U.S. sector was formed by the southern districts of Zehlendorf, Steglitz, Tempelhof, Neukölln, Kreuzberg, and Schöneberg. The British sector embraced the central and western districts of the Tiergarten, as well as Wilmersdorf, Charlottenburg, and Spandau. The French were allotted the northern Wedding and Reinickendorf districts.

This apportioning was based on an agreement reached in London in 1944 by the United States, Britain, and the Soviet Union, acting on a British plan that divided Germany into occupation zones and Greater Berlin into sectors within, but not part of, the Soviet zone of occupation. A significant feature of the agreements concerning Berlin was the inability of the Western side to get a written Soviet guarantee of access.

Blockade and airlift

In March 1948 the Western powers decided to unite their zones of Germany into a single economic unit. In protest, the Soviet representative withdrew from the Allied Control Council. In June 1948, the West introduced a currency reform in the western sectors of occupied Germany, including West Berlin. The Soviet Union responded by launching a land blockade of West Berlin.

A great airlift broke this attempt to cut off the city from vital supplies, Western Allied planes hauling 1,831,200 tons of food, coal, and other necessities. The Soviets abandoned the blockade on May 12, 1949, but the Western Allies kept flying until September, building up a year's supply of essential goods. From time to time afterward access by road to West Berlin was denied for short periods as a form of harassment. In the meantime, on Nov. 30, 1948, a separate municipal government with its own chief burgomaster, or mayor, was set up in East Berlin, and this

completed the splitting of Berlin between East and West.

On June 17, 1953, some 50,000 workers, reacting to restrictive policies, rebelled in East Berlin. The uprising, which spread throughout East Germany, was crushed by Soviet intervention. In August 1961 West Berlin was physically separated from East Berlin by the erection of a wall topped with barbed wire and was isolated from the East German countryside elsewhere mainly by an encircling wire-mesh fence surmounted by watchtowers overlooking a strip of cleared ground. All of the 103-mile barrier was lighted at night. Patrols on foot and on small vehicles passed regularly. Trained dogs on long leashes guarded areas difficult to monitor by sight. Before its sudden transformation into an anachronism, the Berlin Wall had grown to about 12 feet (3.7 metres) in height, uniformly whitewashed on its eastern side, in contrast with the lively graffiti on the west. The barrier had reduced escape attempts to a trickle, until the changed political condition of Germany's neighbours to the east and south made indirect flight possible, notably through Hungary into Austria and then West Germany. In an effort to ease tensions, the United States, Britain, and France had joined the Soviet Union in signing a Berlin agreement in 1971. The working out of practical details for easier access to West Berlin from the West and for visits beyond the wall by West Germans and West Berliners was left to officials of the two German states. This resulted in an eastward flow by West Germans and West Berliners that numbered 9 million visits inside East Germany and another 27 million overland transit trips to and from West Berlin by car or train in 1988.

However, small steps toward normalization were not nearly enough, especially for the shut-in East German populace. In East Berlin and elsewhere, tens of thousands of demonstrators took to the streets in the autumn of 1989, demanding first the removal of the communist regime and then reunification with West Germany.

The opening of the wall

On November 9 the partition of Berlin dissolved with a speed few had thought possible. During the three days following, more than two million East Berliners and East Germans engulfed West Berlin, coming by car, on foot, and by subway and elevated trains. A provisional East German government then permitted West Berliners and West Germans to enter East Berlin freely, so that by New Year's Day 1990 east–west movement in Berlin had become an uncontrollable flood.

By mid-1990, whole sections of the wall, including that paralleling the mid-city Brandenburg Gate, had been removed. With crossing points proliferating, Berlin became in effect one city, although two governments still separately governed. Under the terms of the treaty enacting the reunification of the two German states, Berlin once again became the national capital.

BIBLIOGRAPHY

KARL BAEDEKER, *Berlin: Handbook for Travellers,* 7th ed. (1965), is one of the most famous guidebooks of its kind; although there are later editions of this guide, the cited edition contains particularly detailed information. See also *Berlin* (1983), a book from the Berlitz series of travel guides. On architecture, see WERNER HEGEMANN, *Das Steinerne Berlin,* 2nd ed. (1963, reprinted 1976), which concentrates on the rental buildings of the city; PAUL ORTWIN RAVE, *Berlin in der Geschichte seiner Bauten,* 3rd ed. (1976), mainly a pictorial work showing historical buildings still standing; RICHARD SCHNEIDER (ed.), *Berlin: Bauwerke de Neugotik* (1984), observations on modern architecture, illustrated; and STEPHEN D. HELMER, *Hitler's Berlin: The Speer Plan for Reshaping the Central City* (1985). Culture and art are discussed in WOLF VON ECKARDT and SANDER L. GILMAN, *Bertolt Brecht's Berlin* (1975), a survey of social and artistic developments in the 1920s; WOLFGANG SCHNEIDER, *Berlin: Eine Kulturgeschichte in Bildern und Dokumenten* (1981), a pictorial history of intellectual life in the city; and ROY F. ALLEN, *Literary Life in German Expressionism and the Berlin Circles* (1983), which explores literary trends.

For history, see OTTO FRIEDRICH GANDERT *et al., Heimatchronik Berlin* (1962), a scholarly treatment; HELLMUT KOTSCHENREUTHER, *Kleine Geschichte Berlins,* 3rd ed. (1976), a chronology; KARL SCHWARZ (ed.), *Berlin: Von der Residenzstadt zur Industriemetropole,* 3 vol. (1981), on economic development; and ALEXANDER REISSNER, *Berlin, 1675–1945* (1984), a brief overview. Postwar events are discussed in WALTER KRUMHOLZ, *Berlin ABC,* 2nd ed. (1968), a government-sponsored work specializing in events since 1945; and MARTIN J. HILLENBRAND, *The Future of Berlin* (1980), which discusses the city in view of its division and international status. Partition of the city is also the topic of JOHN W. KELLER, *Germany, the Wall and Berlin* (1964), charting the background of events; and RICHARD L. MERRITT and ANNA J. MERRITT (eds.), *Living with the Wall: West Berlin, 1961–1985* (1985). (H.J.Er./N.C.)

Beverage Production

The preparation of potable liquids from water and plant matter reflects a recurrent human predilection for variety, stimulation, and conviviality, as well as an ingenuity in exploiting resources toward those ends. This article treats beverages made by infusion, such as tea and coffee; alcoholic beverages, including beer and wine, made by fermentation; distilled spirits, requiring fermentation and distillation; and soft drinks, consisting of carbonated water flavoured with sweetened syrup.

For information on the cultivation of tea, coffee, and other crops important to beverage production, see the *Macropædia* article FARMING. For information on milk beverages, see FOOD PROCESSING. Information on fruit juices may be found in the entries on individual fruits in the *Micropædia.*

For coverage of related topics in the *Macropædia* and *Micropædia,* see the *Propædia,* section 731. (Ed.)

This article is divided into the following sections:

Tea

Tea is made from the young leaves and leaf buds of the tea plant, *Camellia sinensis*. Two principal varieties are used, the small-leaved China plant (*C. sinensis sinensis*) and the large-leaved Assam plant (*C. sinensis assamica*). Hybrids of these two varieties are also grown.

HISTORY OF THE TEA TRADE

According to legend tea has been known in China since about 2700 BC. For millennia it was a medicinal beverage obtained by boiling fresh leaves in water, but around the 3rd century AD it became a daily drink, and tea cultivation and processing began. The first published account of methods of planting, processing, and drinking came in AD 350. Around 800 the first seeds were brought to Japan, where cultivation became established by the 13th century. Chinese from Amoy brought tea cultivation to the island of Formosa (Taiwan) in 1810. Tea cultivation in Java began under the Dutch, who brought seeds from Japan in 1826 and seeds, workers, and implements from China in 1833.

Tea in India and Ceylon

In 1824 tea plants were discovered in the hills along the frontier between Burma and the Indian state of Assam. The British introduced tea culture into India in 1836 and into Ceylon (Sri Lanka) in 1867. At first they used seeds from China, but later seeds from the Assam plant were used.

The Dutch East India Company carried the first consignment of China tea to Europe in 1610. In 1669 the English East India Company brought China tea from ports in Java to the London market. Later, teas grown on British estates in India and Ceylon reached Mincing Lane, the centre of the tea trade in London.

By the late 19th and early 20th centuries, tea growing had spread to Russian Georgia, Sumatra, and Iran and extended to non-Asian countries such as Natal, Malawi, Uganda, Kenya, Congo, Tanzania, and Mozambique in Africa, to Argentina, Brazil, and Peru in South America, and to Queensland in Australia.

CLASSIFICATION OF TEAS

Teas are classified according to region of origin, as in China, Ceylon, Japanese, Indonesian, and African tea, or by smaller district, as in Darjeeling, Assam, and Nilgris from India, Uva and Dimbula from Sri Lanka, Keemun from Chi-men in China's Anhwei Province, and Enshu from Japan.

Teas are also classified by the grade, or size, of the processed leaf. Traditional sifting operations result in larger leafy grades and smaller broken grades. The leafy grades are: flowery pekoe (FP), orange pekoe (OP), pekoe (P), pekoe souchong (PS), and souchong (S). The broken grades are: broken orange pekoe (BOP), broken pekoe (BP), BOP fanning, fannings, and dust. Broken grades usually have substantial contributions from the more tender shoots, while leafy grades come mainly from the tougher and maturer leaves. In modern commercial grading, 95 to 100 percent of production belongs to broken grades, whereas earlier a substantial quantity of leafy grades was produced. This shift has been caused by an increased demand for teas of smaller particle size, which produce a quick, strong brew.

Black, green, and oolong teas

The most important classification is by the manufacturing process, resulting in the three categories of fermented (black), unfermented (green), and semifermented (oolong or pouchong). Green tea is usually produced from the China plant and is grown mostly in Japan, China, and to some extent Malaysia and Indonesia. The infused leaf is green, and the liquor is mild, pale green or lemon-yellow, and slightly bitter. Black tea, by far the most common type produced, is best made from Assam or hybrid plants. The infused leaf is bright red or copper coloured, and the liquor is bright red and slightly astringent but not bitter, bearing the characteristic aroma of tea. Oolong and pouchong teas are produced mostly in southern China and Taiwan from a special variety of the China plant. The liquor is pale or yellow in colour, as in green tea, and has a unique malty, or smoky, flavour.

PROCESSING THE LEAF

In tea manufacture, the leaf goes through some or all of the stages of withering, rolling, fermentation, and drying. The process has a twofold purpose: (1) to dry the leaf and (2) to allow the chemical constituents of the leaf to produce the quality peculiar to each type of tea.

The best-known constituent of tea is caffeine, which gives the beverage its stimulating character but contributes only a little to colour, flavour, and aroma. About 4 percent of the solids in fresh leaf is caffeine, and one teacup of the beverage contains 60 to 90 milligrams of caffeine. The most important chemicals in tea are the tannins, or polyphenols, which are colourless, bitter-tasting substances that give the drink its astringency. When acted upon by an enzyme called polyphenol oxidase, polyphenols acquire a reddish colour and form the flavouring compounds of the beverage. Certain volatile oils contribute to the aroma of tea, and also contributing to beverage quality are various sugars and amino acids.

Caffeine content

Only black tea goes through all stages of the manufacturing process. Green tea and oolong acquire their qualities through variations in the crucial fermentation stage.

Black tea. *Withering.* Plucking the leaf initiates the withering stage, in which the leaf becomes flaccid and loses water until, from a fresh moisture content of 70 to 80 percent by weight, it arrives at a withered content of 55 to 70 percent, depending upon the type of processing.

In the traditional process, fresh leaf is spread by hand in thin layers onto trays or sections of coarse fabric called tats. It is then allowed to wither for 18 to 20 hours, depending upon several factors that include the temperature and humidity of the air and the size and moisture content of the leaf. Withering in the open air has been replaced by various mechanized systems. In trough withering, air is forced through a thick layer of leaf on a mesh in a trough. In drum withering, rotating, perforated drums are used instead of troughs, and in tunnel withering, leaf is spread on tats carried by mobile trolleys and is subjected to hot-air blasts in a tunnel. Continuous withering machines move the leaf on conveyor belts and subject it to hot air in an enclosed chamber, discharging withered leaf while fresh leaf is simultaneously loaded.

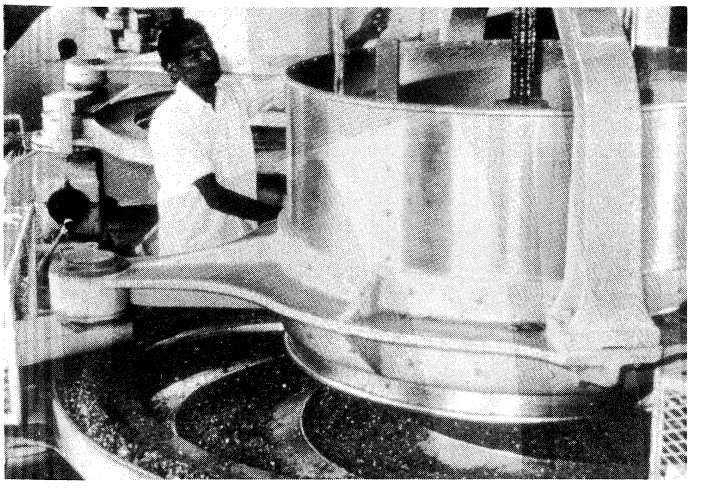
Importance of withering

Mechanized systems greatly reduce withering time, but they can also lower the quality of the final product by reducing the time for chemical withering, during which proteins and carbohydrates break down into simpler amino acids and sugars, and the concentration of caffeine and polyphenols increases.

Rolling. At this stage, the withered leaf is distorted, acquiring the distinctive twist of the finished tea leaf, and leaf cells are burst, resulting in the mixing of enzymes with polyphenols.

The traditional method is to roll bunches of leaves between the hands, or by hand on a table, until the leaf is twisted, evenly coated with juices, and finally broken into pieces. Rolling machines consist of a circular table fitted in the centre with a cone and across the surface with slats called battens. A jacket, or bottomless circular box with a pressure cap, stands atop the table. Table and jacket rotate eccentrically in opposite directions, and the leaf placed in the jacket is twisted and rolled over the cone and battens in a fashion similar to hand rolling. Lumps of rolled leaf are then broken up and sifted. The smaller leaf passing through the sieve—called the fines—is transferred to the fermentation room, and the remaining coarse leaf is rolled again.

In many countries, rolling the leaf has been abandoned in favour of distortion by a variety of machines. In the Legg cutter (actually a tobacco-cutting machine), the leaf is forced through an aperture and cut into strips. The crushing, tearing, and curling (CTC) machine consists of two serrated metal rollers, placed close together and revolving at unequal speeds, which cut, tear, and twist the leaf. The Rotorvane consists of a horizontal barrel with a feed hopper at one end and a perforated plate at the other. Forced through the barrel by a screw-type rotating shaft fitted with vanes at the centre, the leaf is distorted by resistor plates on the inner surface of the barrel and is cut at the end plate. The nontraditional distorting machines



A rolling machine twisting tea leaves, coating them with juices, and breaking them into pieces; Sri Lanka.
By courtesy of the Tea Council U.S.A.

can burst leaf cells so thoroughly that in many cases they render the withering stage unnecessary. However, unlike traditional rolling, they do not produce the larger leafy grades of tea.

Fermentation. Fermentation commences when leaf cells are broken during rolling and continues when the rolled leaf is spread on tables or perforated aluminum trays under controlled conditions of temperature, humidity, and aeration. The process actually is not fermentation at all but a series of chemical reactions. The most important is the oxidation by polyphenol oxidase of some polyphenols into compounds that combine with other polyphenols to form orange-red compounds called theaflavins. The theaflavins react with more units to form the thearubigins, which are responsible for the transformation of the leaf to a dark brown or coppery colour. The thearubigins also react with amino acids and sugars to form flavour compounds that may be partly lost if fermentation is prolonged. In general, theaflavin is associated with the brightness and brisk taste of brewed tea, while thearubigin is associated with strength and colour.

The formation of flavour and colour

In traditional processing, optimum fermentation is reached after two to four hours. This time can be halved in fermenting leaf broken by the Legg cutter, CTC machine, and Rotorvane. In skip fermentation, the leaf is spread in aluminum skips, or boxes, with screened bottoms. Larger boxes are used in trough fermentation, and in continuous fermentation the leaf is spread on trays on a conveyor system. In all of these fermentation systems the leaf is aerated by forced air (oxygen being necessary for the action of the enzymes), and it is brought by automated conveyor to the dryer.

Drying. At this stage, heat inactivates the polyphenol enzymes and dries the leaf to a moisture content of about 3 percent. It also caramelizes sugars, thereby adding flavours to the finished product, and imparts the black colour associated with fermented tea.

Traditionally, fermented leaf was dried on large pans or screens over fire, but since the late 19th century, heated forced air has been used. A mechanized drier consists of a large chamber into the bottom of which hot air is blown as the leaf is fed from the top on a series of descending conveyors. The dried leaf is then cooled quickly to prevent overdrying and loss of quality. Modern innovations on the drier are the hot-feed drier, where hot air is supplied separately to the feeder to arrest fermentation immediately as the leaf is fed, and the fluid-bed drier, where the leaf moves from one end of the chamber to the other over a perforated plate in a liquid fashion.

Prevention of fermentation

Green tea. In preparing unfermented tea, the oxidizing enzymes are killed by steamblasting the freshly plucked leaf in perforated drums or by roasting it in hot iron pans prior to rolling. The leaf is then subjected to further heating and rolling until it turns dark green and takes a bluish tint. The leaves are finally dried to a moisture content of 3 to 4 percent and are either crushed into small pieces or ground to a powder.

With the inactivation of polyphenol oxidase, the polyphenols are not oxidized and therefore remain colourless, allowing the processed leaf to remain green. The absence of theaflavins and thearubigins in the finished leaf also gives the beverage a weaker flavour than black tea.

Oolong tea. After a brief withering stage, the leaf is lightly rolled by hand until it becomes red and fragrant. For oolong it is then fermented for about one-half, and for pouchong for one-quarter, of the time allowed for black tea. Fermentation is stopped by heating in iron pans, and the leaf is subjected to more rolling and heating until it is dried.

PACKAGING

Sorting and grading. The first step in packaging tea is grading it by particle size, shape, and cleanliness. This is carried out on mechanical sieves or sifters fitted with meshes of appropriate size. With small-sized teas in demand, some processed teas are broken or cut again at this stage to get a higher proportion of broken grades. Undesirable particles, such as pieces of tough stalk and fibre, are removed by hand or by mechanical extractor. Winnowing by air removes dust, fibres, and fluff.

Packing. Teas are packed in airtight containers in order to prevent absorption of moisture, which is the principal cause of loss of flavour during storage. Packing chests are usually constructed of plywood, lined with aluminum foil and paper, and sealed with the same material. Also used are corrugated cardboard boxes lined with aluminum foil and paper or paper sacks lined with plastic.

Preventing exposure to moisture

Blended teas are sold to consumers as loose tea, which is packed in corrugated paper cartons lined with aluminum foil, in metal tins, and in fancy packs such as metallized plastic sachets, or they are sold in tea bags made of special porous paper. Tea bags are mainly packed with broken-grade teas.

Instant tea. Instant teas are produced from black tea by extracting the liquor from processed leaves, tea wastes, or undried fermented leaves, concentrating the extract under low pressure, and drying the concentrate to a powder by freeze-drying, spray-drying, or vacuum-drying. Low temperatures are used to minimize loss of flavour and aroma. Instant green teas are produced by similar methods, but hot water is used to extract liquor from powdered leaves. Because all instant teas absorb moisture, they are stored in airtight containers or bottles.

PREPARING THE BEVERAGE

Blending. Tea sold to the consumer is a blend of as many as 20 to 40 teas of different characteristics, from a variety of estates, and from more than one country. Price is an important factor, with cheap teas (called fillers) used to round off a blend and balance cost. Blends are often designed to be of good average character without outstanding quality, but distinctive blends—for example, with a flavour of seasonal Ceylon tea or the pungency and strength of Assam tea—are also made.

Brewing. A tea infusion is best made by pouring water just brought to the boil over dry tea in a warm teapot and steeping it from three to five minutes. The liquor is separated from the spent leaves and may be flavoured with milk, sugar, or lemon.

Tasting. Professional tasters, sampling tea for the trade, taste but do not consume a light brew in which the liquor is separated from the leaf after five to six minutes. The appearance of both the dry and infused leaf is observed, and the aroma of vapour, colour of liquor, and creaming action (formation of solids when cooled) are assessed. Finally the liquor is taken into the mouth with a sucking noise, swirled around the tongue, brought into contact with the palate, cheek, and gums, and then drawn to the back of the mouth and up to the olfactory nerve in the nose before being expectorated. The liquor is thus felt, tasted, and smelled. Tasters have a large glossary of terms for the evaluation of tea, but the less-demanding consumer drinks it as a thirst quencher and stimulant and for its distinctive sour-harsh taste. (S.Si.)

Professional tea tasting

Coffee

Coffee, a beverage brewed from the roasted and ground seeds of the tropical evergreen coffee plant of African origin, is consumed either hot or cold by about one-third of the people in the world, in amounts larger than those of any other drink. Its popularity can be attributed to its invigorating effect, which is produced by caffeine, an alkaloid present in green coffee in amounts between 0.8 and 1.5 percent for the Arabica varieties and 1.6 to 2.5 percent for Robusta.

Two principal coffee plants

Two species of the coffee plant, *Coffea arabica* and *C. canephora,* supply almost all of the world's consumption. Arabica coffee, which is divided between Brazilians and milds, is considered to brew a more flavourful and aromatic beverage than Robusta, the main variety of *C. canephora.* Arabicas are grown in Central and South America, the Caribbean, and Indonesia, while Robustas are grown mainly in Africa.

HISTORY

Wild coffee plants, probably from Kefa (Kaffa), Ethiopia, were taken to southern Arabia and placed under cultivation in the 15th century. One of many legends about the discovery of coffee is that of Kaldi, an Arab goatherd, who was puzzled by the queer antics of his flock. About AD 850, Kaldi supposedly sampled the berries of the evergreen bush on which the goats were feeding and, on experiencing a sense of exhilaration, proclaimed his discovery to the world.

Whatever its historical origin, the stimulating effect of coffee undoubtedly made it popular, especially in connection with the long religious service of the Muslims. The orthodox priesthood pronounced it intoxicating and therefore prohibited by the Qur'ān, but despite the threat of severe penalties, coffee drinking spread rapidly among Arabs and their neighbours.

During the 16th and 17th centuries, coffee was introduced into one European country after another; many accounts are recorded of its prohibition or approval as a religious, political, and medical potion. Coffee gained popularity as a beverage in the London coffeehouses, which became centres of political, social, literary, and eventually business influence. The first coffeehouse in London was established about 1652. In Europe, too, the coffeehouse flourished later in the 17th century. In such North American cities as Boston, New York City, and Philadelphia, coffeehouses became popular beginning in the late 1600s.

Until the close of the 17th century, the world's limited supply of coffee was obtained almost entirely from the province of Yemen in southern Arabia. But, with the increasing popularity of the beverage, the propagation of the plant spread rapidly to Java and other islands of the Indonesian archipelago in the 17th century and to the Americas in the 18th century. Coffee cultivation was started in the Hawaiian Islands in 1825.

By the 20th century the greatest concentration of production was centred in the Western Hemisphere—particularly Brazil. In the late 19th and early 20th centuries, industrial roasting and grinding machines came into use, vacuum-sealed containers were invented for ground roasts, and decaffeination methods for green coffee beans were developed. After 1950 the production of instant coffee was perfected. The popularity of instant coffee led to increased production of the cheaper Robusta beans in Africa.

(T.C.W./Ed.)

PROCESSING GREEN COFFEE

Hulling. The ripened fruits of the coffee shrubs, known as coffee cherries, are processed by disengaging the coffee seeds from their coverings and from the pulp and by drying the seeds from an original moisture content of 65–70 percent water by weight to 12–13 percent. Two different techniques are used: a wet process (used mainly for the mild Arabica coffees) and a dry process (used for Brazilians and Robustas).

Two methods of hulling

The wet process. First the skin and pulp of the fresh fruit is removed by a pulping machine, which consists of a rotating drum or disk that presses the fruit against a sharp-edged or slotted plate, disengaging the pulp from the seed. Pulp still clings to the coffee, however, as a thin, mucilaginous layer. This is eliminated by fermentation, actually a form of digestion in which naturally occurring pectic enzymes decompose the pulp while the wetted seeds are held in tanks for one to three days. Washing clears all remaining traces of pulp from the coffee seeds, which are then dried either by exposure to sunlight on concrete terraces or by passing through hot-air driers. The dry skin around the seed, called the parchment, is then mechanically removed, sometimes with polishing.

The dry process. In this process, the fruits are immediately placed to dry either in sunlight or in hot-air driers. Although mechanical drying is replacing the labour- and time-consuming sun drying, more time and equipment are required than in drying pulped seeds in the wet process. When the fruits have been dried to a water content of 12 percent, they are mechanically hulled to free the seeds from their coverings.

Grading and storage. The practice of grading coffee gives sellers and buyers a guarantee concerning the origin, nature, and quality of the product to aid their negotiations. Each country has a certain number of defined types and grades, but there are no international standards outside the contract market.

The prolonged storage of coffee in the producing countries presents problems, especially in the warm and humid coastal regions, where molds and parasites may develop and cause damage; for this reason coffee from these areas is exported as quickly as possible. In moderate climates, the conservation of dry lots does not pose a problem as long as they are stocked in well-ventilated places.

PROCESSING THE BEAN

Decaffeination. Caffeine can be removed from the green coffee by a variety of methods. In the most common, solvent extraction, the beans are steamed to raise the moisture content and bring the dissolved caffeine to the surface of the beans. They are then washed by an organic solvent such as methylene chloride, which extracts the caffeine. The solution is removed by steam, and the beans are dried.

Roasting. The aromatic and gustatory qualities of coffee are developed by the high temperatures to which they are subjected during roasting or broiling.

Temperatures are raised progressively to about 220°–230° C (430°–440° F). This releases steam, carbon dioxide, carbon monoxide, and other volatiles from the beans, resulting in a loss of weight between 14 and 23 percent. Internal pressure of gas expands the coffee beans by 30 to 100 percent. The beans become a deep, rich brown, and their texture becomes porous and crumbly under pressure. But the most important phenomenon of roasting is the appearance of the characteristic aroma of coffee, which arises from very complex chemical transformations within the bean. Roasting too long can destroy volatile flavour and aroma compounds. For this reason, Robusta beans are often over-roasted (as in the dark French and Italian roasts) to rid the coffee of its natural harshness.

Effects of over-roasting

In the oldest method of roasting, a metal cylinder, or sphere, containing the coffee is rotated above a source of heat such as charcoal, gas, or electricity. In modern roasters, hot air is propelled by a blower into a rotating metal cylinder containing the coffee. The tumbling action of rotation ensures that all beans are roasted evenly.

Regardless of the method used, the coffee, after leaving the industrial roasters, is rapidly cooled in a vat, where it is stirred and subjected to cold air propelled by a blower. Good-quality coffees are then sorted by electronic sorters to eliminate those seeds, either too light or too dark, that roasted badly and whose presence depreciates the quality.

Grinding. Some coffees are left as whole beans to be ground at the time of purchase or by the consumer at home. But a large part of the coffee is ground, or milled, by the manufacturer immediately after roasting. In most modern roasting plants, grinding is accomplished by feeding the coffee through a series of serrated or scored rollers, set at progressively smaller gaps, that first crack the beans and then cut them to the desired particle size.

The degree of fineness is important. If a coffee is too coarse, water filters through too fast to pick up flavour; if it is too fine, water filters through too slowly and retains particles that deposit at the bottom of the cup.

PACKAGING AND BREWING

Packaging. Effective packaging prevents air and moisture from reaching the coffee. Ground coffee alters rapidly and loses its aromatic qualities within a few days if it is not put into hermetically sealed containers immediately. The air, especially in humid atmospheres, causes rancidity through oxidation of fatty components. Modern packaging materials, plastic films like polyethylene and complexes of aluminum and cellulose, are capable of conserving the quality of coffee for a time. But the most satisfactory solution to the problem is packing under vacuum or in an inert gas, in rigorously impervious containers.

Brewing. There are several methods of extracting flavour and aroma from ground coffee. In steeping or boiling, pulverized coffee is measured into hot water, which is set or boiled before being poured off the grounds. In percolation, water is brought to the boil in an urn and fed up a tube to a basket holding the coffee. After filtering through the coffee, the water drips back to the urn, where it is forced back up the tube and recirculated until the brew reaches the desired strength. In the filter, or drip, method, hot water is slowly filtered through the coffee and dripped into a receptacle; it is not recirculated. The espresso machine forces boiled water under pressure through finely ground coffee; because the water has only brief contact with the grounds, it extracts a highly flavoured brew with little bitterness.

Caffeine content

Caffeine content varies with the variety of bean and method of brewing. One serving (five fluid ounces) of Arabica instant coffee contains about 70 milligrams of caffeine, while a serving of brewed Robusta may contain 200 milligrams.

Instant coffee. In the manufacture of instant coffee (called soluble coffee in the industry), a liquid concentration of coffee prepared with hot water is dehydrated. This can be done by spray drying in hot air, by drying under vacuum, or by lyophilization (freeze drying). The operations are complex and methods vary among manufacturers. The resulting soluble powder, on the addition of hot water, forms reconstituted coffee. The average yield is 25 to 30 percent by weight of the ground coffee. Because it picks up moisture readily, instant coffee needs special vacuum packages. (R.C./Ed.)

Beer

Beer is an alcoholic beverage produced by extracting raw materials with water, boiling (usually with hops), and fermenting. In some countries beer is defined by law—as in Germany, where the standard ingredients, besides water, are germinated barley, hops, and yeast.

HISTORY OF BREWING

Before 6000 BC, beer was made from barley in Sumeria and Babylonia. Reliefs on Egyptian tombs dating from 2400 BC show that barley or partly germinated barley was crushed, mixed with water, and dried into cakes. When broken up and mixed with water, the cakes gave an extract that was fermented by microorganisms accumulated on the surfaces of fermenting vessels.

The basic techniques of brewing came to Europe from the Middle East. The Roman historians Pliny (in the 1st century BC) and Tacitus (in the 1st century AD) reported that Saxons, Celts, and Nordic and Germanic tribes drank ale. In fact, many of the English terms used in brewing (malt, mash, wort, ale) are Anglo-Saxon in origin. During the Middle Ages, the monastic orders preserved brewing as a craft. Hops were in use in Germany in the 11th century, and in the 15th century they were introduced into Britain from Holland. In 1420 beer was made in Germany by a bottom fermentation process; before that, yeast rose to the top of the fermenting product and was allowed to overflow or was manually skimmed. Brewing was a winter occupation, and ice was used to keep beer cool during the summer months. Such beer came to be called lager (from German *lagern,* "to store"). The term lager is still used to denote beer produced from bottom-fermenting yeast, and the term ale is now used for top-fermented British types of beer.

Lagers and ales

The Industrial Revolution brought the mechanization of brewing. Better control over the process, with the use of the thermometer and saccharometer, was developed in Britain and transferred to the Continent, where the development of ice-making and refrigeration equipment in the late 19th century enabled lager beers to be brewed in summer. In the 1860s the French chemist Louis Pasteur established many of the microbiological practices still used in brewing. The Danish botanist Emile Hansen devised methods for growing yeasts in culture free of other yeasts and bacteria. This pure-culture technology was quickly taken up by continental lager brewers but not until the 20th century by the ale brewers of Britain. Meanwhile, German-style bottom-fermented lagers fermented by pure yeast cultures became dominant in the Americas.

Brewing in the 20th century is a large-scale industry. Modern breweries use stainless-steel equipment and computer-controlled automated operations, and they package beer in metal casks, glass bottles, aluminum cans, and plastic containers. Beers are now exported worldwide and are produced under license in foreign countries.

TYPES OF BEER

Beverages similar to beer are produced in Japan (sake, from rice) and Mexico (pulque, from cactus). In Nigeria and South Africa, malted sorghum is used to produce *burukutu* and Kaffir beers.

In the West, the properties of the water used for brewing, the types of malt, the brewing practices, and the yeast strains have contributed to traditional distinctions between beers. Early British beers were made from successive extracts of a single batch of brown malt. The first and strongest extract gave the best quality beer, called strong beer, and a third extract yielded the poorest quality beer, called small beer. In the 18th century London brewers departed from this practice and produced porter. Made from a mixture of malt extracts, porter was a strong, dark-coloured, highly-hopped beer consumed by the market porters in London. Brewers in Burton upon Trent, using the famous hard waters of that region and pale malts roasted in coke-fired kilns, created pale ales, also called best bitter. Pale ale is less strong, less bitter, paler in colour, and clearer than porter. Mild ales—weaker, darker, and sweeter than bitter—are a common variation; more colour is obtained by special malts, roasted barley, or caramels, less hops are used, and cane sugar is added to impart sweetness and aid maturation. Stouts are stronger versions of mild ale; some, such as milk stouts, contain lactose (milk sugar) as a sweetener.

Pale ale, mild ale, and stout

Bottom-fermented lagers have their origins in different regions of continental Europe. Brewers in Plzeň (now in Czechoslovakia) used local soft waters to produce the famous Pilsner beer, which became the standard for highly hopped, pale-coloured, dry lagers. *Dortmunder* is a pale lager of Germany, while Munich has become associated with dark, strong, slightly sweet beers with less hop character. The dark colour comes from highly roasted malt, and other characteristic flavours arise during the decoction mashing process. Bock is an even stronger, heavier Munich-type beer that is brewed in winter for consumption in the spring. *Märzbier* ("March beer") is a lighter brew produced in the spring. While all German lagers are made with malted barley, a special brew called weiss beer (*Weissbier;* "white beer") is made from malted wheat. In other countries such as Denmark and The Netherlands, other cereals are used in lighter coloured larger beers.

The 20th century has seen the erosion of traditional distinctions based on place of manufacture, raw materials, and brewing methods. This has caused a reaction among a small body of consumers. In Britain it has encouraged support for smaller, traditional ale breweries. In the United States a growing number of "microbreweries" make beers with more flavour and colour.

The strength of beer may be measured by the percentage

by volume of ethyl alcohol. Strong beers are in excess of 4 percent, and so-called barley wines 8 to 10 percent. Diet beers or light beers are fully fermented, low-carbohydrate beers in which enzymes are used to convert normally unfermentable (and high-calorie) carbohydrates to fermentable form. In low-alcohol beers (0.5 to 2.0 percent alcohol) and "alcohol-free" beers (less than 0.1 percent alcohol), alcohol is removed after fermentation by low-temperature vacuum evaporation or by membrane filtration.

THE BREWING PROCESS

Beer production involves malting, milling, mashing, extract separation, hop addition and boiling, removal of hops and precipitates, cooling and aeration, fermentation, separation of yeast from young beer, aging, maturing, and packaging. The object of the entire process is to convert grain starches to sugar, extract it with water, and then ferment it with yeast to produce the alcoholic, lightly carbonated beverage.

Malting. Malting modifies barley to green malt, which can then be preserved by drying. The process involves steeping and aerating the barley, allowing it to germinate, and then drying and curing the malt.

Barley. In order to be fermented by yeast, the food reserve of barley, starch, must be converted by enzymes into simple sugars. Two enzymes, α- and β-amylases, carry out the conversion. The latter is present in barley, but the former is only made during germination of the grain. Specially bred strains of barley (generally low in nitrogen content) are used for malting. Other important characteristics are yield, even germination, ability to produce enzymes, and a highly extractable malt.

Steeping. Malting begins by immersing barley harvested at less than 12 percent moisture in water at 12° to 15° C (55° to 60° F) for 40 to 50 hours. During this steeping period, the barley may be drained and given air rests, or the steep may be forcibly aerated. As the grain imbibes water, its volume increases by about 25 percent, and its moisture content reaches about 45 percent. A white root sheath, called a chit, breaks through the husk, and the chitted barley is then removed from the steep for germination.

Germination. Activated by water and oxygen, the root embryo of the barleycorn secretes a plant hormone called gibberellic acid, which initiates the synthesis of α-amylase.

Conversion of starch into sugar

The α- and β-amylases then convert the starch molecules of the corn into sugars that the embryo can use as food. Other enzymes, such as the proteases and β-glucanases, attack the cell walls around the starch grains, converting insoluble proteins and complex sugars (called glucans) into soluble amino acids and glucose. These enzymatic reactions are called modification. The more germination proceeds, the greater the modification. Overmodification leads to malting loss, in which rootlet growth and plant respiration reduce the weight of the grain.

In traditional malting, the steeped barley was placed in heaps called couches and, after 24 hours, spread on a floor to permit germination. Because respiration of the grain causes oxygen to be taken up and carbon dioxide and heat to be produced, control of aeration, ventilation, and temperature was achieved by manually turning the grain. Large-scale floor maltings with mechanical turners were introduced and then replaced by pneumatic maltings, in which germination occurred in boxes with the bed automatically turned, aerated, and ventilated with forced air. In some malting operations gibberellic acid is sprayed onto the barley to speed germination, and bromates are used to suppress rootlet growth and malting loss. Although less-modified malts are traditionally used in lagers, and well-modified malts in ales, it is now usual to produce well-modified malts regardless of whether lager or ale is to be made.

Kilning. Green malt is dried to remove most of the moisture, leaving 5 percent in lager and 2 percent in traditional ale malts. This process arrests enzyme activity but leaves 40 to 60 percent in an active state. Curing at higher temperatures promotes a reaction between amino acids and sugars to form melanoidins, which give both colour and flavour to malt.

In the first stage of kilning a high flow of dry air at 50° C (120° F) for lager malt and 65° C (150° F) for ale malt is maintained through a bed of green malt. This lowers the moisture content from 45 to 25 percent. A second stage of drying removes more firmly bound water, the temperature rising to 70°–75° C (160°–170° F) and the moisture content falling to 12 percent. In the final curing stage, the temperature is raised to 75°–90° C (170°–195° F) for lager and 90°–105° C (195°–220° F) for ale. The finished malt is then cooled and screened to remove rootlets.

Roasting special malts

Special malts are made by wetting and heating green malt in closed drums at high temperatures. Made in this way are crystal (caramel), chocolate (black), and amber malts; used in small and varying proportions (2 to 3 percent of brewing malt), they introduce considerable variations in colour and flavour to finished beers. Chocolate malt and roasted ungerminated barley are used at a high proportion (25 percent) to make stouts and porters. The use of unmalted cereals has also become common, because they are less expensive sources of starch and can be used to dilute malt colour and flavour, thereby yielding fresher, lighter beers.

Modernization. Modern maltings can produce malt in four to five days, and technological improvements give precise control over temperature, humidity, and use of heat. Tower maltings have been developed with an uppermost floor for steeping and lower floors for germination and kilning, producing a compact, semicontinuous operation that is also fully automated.

Mashing. After kilning, the malt is mixed with water, and the enzymatic conversion of starch into fermentable sugar is completed.

Milling. For efficient extraction with water, malt must be milled. Early milling processes used stones driven manually or by water or animal power, but modern brewing uses mechanically driven roller mills. The design of the mill and the gap between the rolls are important in obtaining the correct reduction in size of the malt. The object is to retain the husk relatively intact while breaking up the brittle, modified starch into particles.

Mixing the mash. The milled malt, called grist, is mixed with water, providing conditions in which starch, other molecules, and enzymes are dissolved and rapid enzyme action takes place. The solute-rich liquid produced in mashing is called the wort. Traditionally, mashing may be of one of two distinct types. The simplest process, infusion mashing, uses a well-modified malt, two to three volumes of water per volume of grist, a single vessel (called a mash tun), and a single temperature in the range of 62° to 67° C (145°to 150° F). With well-modified malt, breakdown of proteins and glucans has already occurred at the malting stage, and at 65° C the starch readily gelatinizes and the amylases become very active. Less well-modified malt, however, benefits from a period of mashing at lower temperatures to permit the breakdown of proteins and glucans. This requires some form of temperature programming, which is achieved by decoction mashing. After grist is mashed in at 35° to 40° C (95° to 105° F), a proportion is removed, boiled, and added back. Mashing with two or three of these decoctions raises the temperature in stages to 65° C. The decoction process, traditional in lager brewing, uses four to six volumes of water per volume of grist and requires a second vessel called the mash cooker.

Two mashing methods

Other sources of starch that gelatinize at 55° to 65° C can be mashed along with malt. Wheat flour and corn (maize) flakes may be added directly to the mash, whereas corn grits and rice grits must first be boiled in order to gelatinize. Their use requires a third vessel, the cereal cooker.

Modern mashing systems use mixed grists and mash mixers, which are efficiently stirred and temperature-programmed mashing vessels. Enzymes of bacterial and fungal origin may be added as aids. Ale and lager are mashed in the same equipment, but they require different temperature programs and grist composition. Modern breweries often practice high-gravity brewing, in which highly concentrated worts are made, fermented, and then diluted, allowing more beer to be brewed on the same equipment.

Separating the wort. The mash tun used in infusion mashing is fitted with a false base containing precisely-

machined slots through which the husk, preserved during milling, cannot pass. The trapped husk thus forms a filter bed that removes solids from the wort as it is drained, leaving a residue of spent grains. Wort separation takes four to 16 hours. For thorough extraction, the solids are sprayed, or sparged, with water at 70° C.

The decoction brewer transfers the mash to a separation vessel called the lauter tun, where a shallow filter bed is formed, allowing a more rapid runoff time of about 2½ hours. Large modern breweries use either lauter tuns or special mash filters to speed up the runoff and conduct 10 or 12 mashes a day. As much as 97 percent of the soluble material is obtained, and 75 percent of this is fermentable. Wort is approximately 10 percent sugar (mainly maltose and maltotriose), and it contains amino acids, salts, vitamins, carbohydrate, and small amounts of protein.

Boiling. After separation, the wort is transferred to a vessel called the kettle or copper for boiling, which is necessary to arrest enzyme activity and to obtain the bitterness value of added hops.

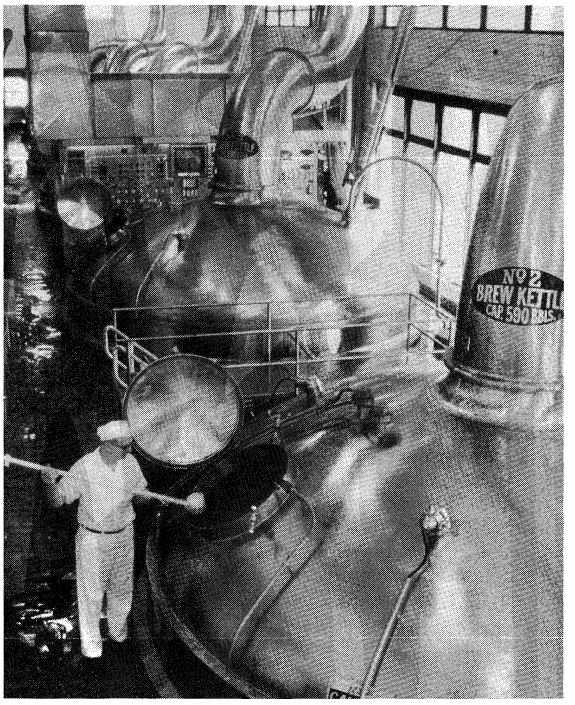
Flavouring value of hops

Hops. Several varieties of the hop (*Humulus lupulus*) are selected and bred for the bitter and aromatic qualities that they lend to brewing. The female flowers, or cones, produce tiny glands that contain the chemicals of value in brewing. Humulones are the chemical constituents extracted during wort boiling. One fraction of these, the α-acids, is isomerized by heat to form the related iso-α-acids, which are responsible for the characteristic bitter flavour of beer.

Traditionally the dried hop cones are added whole to the boiling wort, but powdered compressed hops are often used because they are more efficiently extracted. In addition, the hop components may be extracted by solvents such as liquid carbon dioxide and added in this form to the wort or, after isomerization, to the finished beer.

Heating and cooling. The kettle boil lasts 60 to 90 minutes, sterilizing the wort, evaporating undesirable aromas, and precipitating insoluble proteins (known as hot break, or trub). Trub and spent hops are then removed in a separator where the hop cones form the filter bed. In modern practice a more rapid whirlpool separator is also used. This device is a cylindrical vessel into which wort is pumped at a tangent, the circulating whirlpool movement causing solids to form a cone at the bottom. Clarified wort is cooled, formerly in shallow troughs or by trickling down an inclined cooled plate but now in a plate heat exchanger. This last is an enclosed, hygienic vessel in which hot wort runs along plates while cold water passes along the other side in the opposite direction. Oxygen is added at this stage, and the cooled wort passes to fermentation vessels.

By courtesy of the Miller Brewing Company, Wisconsin; photo by Scott J. Witte



Testing the wort before adding hops and boiling the kettle.

Fermentation. In this most important stage of the brewing process, the simple sugars in wort are converted to alcohol and carbon dioxide. Fermentation is carried out by yeast, which is added, or pitched, to the wort at three kilograms per hectolitre (about four ounces per gallon), yielding 10,000,000 cells per millilitre of wort.

Yeast. Yeasts are classified as fungi; those strains used for fermentation are of the genus *Saccharomyces* (meaning "sugar fungus"). In brewing it is traditional to refer to ale yeasts used predominantly in top fermentation as top strains of *S. cerevisiae* and to lager yeasts as bottom strains of *S. carlsbergensis.* Modern yeast systematics, however, classifies all brewing strains as *S. cerevisiae,* and many ales are made by bottom fermentation with what were originally top strains.

More than 400 simple organic compounds have been characterized in beer and many more identified, and the majority of these are produced by yeast. The bitter substances of hops, ethyl alcohol, and carbon dioxide have the greatest effects on the senses of taste and smell. Other compounds giving a beer its character include: esters such as isoamyl acetate (banana), ethyl hexanoate (apple), and ethyl acetate (solvent); higher alcohols such as isoamyl alcohol and 2-phenyl ethanol; acids such as octanoic, acetic, isovaleric, butyric malic, and citric; dialkyl sulfides such as dimethyl sulfide; and diketones such as diacetyl. The ester ethyl isovalerate and the aldehyde nonenal contribute to stale and oxidized flavours. The mechanisms of metabolism leading to the formation of these flavouring agents is neither well understood nor easily changed. Until new processes (perhaps genetic engineering) can produce changes in brewer's yeast, brewers will attach great value to known yeast strains and will maintain selected strains for brewing particular beers.

Formation of flavouring compounds

Fermenting methods. Brewing is unique among the fermentation industries in that yeast from one fermentation is used to pitch the next. This means that hygienic conditions and rigorous quality control are necessary. A high proportion of live cells and freedom from bacteria and other yeasts are important quality considerations.

Traditional open-topped earthenware fermentation vessels gave way to round, wooden and later square, copper-lined fermentors, and brewery fermentation systems evolved around the mechanism used to separate yeast from freshly fermented, or green, beer. Top fermentations, in which yeast rises to the surface, require the most elaborate systems, but most brewing operations now use more hygienically operated closed vessels and bottom fermentation. These vessels, erected outside the brewery, are several thousand hectolitres in capacity (one hectolitre = 26 gallons) and are made of stainless steel. Temperature control is achieved automatically by circulating cold liquid in jackets fitted to the wall of the vessel. Large ale breweries also use this system, removing ale yeast (*S. cerevisiae*) from the bottom of the vessel.

The temperature of the wort at pitching is 15° to 18° C (59° to 65° F) for ale and 7° to 12° C (50° to 63° F) for lager. As fermentation proceeds, the specific gravity falls as the sugars are converted by the yeast. The extent of fermentation is governed by the wort composition and by the amount of fermentable sugar to remain in maturing beer. During fermentation, yeast multiples five- to eightfold and generates heat. The temperature is allowed to rise until it reaches 20° to 23° C (68° to 74° F) for ale and 12° to 17° C (54° to 63° F) for lager. At that point the fermentation is cooled to 15° C (59° F) for ale and 4° C (39° F) for lager, considerably slowing yeast action. Yeast is then removed and the green beer, still containing about 500,000 yeast cells per millilitre, is transferred to a conditioning or maturation vessel, where a secondary fermentation may take place. In traditional brewing, the primary stage of fermentation took seven days for ale and three weeks or more for lager. These times have been shortened to two to four days and seven to 10 days by modern practices using more efficient fermentation vessels.

MATURATION AND PACKAGING

Priming and krausening

A slow secondary fermentation of residual or added sugar (called primings) or, in lager brewing, the addition of actively fermenting wort (called krausen) generates carbon dioxide, which is vented and purges the green beer of undesirable volatile compounds. Continued yeast activity also removes strong flavouring compounds such as diacetyl. Allowing pressure to build up in the sealed vessel then increases the level of carbonation, giving the beer its "condition." In traditional brewing, large volumes of ale were conditioned in tanks for seven days at 15° C, whereas lagers were matured at 0° C (32° F) for up to three months. These long maturation periods were caused by the precipitation of protein–tannin complexes, which at low temperature form "chill hazes" that are slow in settling out. Modern practice speeds up this process by adding excess tannin, clarifying with protein or tannin adsorbents, or using enzymes to degrade the proteins.

Traditional, or "real," ales are packaged into casks. Sugar primings, clarifying agents such as isinglass finings, and whole hops are added, and the beer is transferred to the point of sale, where it is carefully vented to the proper level of conditioning before being sold. Some British and Australian ales are packaged in bottles together with yeast to make "bottle-conditioned" beer.

Beer produced on a large scale in modern breweries is kept free of oxygen (which ultimately spoils beer), filtered through cellulose or diatomaceous earth to remove all yeast, and packaged at 0° C under pressure of carbon dioxide. Beer produced by high-gravity brewing is diluted to the desired alcohol concentration, immediately prior to packaging, with oxygen-free, carbonated water. Most beers packaged in bottles or metal cans are pasteurized in pack by heating to 60° C for five to 20 minutes. Beer is also packaged into metal kegs of 50-litre (in the United States, 15-gallon) capacity after pasteurization at 70° C for five to 20 seconds. Modern packaging machinery is designed to operate hygienically, exclude air, and run at rates of 2,000 cans or bottles per minute. (T.W.Y.)

Wine

Wine is the fermented juice of the grape. Of the grape genus *Vitis,* one species, *V. vinifera* (often erroneously called the European grape), is used almost exclusively. Beverages produced from *V. labrusca,* the native American grape, and from other grape species are also considered wines. When other fruits are fermented to produce a kind of wine, the name of the fruit is included, as in the terms peach wine and blackberry wine.

HISTORY

The spread of viticulture. *Vitis vinifera* was being cultivated in the Middle East by 4000 BC, and probably earlier. Egyptian records dating from 2500 BC refer to the use of grapes for wine making, and numerous Old Testament references to wine indicate the early origin and significance of the industry in the Middle East. The Greeks carried on an active wine trade and planted grapes in their colonies from the Black Sea to Spain. The Romans carried grape growing into the valleys of the Rhine and Moselle (which became the great regions of Germany and Alsace), the Danube (in modern-day Romania, Yugoslavia, Hungary, and Austria), and the Rhône, Saône, Garonne, Loire, and Marne (which define the great French regions of Rhône, Burgundy, Bordeaux, Loire, and Champagne, respectively). The role of wine in the Christian mass helped maintain the industry after the fall of the Roman Empire, and monastic orders preserved and developed many of the highly regarded wine-producing areas in Europe.

Greek and Roman wine trade

Following the voyages of Columbus, grape culture and wine making were transported from the Old World to the New. Spanish missionaries took viticulture to Chile and Argentina in the mid-16th century and to lower California in the 18th. With the flood of European immigration in the 19th and early 20th centuries, modern industries, based on imported *V. vinifera* grapes, were developed. The prime wine-growing regions of South America were established in the foothills of the Andes Mountains. In California, the centre of viticulture shifted from the southern missions to the Central Valley and the northern counties of Sonoma, Napa, and Mendocino.

British settlers planted European vines in Australia and New Zealand in the early 19th century, and Dutch settlers took grapes from the Rhine region to South Africa as early as 1654.

The introduction of the eastern American root louse, phylloxera, seriously threatened wine industries around the world between 1870 and 1900, destroying vineyards almost everywhere that *V. vinifera* was planted but especially in Europe and parts of Australia and California. To combat this parasite, *V. vinifera* scions (detached shoots including buds) were grafted to species native to the eastern United States, which proved almost completely resistant to phylloxera. After the vineyards recovered, European governments protected the reputations of the great regions by enacting laws that allotted regional names and quality rankings only to those wines produced in specific regions under strictly regulated procedures. Today, newer wine-producing countries have passed similar regulations.

Enology: scientific winemaking. Prior to the 19th century little was known about the process of fermentation or the causes of spoilage. The Greeks stored wine in earthenware amphorae, and the Romans somewhat extended the life of their wines with improved oaken cooperage, but both civilizations probably drank almost all of their wines within a year of vintage and disguised spoilage by adding such flavourers as honey, herbs, cheese, and salt water. Wooden barrels remained the principal aging vessels until the 17th century, when mass production of glass bottles and the invention of the cork stopper allowed wines to be aged for years in bottles.

Early storage

In the mid-19th century the French chemist Louis Pasteur and others explained the nature of fermentation and identified the yeasts responsible for it. Pasteur also identified the bacteria that spoil wine and devised a heating method (later called pasteurization) to kill the bacteria. Later in the century, methods were developed for growing pure strains of specific yeasts in culture. Advances in plant physiology and plant pathology also led to better vine training and less mildew damage to grapes.

Mechanized innovations in the 20th century have mainly contributed to quality control. Stainless steel fermentation and storage tanks are easily cleaned and can be refrigerated to precise temperatures. Automated, enclosed racking and filtration systems reduce contact with bacteria in the air. Beginning in the 1960s, the use of mechanical grape harvesters and field crushers allowed quick harvesting and immediate transfer to fermentation tanks.

THE WINE GRAPE

The thousands of grape varieties that have been developed, with 5,000 reported for *V. vinifera* alone, differ from one another in such characteristics as colour, size, and shape of berry; juice composition (including flavour); ripening time; and disease resistance. They are grown under widely varying climatic conditions, and many different processes are applied in producing wines from them. All of these possible variations contribute to the vast variety of wines available.

Species and varieties. *Vitis vinifera,* probably originating in the Caucasus Mountains, is the principal wine-producing plant, with most of the world's wine still made from varieties of this species. *V. labrusca* and *V. rotundifolia* have been domesticated in the eastern United States, the domestication of *V. amurensis* has been reported in Japan, and various interspecies hybrids have been used for wine production. The high sugar content of most *V. vinifera* varieties at maturity is the major factor in the selection of these varieties for use in much of the world's wine production. Their natural sugar content, providing necessary material for fermentation, is sufficient to produce a wine with alcohol content of 10 percent or higher; wines containing less alcohol are unstable because of their sensitivity to bacterial spoilage. The moderate acidity of ripe grapes of the *V. vinifera* varieties is also favourable to wine making; the fruit has an acidity of less than 1 percent (calculated as tartaric acid, the main acid in grapes) and a

Qualities of *Vitis vinifera*

pH of 3.1 to 3.7 (mildly acid). Malic acid is also an important acid; only small amounts of citric acid are present.

A third factor attracting wine makers to this grape is its tremendous range in composition. The pigment pattern of the skin varies from light greenish yellow to russet, to pink, red, reddish violet, or blue-black; the juice is generally colourless, although some varieties have a pink to red colour, and the flavour varies from quite neutral to strongly aromatic (Gewürztraminer, Cabernet Sauvignon, Zinfandel). Some varieties, such as Pinot Noir, having rather neutral flavoured juice, develop a characteristic flavour when fermented on the skins and aged.

The species *V. labrusca* and *V. rotundifolia* seldom contain sufficient natural sugar to produce a wine with alcohol content of 10 percent or higher, and additional sugar is usually required. Their acidity at maturity is often excessive, with a low pH. Varieties of these species usually have distinctive flavours. The flavours of *V. labrusca,* owing to methyl anthranilate and other compounds, are considered too pronounced by some consumers. This flavour, especially prevalent in wines made from the Concord-type varieties, is commonly called "foxy."

Cultivation. Grapes, although primarily a temperate-zone plant, can be grown under semitropical conditions. They are not adapted to the cooler parts of the temperate zone, where growing seasons may be too short to allow the fruit to reach maturity or where low winter temperatures (less than −7° C [20° F]) may kill the vine or its fruitful buds. *V. vinifera* is more susceptible to damage from winter conditions than is *V. labrusca.*

Climate strongly influences the composition of mature grapes. A major cause of the variation among grapes from different areas is the differing quantities of heat received by the vines during the growing season. Other important factors include differences in night and day temperature, hours of sun, and soil temperature.

Grapes begin their growth cycle in the spring when average daily temperature is about 10° C (50° F). To reach maturity, they require a certain amount of heat above 10° C during the growing season. This amount of heat, called the heat summation, is calculated by totaling the number of degrees of average daily temperature over 10° C for each day of the growing season. A heat summation of about 1,800° is required for successful growth. If the heat summation is less than required, the grapes will not ripen; they will reach the end of the growing season with insufficient sugar and too much acidity. This condition, frequently occurring in the eastern United States, Switzerland, and other cool regions, can be corrected by adding sugar to the crushed grapes. Where the heat summation is much greater than required, as in Algeria and parts of California, the grapes mature earlier and with less acidity and colour than those produced under cooler conditions.

The heat summation

Factors influencing the heat summation of a vineyard and, therefore, grape composition include: exposure (in Europe, best from the east); air drainage (preferably from the slopes to the valley); soil temperature (above 10° C during the growing season); and soil moisture content (not too dry at any time and not waterlogged for more than short periods).

Seasonal conditions also can be critical, especially in regions of low heat summation, as found in parts of France and Germany. When the growing season in such areas is warmer than usual, the fruit produced is riper and better balanced than is usual in cool seasons. In warm regions the sweeter dessert wines may benefit from somewhat low heat summation, resulting in less berry raisining (moisture loss) and giving the fruit better colour and acidity than is achieved when the growing season is excessively warm.

Such cultivation practices as weeding and pruning also may influence the mature fruit composition. Although the composition of the soil has an influence on soil temperature, root penetration, water-holding capacity, and vine nutrition, its effect on the quality of wine, varying from region to region, is poorly understood.

WINE REGIONS AND VARIETIES

Almost all wines are labeled by the region of production, maturity of the fruit, variety of grape or type of wine, and year of production, and they can be further distinguished by such characteristics as colour, sweetness, and varietal aroma. Specific characteristics are traditionally associated with certain wines, and in many cases these traditions are guaranteed by law.

Discussed below are the wines and viticultural laws of France, Italy, West Germany, the United States, Australia, and South Africa. Many other countries produce enormous quantities of table wines. In Europe there are Spain, Portugal, Switzerland, Hungary, Yugoslavia, Romania, Bulgaria, Greece, and the Soviet Union. In North Africa and the Middle East there are Algeria, Tunisia, and Israel. In South America there are Brazil, Peru, Chile, and Argentina. In Asia Japan is the largest producer.

Europe. In Europe wines are primarily distinguished by the region where they are produced.

France. Most French wines are everyday *vins ordinaires,* of no outstanding regional, varietal, or vintage characteristics. The finest wines are entitled to the *appellation d'origine contrôlée* (AOC; "controlled name of origin"), which is based on a hierarchy of specific geographic areas known to produce the best wines. The largest area in the hierarchy is the region; allowing for some variation, within the regions are districts, within the districts are communes, and within the communes are vineyards, or châteaus. To receive any of these successively more rigorous *appellations,* wines must be produced within specific areas and must meet standards of grape variety, alcoholic content, quantity of harvest, and techniques of vine growing and wine making. Of the smaller areas, some châteaus and communes receive rankings of quality such as *villages, supérieure,* and *grand cru* ("great vintage").

The French *appellation d'origine contrôlée*

The greatest regions of France are Bordeaux, Burgundy, Rhône, Loire, Champagne, and Alsace. Following the AOC hierarchy, Bordeaux contains such districts as Médoc, which contains the commune Pauillac, which in turn contains three *grand cru* châteaus. Bordeaux wines are mainly red and dry (except for those of the district of Sauternes, which are white and sweet). Primary varieties for the red wines are Cabernet Sauvignon, Cabernet Franc, and Merlot; for the white, Sauvignon Blanc and Sémillon.

Burgundy is smaller than the Bordeaux region. It comprises the districts of Chablis (dry white wines), Côte d'Or (red and white), Beaujolais (red), and Mâcon (white and red). The white wines are made from Chardonnay or Aligoté, the reds from Pinot Noir or (in Beaujolais) Gamay.

The Rhône region produces mostly strong, full-bodied red wines from the Syrah grape. The Loire is known for its white wines, the district of Pouilly-Fumé using Sauvignon Blanc grapes and Vouvray using Chenin Blanc. In the Champagne, legal definitions extend to the bottle-fermentation process by which the sparkling wine is produced; Pinot Noir and Chardonnay are the principal varieties. Alsace differs from other French regions by defining its mostly dry white wines primarily by grape variety, producing Alsatian Riesling, Gewürztraminer, Pinot Gris, and Sylvaner.

Wines receiving the classification *vins délimités de qualité supérieure* (VDQS; "delimited wines of superior quality"), must meet standards of region, variety, alcohol content, and sensory quality that result in good quality but are less severe than those of the AOC.

Italy. Known for its huge output of everyday red *vini da tavola* ("table wines"), Italy labels its best traditional wines as *denominazione di origine controllata* (DOC) or *denominazione di origine controllata e garantita* (DOCG). These wines must be produced in specific regions and must adhere to standards similar to the French AOC. Labels may indicate the grape variety—as in Barbera d'Alba, a red wine of the Barbera grape grown in the district of Alba in the Piedmont region.

Piedmont produces red Barolo and Barbaresco and the white, sparkling Asti Spumante. Vermouth, the flavoured dessert wine of Italy, originated in Turin, the principal Piedmontese city. From the district of Verona in the Veneto region come the red wines of Valpolicella and Bardolino and the whites of Soave. Tuscany is famous for the red wines of the various Chianti zones. Dry white Frascati wines come from the Latium region near Rome, while

Marsala, the fortified wine sweetened with concentrated grape juice, comes from Sicily.

Germany. The prime viticultural areas of West Germany fall into 11 regions, which are divided into districts, villages, and vineyards. A wine of better quality than the everyday *Tafelwein* and *Landwein* may receive the classification *Qualitätswein bestimmter Anbaugebiete* (QbA; "quality wine from a designated region") if it is produced in a specific region and meets standards of taste and alcohol content. Sugar may be added in the production of QbA wines to make up for Germany's short, cool growing season. Wines of the highest category, *Qualitätswein mit Prädikat* (QmP; "quality wine with special attributes"), must come from specific districts and be fermented from their natural sugar. The various *Prädikate* reflect the ripeness of the grape and, therefore, the sweetness of the wine. In order of increasing sugar content, they are: *Kabinett* (ripe harvest); *Spätlese* (late harvest); *Auslese* (selected late harvest); *Beerenauslese* (selected overripe); and *Trockenbeerenauslese* (from berries dried on the vine).

German *Qualitätswein*

About 90 percent of German wines are white. Riesling, Sylvaner, Müller-Thurgau, and Gewürztraminer grapes create the soft, fragrant, low-alcohol wines for which the country is famous.

Regions outside Europe. The newer wine-producing countries, lacking the centuries-old viticultural regions of Europe, emphasize the grape variety in their production of fine wines. Beginning in the 1960s, some of these countries enacted regulations guaranteeing the authenticity of these wines.

United States. Much American wine is mass-produced generic wine, often given such European-derived names as chablis, burgundy, and port. These brands must include an appellation of origin, such as California chablis, on the label.

American labeling laws

Varietal wines may be labeled after a *V. vinifera* grape if the designated variety makes up at least 75 percent of the product. It must then claim an appellation of origin. If the appellation is a county, state, or even the country, then no less than 75 percent of the wine's grapes must come from that area. If the appellation is one of the growing number of approved viticultural areas, then that area must account for 85 percent or more of the grapes. Wines may bear a vintage date if at least 95 percent of their grapes are harvested in that year.

California produces about 90 percent of American wines. The Napa Valley, Sonoma County, and other cooler areas of the north coast region produce the best wines. Cabernet Sauvignon and Chardonnay are the most prestigious, followed by Sauvignon Blanc and Pinot Noir. The Zinfandel, grown almost exclusively in California, produces a wine equal to those of the classic European grapes. California wines tend to be of higher alcoholic content and more pronounced varietal aroma and flavour than their European counterparts.

Australia. The main regions are found in an arc rimming the cooler southern states of New South Wales, Victoria, and South Australia. The Shiraz grape produces fine red wines, as does the Cabernet Sauvignon. Prominent white wines are Sémillon and Chardonnay. Sweet dessert wines are produced from Muscat and other grapes.

South Africa. Under the Wines of Origin laws, 75 percent of a varietal wine must come from the designated variety. The wine may claim one of many designated regions of origin only if all of the grapes come from that region (80 percent for fortified wines).

Long famous for sherry-type wines made from the Chenin Blanc (also called the Steen), South Africa also produces wines from several other noble varieties in areas along the cooler southwestern Cape.

THE WINE-MAKING PROCESS

Harvesting. Fresh and fully ripened wine grapes are preferred as raw material for wine making. In cool climates, as in northern Europe and the eastern United States, however, lack of sufficient heat to produce ripening may necessitate harvesting the grapes before they reach full maturity. The resulting sugar deficiency may be corrected by direct addition of sugar or by the addition of a grape juice concentrate. Grapes allowed to reach full maturity on the vine, or partially dried by exposure to sun after harvesting, are high in sugar content as a result of natural moisture loss (partial raisining as in the production of Málaga wines in Spain). A beneficent mold, *Botrytis cinera,* may also be employed to hasten moisture loss (as in the production of Sauternes in France). These grapes are used to produce sweet table wines. Special methods employed to produce these wines include the addition of sulfur dioxide, the use of small fermenting vessels during processing, or the use of cool temperatures—the objective being to stop the fermentation before all the sugar is fermented.

Because of the effect upon grape composition, proper timing of the harvest is of great importance. Premature harvesting results in thin, low-alcohol wines; very late harvesting may yield high-alcohol, low-acid wines.

Harvesting may be completed in one picking or in several. The grape cluster is cut from the vine with a special knife, or shears may be used for their convenience in removing rotten berries. The clusters, placed in buckets or boxes, are transferred to larger containers (large tubs in Europe, metal gondola trucks in California and elsewhere) for transportation to the winery. Mechanical harvesting systems, based on shaking the berries from the clusters or on breaking the stems, are widely used in California, Australia, France, and elsewhere.

Mechanical harvesting

At the winery the grapes may be dumped directly into the crusher or may be unloaded into a sump and carried to the crusher by a continuous conveyor system.

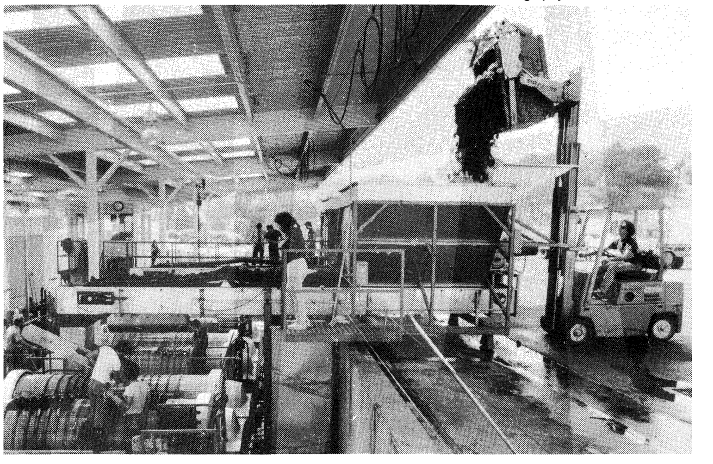
Crushing. In modern mechanized wine production, the grapes are normally crushed and stemmed at the same time by a crusher-stemmer, usually consisting of a perforated cylinder containing paddles revolving at 600 to 1,200 revolutions per minute. The grape berries are crushed and fall through the cylinder perforations; most of the stems pass out of the end of the cylinder. A roller-crusher may also be used. Ancient methods of crushing with the feet or treading with shoes are rare.

When red grapes are used to produce a white juice, as in the Champagne region of France, crushing is accomplished by pressing.

Red grapes are sometimes introduced whole into tanks, which are then closed. The resulting respiration in the fruit, consuming oxygen and producing carbon dioxide, kills the skin cells, which lose their semipermeability, allowing easy colour extraction. There is also some intracellular respiration of malic acid. This respiration process is slow and in warm regions may result in wines of low colour and acidity and distinctive odour.

Juice separation. When the juice of white grapes is processed or a white wine is desired, the juice is usually separated from the skins and seeds immediately after crushing. Occasionally, to increase flavour extraction, the white skins may be allowed to remain in contact with the juice for 12 to 24 hours, but this procedure also increases colour extraction, sometimes undesirably.

Alan Pitcairn/Grant Heilman Photography



Pouring Pinot Noir grapes into the press for the production of sparkling wine, Napa Valley, Calif.

Two main procedures are employed to separate the juice from the solids. Much of the juice may be drained off by placing the crushed grapes in a container having a false bottom and often false sides. This juice is called the free run juice, and the mass of crushed grapes is called the must, a term also used to refer to the unfermented grape juice, with or without skins.

Pressing the grapes

More commonly, the crushed grapes are placed in a press. The traditional basket press is gradually being supplanted by a horizontal basket press, applying pressure from both ends. Continuous screw-type presses are also employed, especially for drained pulp. The Willmes press, widely employed for white musts, consists of a perforated cylinder containing an inflatable tube. The crushed grapes are introduced into the cylinder, and the tube is inflated, pressing the grapes against the rotating cylinder sides and forcing the juice out through the perforations. Several pressings may be made without the extensive hand labour required for basket presses.

Continuous presses are practical for production of red wines, in which skins, seeds, and juice are all fermented together. Separation of the juice is simplified because fermentation makes the skins less slippery, and the amount of free run juice obtained is, therefore, much greater than for unfermented musts. Separation of the less slippery solids from the juice by pressing is also simplified.

The drained pomace (crushed mass remaining after extraction of the juice from the grapes), from white or red fermentations, may be used to provide distilling material for production of wine spirits. Water is usually added, the fermentation is completed, and the low-alcohol wine is drained off. The pomace may be further washed and pressed or may be distilled directly in special stills.

Must treatment. White musts are often turbid and cloudy, and settling is desirable to allow separation of the suspended materials. Such measures as prior addition of sulfur dioxide and lowering of the temperature during settling help prevent fermentation and allow the suspended material to settle normally. In many areas wineries centrifuge the white must to remove the solids. In this process a strong pulling force is created by circular motion. Musts are sometimes pasteurized, inactivating undesirable enzymes that cause browning. The addition of pectin-splitting enzymes to the musts to facilitate pressing is uncommon. Bentonite, a type of clay, may be added to musts to reduce total nitrogen content and facilitate clarification.

There is renewed interest in the prefermentation heat treatment of red musts to extract colour and deactivate enzymes. This process, when performed rapidly at moderate temperatures and without undue oxidation, may be particularly desirable in the production of red sweet wines, which employs short periods of fermentation on the skins, and for use with red grapes that have been attacked by the parasitic fungus *Botrytis cinera,* which contains high amounts of the polyphenol oxidase type of enzymes that cause browning.

Fermentation. The process of alcoholic fermentation requires careful control for the production of high quality wines. Requirements include suppression of the growth of undesirable microorganisms, presence of adequate numbers of desirable yeasts, proper nutrition for yeast growth, temperature control for prevention of excessive heat, prevention of oxidation, and proper management of the cap of skins floating in red musts.

Grape skins are normally covered with bacteria, molds, and yeast. The wild yeasts such as *Pichia, Kloeckera,* and *Torulopsis* are often more numerous than the wine yeast *Saccharomyces.* Although species of *Saccharomyces* are generally considered more desirable for efficient alcoholic fermentation, it is possible that other yeast genera may contribute to flavour, especially in the early stages of fermentation. *Saccharomyces* is preferred because of its efficiency in converting sugar to alcohol and because it is less sensitive to the inhibiting effect of alcohol. Under favourable conditions, strains of *Saccharomyces cerevisiae* have produced up to 18 percent (by volume) of alcohol, although 15 to 16 percent is the usual limit.

Wine yeast

Use of the yeast *Schizosaccharomyces pombe* has been proposed for the early stages of alcoholic fermentation. Because it metabolizes malic acid, this yeast would be useful in excessively acid musts, but commercial applications have not yielded consistently favourable results. The addition of lactic-acid bacteria to musts, using strains metabolizing malic acid, is now common.

The number of undesirable microorganisms is greatest in partially rotted or injured grapes. Such damage may occur in harvesting or during transportation, particularly in warm climates. Suppression of undesirable microorganism growth is required, and the most common method used is the addition of sulfur dioxide to the freshly crushed grapes at the rate of about 100 to 150 milligrams per litre. Sulfur dioxide is more toxic to undesirable microorganisms than to desirable microorganisms. When it is used in musts, an inoculum of the desired yeast strain, usually called a pure yeast culture, is added. Musts are rarely pasteurized, although this process may be applied when they contain undesirable amounts of oxidizing enzymes from moldy grapes.

Enologists, technicians in the science of wine making, do not agree on the most desirable yeast species and strain, but strains of *S. cerevisiae* are generally used. The chosen strain is allowed to multiply as much as possible in sterilized grape juice and is then transferred to larger containers of sterilized grape juice, where it continues to grow until the desired volume is reached. Suitable pressed yeasts of desirable strains are added directly, avoiding the troublesome practice of building up and maintaining a pure yeast culture. About 1 to 3 percent of a pure yeast culture, or sufficient pressed yeast to provide a population of 1,000,000 cells per millilitre, is used.

Temperature control during alcoholic fermentation is necessary to (1) facilitate yeast growth, (2) extract flavours and colours from the skins, (3) permit accumulation of desirable by-products, and (4) prevent undue rise in temperature, killing the yeast cells.

Optimum temperature for growth of common wine yeasts is about 25° C (77° F), and in many viticultural areas of the cooler temperate zone, grapes are crushed at about this temperature. Fermentation is seldom started at so high a temperature, however, because it is then difficult to prevent the temperature from exceeding 30° C during fermentation.

Extraction of flavours and colours is not a problem in white musts; the crushed grape mass is usually separated from the skins before fermentation. Fermentation of white musts at relatively cool temperature (about 10° to 15° C [50° to 60° F]) apparently results in greater formation and retention of desirable by-products. An undesirable feature of such relatively low-temperature fermentations is the longer period required for completion (six to 10 weeks compared to one to four weeks at higher temperatures) and the tendency for the fermentation to stop while residual sugar remains. (This is not always considered undesirable—*i.e.,* in German wine production.) In practice white table wines are usually fermented at about 20° C.

Optimum temperatures for fermentation

In red wine musts, the optimum colour extraction consistent with yeast growth occurs at about 22° to 28° C (72° to 82° F). Alcoholic fermentation produces heat, however, and careful temperature control is required to prevent the temperature from reaching a point (about 30° C) where yeast growth is seriously restricted. At still higher temperatures, growth will stop completely. Modern temperature control is accomplished by use of heat exchangers. Older methods include placing the fermenters in a cold room; using cold pipes in the fermenter; pumping the must through double-walled pipes, with cold water in the surrounding pipe; pumping the must through a sump containing cooling coils; and pumping the coolant through jackets surrounding the tank.

Contact with air must be restricted to prevent oxidation during fermentation. In very large containers, the volume of carbon dioxide given off is sufficient to prevent entry of air. In small fermenters, fermentation traps are inserted, preventing entry of air but permitting exit of carbon dioxide. These traps are particularly desirable during the final stage of fermentation, when carbon dioxide evolution is slow. Following fermentation, small amounts of sulfur

dioxide are added to help prevent oxidation. Ascorbic acid (50 to 100 milligrams per litre) is sometimes employed to decrease the oxidation and thus the amount of sulfur dioxide required as an antioxidant, but is not generally recommended.

The cap of skins and pulp floating on top of the juice in red-wine fermentation inhibits flavour and colour extraction, may rise to an undesirably high temperature, and may acetify if allowed to become dry. Such problems are avoided by submerging the floating cap at least twice daily during fermentation. This operation, comparatively easy with small fermenters, becomes difficult with large, tall fermenters of up to 100,000-gallon (380,000-litre) capacity. In large units the fermenting must is drawn off near the bottom and pumped back over the cap. The use of small fermentation vessels permits a greater percentage of heat loss to the surrounding atmosphere, simplifying temperature control.

Postfermentation treatment. With appropriate must composition, yeast strain, temperature, and other factors, alcoholic fermentation ceases when the amount of fermentable sugar available becomes very low (about 0.1 percent). Fermentation will not reach this stage when (1) musts of very high sugar content are fermented, (2) alcohol-intolerant strains of yeast are used, (3) fermentations are carried on at too low or high temperatures, and (4) fermentation under pressure is practiced. Fermentation of normal musts is usually completed in 10 to 30 days. In most cases, the major portion of the yeast cells will soon be found in the sediment, or lees. Separation of the supernatant wine from the lees is called racking. The containers are kept full from this time on by "topping," a process performed frequently, as the temperature of the wine, and hence its volume, decreases. During the early stages, topping is necessary every week or two. Later, monthly or bimonthly fillings are adequate.

Racking

Normally the first racking should be performed within one to two weeks after completion of fermentation, particularly in warm climatic regions or in warm cellars, as the yeasts in the thick deposit of lees may autolyze (digest themselves), forming off-odours.

Early racking is not required for wines of high total acidity—*i.e.,* those produced in cool climatic regions or from high-acid varieties. Such wines may remain in contact with at least a portion of the lees for as long as two to four months, permitting some yeast autolysis in order to release amino acids and other possible growth factors favouring growth of lactic-acid bacteria. These bacteria then induce the second, or malolactic, fermentation.

Malolactic fermentation. Enologists have known for some time that young wines frequently have a secondary evolution of carbon dioxide, occurring sometime after the completion of alcoholic fermentation. This results from malolactic fermentation, in which malic acid is broken down into lactic acid and carbon dioxide. The fermentation is caused by enzymes produced by certain lactic-acid bacteria.

Flavour by-products of unknown composition are also produced during this fermentation. Malolactic fermentation is desirable when new wines are too high in malic acid, as in Germany, or when particular nuances of taste and flavour are desired, as in the red wines of Burgundy and Bordeaux in France. In other regions, some producers may encourage malolactic fermentation, and others may discourage it, depending upon the particular character desired in the wine. In all regions, this second fermentation is somewhat capricious. One product, diacetyl (a flavour and aroma agent), is apparently beneficial at low levels and undesirable at higher levels.

At low temperatures, malolactic fermentation proceeds slowly, if at all. German cellars are often equipped with steam pipes, raising the temperature to encourage this fermentation. The bacteria may fail to grow because of a deficiency or complete absence of essential amino acids. Most lactic-acid bacteria growth can be inhibited by the presence of 70 to 100 milligrams per litre of sulfur dioxide.

Excessive malolactic fermentation may produce wines too low in acidity (flat tasting) or with undesirable odours (mousy, sauerkraut, or diacetyl). Such faults may be prevented by earlier racking, filtration, and addition of sulfur dioxide.

Clarification. Some wines deposit their suspended material (yeast cells, particles of skins, etc.) very quickly, and the supernatant wine remains nearly brilliant. This is particularly true when 50-gallon wooden barrels, which have greater surface-to-volume ratio than larger containers, are employed. The rough interior of wooden cooperage facilitates deposition of suspended material. Other wines, particularly in warm regions or when large tanks are used, may remain somewhat cloudy for long periods. Removal of the suspended material during aging is called clarification. The major procedures involved are fining, filtration, centrifugation, refrigeration, ion exchange, and heating.

Removal of suspended material

Fining. Fining is an ancient practice in which a material that aids clarification is added to the wine. The main processes involved are adsorption, chemical reaction and adsorption, and possibly physical movement. Proteins and yeast cells are adsorbed on fining agents such as bentonite (a type of clay formed mainly of montmorillonite) or gelatin. Chemical reactions occurring with tannins and gelatin may be followed by adsorption of suspended compounds. If an inert material, such as silica, is added to a cloudy wine, some clarification will occur simply by the movement of the particles of inert silica through the wine. This action probably occurs to a certain extent with the addition of any fining agent.

Bentonite has largely replaced all other fining agents. Such fining agents as gelatin, casein, isinglass, albumin, egg white, nylon, and PVPP (polyvinyl pyrrolidone) may be used for special purposes, including removal of excess tannin or colour.

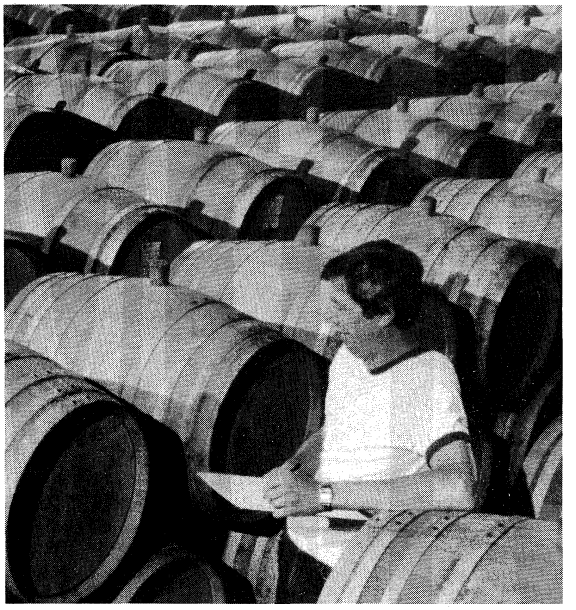
Excessive amounts of metals, particularly iron and copper, may be present in the wine, usually from contact with iron or metal surfaces. These result in persistent cloudiness and require removal by such special fining materials as potassium ferrocyanide (blue fining), long recommended in Germany. Cufex, a proprietary product containing potassium ferrocyanide, may be used in the United States under strict control. Phytates have been used for removing iron. In modern winery operations excessive metal content is rare, mainly owing to the use of stainless steel equipment.

Filtration. Filtration is another ancient practice, and early filters consisted of rough cloth-covered screens through which the wine was poured. Modern filter pads are made of cellulose fibres of various porosities or consist of membrane filters, also in a range of porosities. The pore size of some filters is sufficiently small to remove yeast cells and most bacterial cells, but filters operate not only because of pore size but also by a certain amount of adsorption. Diatomaceous earth-filter aids, commonly added to the wine during filtration, increase the functional life of a filter by retarding pore clogging.

Centrifugation. Centrifugation, or high-speed spinning, used to clarify musts, is also applied to wines that are difficult to clarify by other means. This operation requires careful control to avoid undue oxidation and loss of alcohol during the process.

Refrigeration. Refrigeration aids wine clarification in several ways. Temperature reduction often prevents both yeast growth and the evolution of carbon dioxide, which tends to keep the yeast cells suspended. Carbon dioxide is more soluble at lower temperatures. A major cause of cloudiness is the slow precipitation of potassium acid tartrate (cream of tartar) as the wine ages. Rapid precipitation is induced by lowering the temperature to −7° to −5° C (19° to 23° F) for one or two weeks. If the resulting wine is filtered off the tartrate deposit, tartrate precipitation will not usually cause clouding later.

Ion exchange. Another method of tartrate stabilization is to pass a portion of wine through a device called an ion exchanger. If this ion exchanger is charged with sodium, it will replace the potassium in potassium acid tartrate with sodium, making a more soluble tartrate. Usually, if the potassium content of the blend of either treated or untreated wine is reduced to about 500 milligrams per litre, no further precipitation will occur. Exceptions may occur, however, and to be safe, tartrate and potassium contents



Checking inventory of wine casks in the cellars of a northern California winery.
Comstock

and pH are included in the calculation. The use of ion exchange is illegal in several countries.

Heating. Many wines contain small amounts of proteins that may cause clouding either by precipitation or by reacting with copper or other metals to form aggregates that in turn form clouds. The use of bentonite removes some protein, and protein adsorption is increased if the wine is warm when fined. Pasteurization at 70° to 82° C (158° to 180° F) also can be used to precipitate proteins, but in modern practice this process is seldom employed to aid clarification.

AGING AND BOTTLING

Aging. Many wines improve in quality during barrel and bottle storage. Such wines eventually reach their peak and with further aging begin to decline. During the aging period, acidity decreases, additional clarification and stabilization occur as undesirable substances are precipitated, and the various components of the wine form complex compounds affecting flavour and aroma.

Wines are usually aged in wooden containers made of oak, allowing oxygen to enter and water and alcohol to escape. Extracts from the wood contribute to flavour. Humidity affects the kind of constituents that escape, with alcohol becoming more concentrated in wine stored under conditions of low humidity and weakening with high humidity. As the water and alcohol are released, volume decreases, leaving headspace, or ullage, that is made up by the addition of more of the same wine from another container.

Aging in wood

Some red table wines appreciate in quality, developing less astringency and colour, and a greater complexity of flavour with aging in oak cooperage of up to 500-gallon size for two to three years. In the best red wines, additional improvement may continue with two to 20 years of bottle aging (the rate of aging being lower in the bottle than in the barrel). Many dessert wines improve during cask aging, particularly sweet sherries, but extraction of excessive wood flavour must be avoided. Those rosé and dry red wines that will not improve with long cask and bottle aging are aged for a short period of time, clarified, and then bottled. More than 90 percent of all table wines are probably marketed and consumed before they are two years old. In dry white wines, a fresher flavour is considered desirable, and the chief benefit of aging is greater clarification as various undesirable substances are precipitated. These wines are rarely aged in the wood for long periods, and some are never kept in wood. This change is possible because of the efficiency of new clarification methods. Earlier bottling of white wines reduces costs for storage and for handling in wooden cooperage and produces fresher, fruitier flavour. Sweet white table wines profit by some aging in wood.

Bottling. Before bottling, wine may require blending, filtration, and use of antiseptics to combat microbe development. Often several casks containing the same wine will develop differences during aging, and blending is desirable to ensure uniformity. Wines that are slightly deficient in colour or acid may be blended with special wines as a means of correction. Blending frequently improves quality by adding to the complexity of the wine.

A final polishing filtration is required before bottling, and the amount of sulfur dioxide is adjusted, especially in sweet table wines. Sulfur dioxide is frequently used, but sorbic acid or sorbates are used in sweet table wines to inhibit yeasts, although they are not generally recommended because of the off-odour that may develop. Such operations as the addition of sulfur dioxide, heating (wherever beneficial), and polishing filtration are usually accomplished by a continuous in-line process. Equipment, usually semiautomatic or completely automatic, must be free of undesirable microorganisms and is made of resistant alloys to avoid undesirable metal pickup.

During the actual bottling operation, oxygen pickup must be kept to a minimum. Bottomfilling—that is, inserting a tube into the bottle and filling from the bottom—is often used. In some cases, the bottle may be flushed with carbon dioxide before filling, or the wine may be sparged (agitated) with nitrogen gas. Wines subject to oxidation require special care.

Sterile new bottles are used in the United States. Elsewhere, bottles may be reused after thorough cleaning and sterilization. The bottle shape and colour are dictated by custom and cost. Some white wines, subject to change when exposed to light, are preferably bottled in brown, brownish green, or greenish blue coloured bottles. Although brown glass is probably preferable for Sauternes, custom dictates the use of clear bottles. Glass is still the usual material, although experiments have been made with plastics.

Bottle shape and colour

After bottling, the closure is made. Screw caps are used for standard wines. Cork closures are preferred for wines that will be aged in the bottle. Red wines that may be aged in the bottle for many years are closed with corks two inches (five centimetres) long or longer. Occasionally a cork may communicate an off-odour, called "corked," to the wine, apparently resulting from a contaminant or from a defect that allows mold growth in or on the cork.

A capsule is placed over the closure, the label is applied, and the bottles are packaged in cases for shipment. Wines requiring bottle aging are often not capsuled, labeled, or cased until they have been aged.

Bottled table and dessert wines should be stored on their sides during aging, both at the winery and by the final customer pending consumption. Appropriate storage conditions include absence of light and low, even temperatures maintained at about 12° to 16° C (54° to 61° F). Diurnal fluctuations in temperature lead to rapid aging and early deterioration.

SPECIAL WINES

The procedures discussed above are primarily concerned with the production of still (non-sparkling) table wines. Sparkling, dessert, and flavoured wines require special techniques.

Sparkling wines. Wines containing excess carbon dioxide are called sparkling wines. They are always table wines, usually containing less than 4 percent sugar. The two basic techniques used for their production are a second sugar fermentation, often induced artificially, or direct carbonation, involving the addition of carbon dioxide.

Sparkling wine results when the escape of carbon dioxide from the fermenting liquid is prevented. The basic material is usually a dry white, rosé, or red table wine. Sufficient sugar is added to the basic wine to produce a pressure of about five or six atmospheres (units of pressure, each equal to 14.7 pounds per square inch) following fermentation, assuming there is no loss of carbon dioxide. The size of the fermentation container may vary from 0.1 to 25,000

gallons. Bottles or tanks used for this type of fermentation must be capable of withstanding pressures as high as 10 atmospheres. Use of tanks equipped with pressure gauges allows excess pressure to be let off as needed. The special bottles used for sparkling wines are thicker than normal in order to withstand pressure of seven to nine atmospheres. The neck of the bottle is shaped either for seating a crown cap or with a lip that catches a steel clamp to hold the cork in place.

Special bottles for sparkling wines

The basic wine is clarified before being placed in the fermentation container. Several wines are usually blended to secure a base wine of the proper composition and flavour balance. The original alcohol content should be only 10–11.5 percent; the secondary fermentation will result in an increase of about 1 percent. The pH should be 3.3 or slightly less, with 0.7 percent or more total acidity calculated as tartaric acid, and the wine should have a fresh fruity flavour. No single or pronounced varietal character should predominate in the base wine, except in muscat-flavoured sparkling wines. Special care is necessary to avoid wines with any off character in odour or taste, or any trace of undesirable bacterial activity.

The final clarified wine is placed in the fermentation vessel, and the requisite sugar for the fermentation, about 2.5 percent, is added, along with 1 to 2 percent of an actively growing yeast culture. The strain of yeast selected should be able to ferment adequately in wines of 10 to 11.5 percent alcohol and under pressure. A number of strains have proved satisfactory. The yeast cells should settle (agglutinate) rapidly and completely after fermentation.

The secondary fermentation is carried out at 10° to 12° C (50° to 54° F) for best absorption of the carbon dioxide produced and should be completed in four to eight weeks. To save time, both tank and bottle fermentations are often conducted at temperatures of 15° to 17° C (59° to 63° F) or even higher, and the secondary fermentation is frequently completed in 10 days to two weeks.

Tank fermentation. Additional differences between tank- and bottle-fermented wines may develop after secondary fermentation. Upon completion of fermentation, tank-fermented wines are filtered to remove the yeast deposit and then bottled. The filtration operation can introduce air, sometimes leading to oxidative changes affecting colour and taste. In addition, it is difficult to accomplish the necessary filtration, removing any viable yeast cells, without reducing the level of the pressure that has been built up within the wine. Because of such difficulties, sulfur dioxide may be added to tank-fermented wines in order to prevent refermentation. While still in the tank, the wine is sweetened to the desired level by the addition of inert sugar syrup.

Bottle fermentation. Bottle-fermented wines may also be clarified soon after fermentation. In the transfer process, the bottle-fermented wine is transferred, under pressure, to a second tank, from which it is filtered and bottled. In this case, as with tank-fermented wines, little aging of the wine takes place in contact with the yeast, and sulfur dioxide may be added. The transfer process is widely used in the United States, Germany, and elsewhere.

Classic champagne method

In contrast, in classic bottle fermentation, or *méthode champenoise* ("champagne method"), the wine remains in the bottle, in contact with the yeast, for one to three years. During this period of aging under pressure, a series of complex reactions occurs, involving compounds from autolyzed yeast and from the wine, resulting in a special flavour. Bottle-aged wine is rarely transferred, filtered, or rebottled because the addition of sulfur dioxide, required to prevent oxidation, would interfere with the delicate odour so carefully developed by aging. Aged bottle-fermented wines therefore are usually clarified in the bottle. In this process the bottles are placed neck down in special racks at a 45° angle. Each day the bottle is turned to the right and left, inducing the yeast debris within to move down the side of the bottle onto the cork. This process, riddling or *remuage,* may last from a few weeks to several months. When it is complete, all of the yeast is on the cork, and the bottle is gradually brought to an inverted position of 180°. Mechanical *remuage* in large containers is widely practiced.

Remuage, the turning of champagne bottles to force yeast sediment into the bottle necks, France.
G. Lienhard

In the traditional procedure, the cork is slowly pulled out, and the pressure within the bottle propels the sediment out of the bottle. In the modern procedure, to prevent undue pressure loss, the bottle temperature is lowered to 10° to 15° C (50° to 59° F). The neck of the bottle is placed in a freezing solution and frozen solid. When the crown cap, or cork, is removed and the yeast deposit is ejected, the process is called disgorging, or *dégorgement.* The bottle is quickly turned to an upright position. When performed properly, disgorging (which is usually mechanized), involves the loss of only 3 to 5 percent of the wine. The bottle is held under pressure while it is refilled.

The filling solution is a small amount of sweetening dosage, usually white wine containing 50 percent sugar. The amount added depends on the degree of sweetness the producer desires. Wines labeled brut, or sometimes nature (a term also applied to a still champagne), are extremely dry (very low in sugar content), usually containing 0 to 1.5 percent sugar; wines labeled extra dry or extra sec, or dry or sec, are sweeter, often containing 2 to 4 percent sugar; semi-dry or demi-sec wines may contain 5 percent or more sugar; and sweet or doux wines have about 8 percent sugar. In commercial practice, there is considerable variation in the exact degree of sweetness described by a specific term. If the dosage does not bring the contents to the desired level, more wine of a previously disgorged bottle is added. The closure, made of cork or plastic, is held in place with a wire netting.

If the wine has been aged for two or three years, the sugar in the final dosage does not ferment, as that in the original dosage did, because few viable yeast cells remain. Even in wines aged for shorter periods, skillful disgorging leaves few viable yeast cells on the sides of the neck of the bottle. Furthermore, the wine lacks oxygen to stimulate yeast growth and is lower in growth-promoting nitrogenous constituents and higher in alcohol than the original wine. The high carbon dioxide content also has a repressive effect on yeast growth. When bottle-fermented wines are fermented very rapidly and disgorged early, however, it is customary to add some sulfur dioxide with the final dosage to repress yeast growth.

American labeling laws

In the United States, tank-fermented wines must be labeled "fermented in bulk" or "bulk-fermented." Bottle-fermented wines may be labeled "bottle-fermented," but only wines handled by the classic method may be labeled "fermented in this bottle."

Carbonation. Carbonation is a less involved process but is used infrequently. Carbonated wines have many characteristics of fermented sparkling wines, and this simple physical process is much less expensive. The action of the second fermentation under pressure may produce es-

pecially desirable flavour by-products, however, and there is greater prestige value attached to fermented sparkling wines. In some cases, the wines used as a base for the carbonated sparkling wines may be overmature or otherwise inferior to those used for the fermented sparkling wines.

The base wine used for carbonation, like the base wine for fermented sparkling wines, must be well balanced, with no single varietal flavour predominating. Young fruity wines are preferred, and the wine should not contain any trace of off-odour. Since no secondary fermentation takes place, wines of 11.5 to 12.5 percent alcohol content are used. The wine should be tartrate-stable, metal-stable, and brilliant, and the sulfur-dioxide content should be low. For white wines, the colour should be a light yellow.

A variety of techniques have been used for carbonation. Production of carbonation by passing the wine from one bottle to another, under carbon-dioxide pressure, is now seldom employed because of its slowness. Carbonation has been produced in bottles after deaeration, and this technique could be adapted to multibottle operations. Direct carbonation is frequently practiced with cold wine in pressure tanks, and if the stream of gas is finely divided, good carbonation is obtained. Pinpoint carbonation, spraying the wine into a pressure chamber containing carbon dioxide, may also be employed. Following the carbonation procedure, the wine is bottled under pressure. A cork or plastic or crown-cap closure is applied, the label is affixed, and the wine is cased for distribution.

In many countries, there is a higher tax on fermentation-produced sparkling wines than on carbonated sparkling wines. The two types also have different labeling requirements, and the process of carbonation usually must be stated on the label.

There are a few low-level carbon dioxide wines on the market, produced either by fermentation or by carbonation. In Germany and other areas, tank-fermented wines, or "pearl" wines, of about one atmosphere pressure, are produced. In the United States, Portugal, and Switzerland, a number of wines are lightly carbonated at the time of bottling, adding piquancy.

There are a few wines in which the carbon dioxide comes not from alcoholic fermentation, but from malolactic fermentation of excess malic acid in the wine. The *vinhos verdes* wines of northern Portugal are examples of this type. This fermentation is sometimes responsible for undesirable gassiness in red wines.

Fortified wines. The addition of alcohol during or after alcoholic fermentation produces fortified wines of over 14 percent alcohol, generally called dessert wines in the United States. In most countries, these wines are taxed at higher rates than those of 14 percent or lower alcohol.

Why wines are fortified

Fortification has two purposes: (1) to raise the alcohol content sufficiently (usually 17 to 21 percent) to prevent fermentation of all of the sugar and (2) to produce types with a special alcohol character. The alcohol used for fortification is usually (legally required in most countries) distilled from wine. The distillation of the fortifying spirits is made to a high percent alcohol, usually 95 to 96 percent. Industrial alcohol has also been employed in a few countries.

The repressive effect of alcohol on alcoholic fermentation increases rapidly as the alcohol content is raised above 14 percent, particularly in the presence of sugar. To secure prompt cessation of fermentation, the added alcohol must be rapidly and uniformly mixed with the fermenting must, and this is accomplished by stirring or mixing with compressed air.

In the most simple type of fortification, the initial fermentation is allowed to proceed nearly to, or all the way to, completion. The resulting wine is usually subjected to a baking process, as in Madeiras and California sherries, lasting for one to four months, at 58° to 65° C (136° to 149° F). If the wine is low in sugar content, heating will change the flavour and colour of the wine only slightly; with greater sugar content, a more caramelized flavour, typical of sweet Madeiras and sweet California sherries, is produced.

When white must is fortified during fermentation, the resulting wine is sweet, the degree of sweetness depending on the original sugar content of the must and the time of fortification. Some types, fortified early, produce very sweet wines. Muscatels, produced in many countries, are often of this type.

Red sweet wines, such as port, are more difficult to produce. Although the grapes must be fermented on the skins to extract colour, the fermentation cannot be continued for long if the requisite sugar is to remain in the finished wine. One method of securing sufficient colour is to use grape varieties containing large amounts of pigments in their skins. The skins and juice are sometimes heated to about 65° C (149° F) to extract colour.

The *flor* sherries, such as the dry or fino-type sherry produced in Spain, are a special type of dessert wine. The base wine is fortified to about 15 percent alcohol, and a special alcohol-tolerant film yeast develops as a film on the wine surface. Acetaldehyde, an aldehyde, is one of the flavour products produced by this procedure. Following this process, the alcohol content may be further raised to 16–18 percent. By adjusting the oxygen content, the *flor* yeast may be induced to develop and produce acetaldehyde in a submerged culture, a process used commercially in California.

Marsala, a type of dessert wine produced in Sicily, has a dark amber colour and burnt sugar flavour, derived from the addition of grape juice that has been cooked and reduced to about one-third its original volume.

Aged dessert wines

Dessert wines aged for only short periods lack the complex flavour of those dessert wines aged in small oak cooperage for at least two to four years. During aging, white wines gradually darken in colour, while red wines become less red and more amber. Flavour becomes more complex and mellow as wood flavour is extracted from the container, various substances in the wine become oxidized, and complex compounds of acids and alcohol are formed. If the wood containers are stored in warm, dry rooms, more water than alcohol is lost, and the alcohol content of the wine increases. This effect is common in dessert wines of the south of Spain. At lower storage temperatures and normal humidity, there is little change and sometimes even a slight decrease in alcohol content.

In the production of certain wines, special character is achieved by blending wines of different ages, a technique often used for port blends. By varying the proportion of the various wines, a range of types varying in colour and flavour may be produced. The blending may be performed continuously, as in the solera system common in Spain. This process involves a series of casks graduated according to the age of the wine each contains. One or more times each year, a portion of wine, usually 10 to 25 percent, is taken out of the oldest cask. This is replenished from the next oldest containers, and these in turn from younger containers. After a number of years, depending on the portion withdrawn each year and the number of years since the start, the average age of wine in the oldest container no longer changes. This process is called a fractional-blending system.

Flavoured wines. Vermouth, a flavoured wine product, probably originated in Turin in the 18th century as a sweet dessert wine with various Mediterranean and other herbs and plant materials added. A similar product, lower in sugar content, was produced in the south of France. Although sweet vermouth is often considered an Italian type and dry vermouth usually refers to the French type, both countries now produce both types. Various producers have their own formulas, and the herbs and spices used as flavourings include bitter orange peel, cinnamon, clove, coriander, mace, marjoram, nutmeg, saffron, and wormwood.

French and Italian aperitifs

Aperitif wines, usually taken before meals, are made by adding quinine and other ingredients to sweet, heavy wines. In France they are marketed under such brand names as Byrrh, Dubonnet, Lillet, and Saint Raphaël; in Italy they include Campari and Punt e Mes.

There are various flavoured wine beverages, frequently mixed by the consumer and sometimes bottled by a manufacturer, in which flavouring materials are added after the manufacture of the wine. May wine, of German origin, is a type of punch made with Rhine wine or other light,

dry, white wines, flavoured with the herb woodruff and served chilled and garnished with strawberries or other fruit. Sangria, a popular punch in many Spanish-speaking countries, is made with red or white wine mixed with sugar and plain or sparkling water, flavoured with citrus fruit, and served chilled. Mulled wine is usually made with red wine diluted with water, sweetened with sugar, flavoured with such spices as cloves and cinnamon, and served hot. Glogg, a hot punch of Swedish origin, is frequently made with red wine and contains spices, almonds, and raisins. Wine coolers, popular in the United States, are wines of low alcohol flavoured with fruit juices.

Fruit wines. Fruit wines, derived from fruits other than grapes, include cider, made from apples; perry, produced from pears; plum wine and cherry wine; and wines made from various berries. They are frequently made by home wine makers and have some commercial importance in cold climates where wine grapes are not produced. Cider and perry are important products in England and northern France; fortified cherry and black currant wines are produced in Denmark; and important U.S. fruit wines, mainly produced on the eastern coast, include apple, cherry, blackberry, elderberry, and loganberry wines. Various kinds of fruit wines are exported from The Netherlands, Denmark, Poland, Czechoslovakia, Yugoslavia, and Israel.

Fruit wines usually have sweet flavour and should retain much of the flavour and colour of the original fruit. The musts are high in acid content and require dilution with water and the addition of sugar before fermentation. Many commercial fruit wines contain about 12 percent alcohol. When they are fortified with brandy, derived from the same fruit, alcoholic content is about 20 percent. The alcoholic content of cider and perry is usually 2–8 percent.

(M.A.A.)

Distilled spirits

The production of distilled spirits is based upon fermentation, the natural process of decomposition of organic materials containing carbohydrates. It occurs in nature whenever the two necessary ingredients, carbohydrate and yeast, are available. Yeast is a vegetative microorganism that lives and multiplies in media containing carbohydrates—particularly simple sugars. It has been found throughout the world, including frozen areas and deserts.

Natural occurrence of fermentation

Distilled spirits are all alcoholic beverages in which the concentration of ethyl alcohol has been increased above that of the original fermented mixture by a method called distillation. The principle of alcoholic distillation is based upon the different boiling points of alcohol (78.5° C, or 173.3° F) and water (100° C, or 212° F). If a liquid containing ethyl alcohol is heated to a temperature above 78.5° C but below 100° C, and the vapour coming off of the liquid is condensed, the condensate will have a higher alcohol concentration, or strength.

HISTORY OF DISTILLING

Because the two ingredients necessary to alcoholic fermentation are widely spread and always appear together, civilizations in almost every part of the world developed some form of alcoholic beverage very early in their history. The Chinese were distilling a beverage from rice beer by 800 BC, and arrack was distilled in the East Indies from sugarcane and rice. The Arabs developed a distillation method that was used to produce a distilled beverage from wine. Greek philosophers reported a crude distillation method. The Romans apparently produced distilled beverages, although no references concerning them are found in writings before AD 100. Production of distilled spirits was reported in Britain before the Roman conquest. Spain, France, and the rest of western Europe probably produced distilled spirits at an earlier date, but production was apparently limited until the 8th century, after contact with the Arabs.

The first distilled spirits were made from sugar-based materials, primarily grapes and honey to make grape brandy and distilled mead, respectively. The earliest use of starchy grains to produce distilled spirits is not known, but their use certainly dates from the Middle Ages. Some government control dates from the 17th century. As production methods improved and volume increased, the distilled spirits industry became an important source of revenue. Rigid controls were often imposed on both production and sale of the liquor.

The earliest stills were composed simply of a heated closed container, a condenser, and a receptacle to receive the condensate. These evolved into the pot still, which is still in use, particularly for making malt whiskeys and some gins. The next refinement was heating the alcohol-containing liquid in a column made up of a series of vaporization chambers stacked on top of one another. By the early 19th century large-scale continuous stills, very similar to those used in the industry today, were operating in France and England. In 1831 the Irishman Aeneas Coffey designed such a still, which consisted of two columns in series.

Since distillation requires that the liquid portion of a fermentation mixture be vaporized, considerable heat must be applied to the process. The fuel used in distilling spirits has always been that which has been most readily available at the particular time and place. Peat, coal, and wood were the fuels used historically, while the fuels of choice today are coal, natural gas, and oil. The high steam requirement for continuous-still operation inhibited the development of rectifying columns for production of spirits until after the Industrial Revolution.

Many of the minor components of distilled spirits, which are present only in parts per million, are detectable by the senses of taste and smell, but efforts to identify and quantify these compounds chemically have often been hampered by the lower limits of detection by analytical methods. Classes of compounds such as aldehydes, organic acids, esters, and alcohols were easily identified by conventional methods, but many of them could not be determined until after the development of chromatography. The Russian botanist Mikhail Tsvet was an early pioneer of this measurement technique, reporting his first work in 1903. Refinements in both technique and equipment, made during the first half of the 20th century, allowed numerous flavour components in distilled spirits to be identified by gas chromatography.

PRODUCTION

Raw materials. The raw materials used for making a distilled spirit are of two basic types: (1) those containing a high concentration of natural sugars or (2) those containing other carbohydrates that can easily be converted to sugars by enzymes. Enzymes are proteins that act as catalysts to promote chemical reactions. Very small amounts of an enzyme can cause a fundamental change in a large amount of material. Most enzymes are specific in their action, so that a system of several enzymes is necessary, for example, to convert starch into sugar and ultimately into ethyl alcohol. The amylases are enzymes that convert starches into sugars; sprouting grains—especially barley—are natural sources of these enzymes. Yeast has a complex enzyme system that converts sugar into carbon dioxide and a multiplicity of other products, including ethyl alcohol.

Reduced activity of any enzyme in the system distorts the results, often forming unwanted products. Enzymes are easily poisoned by certain compounds; they are also sensitive to temperature variations and to the degree of acidity of the medium.

Sugary materials. Grapes, cultivated in most of the subtropic and warm temperate zones of the world, are the major fruit employed as the raw material of distilled spirits, and the final product of their fermentation is brandy. Other natural fruits, such as apples and peaches, are used to a lesser extent, and many fruits are limited to local importance.

Sugary vegetables include sugarcane, sugar beets, and *Agave tequilana* (a type of cactus). Sugarcane and its products, including cane juices, molasses, and sugar, are the most important of the vegetable group. Grown throughout the tropics and semitropics, sugarcane is used in making rum and an alcohol derived from rum. Sugarcane juice can be pressed from the cane for use as the base raw ma-

terial for fermentation, or the juice may be concentrated for sugar production, with the molasses residue from the sugar crystalization used as a base for fermentation. This process is also applied to sugar beets.

Starchy materials. For many centuries, it was only feasible to employ local grain crops for liquor production, and, in this way, the basic characteristics of the local distilled beverage were established. Improved transportation removed this restriction, and today economic considerations frequently determine grain selection, with the principal grain used being the one available at the lowest price per unit of fermentable materials.

Cereal grains used for liquor production

Corn (maize) is the most important cereal grain employed; it is produced worldwide. Rye grain, though less efficient in fermentation than corn, is used extensively in whiskey production, primarily for the flavour characteristics it imparts to the final product. It is particularly employed in Canada and the United States. Rice, a widely grown cereal, has limited use in distilled spirits production outside of Asia from India to Japan. Barley grain, probably the first cereal employed for distillation in large quantities, was formerly a major crop throughout Britain, Scotland, Ireland, and western Europe. Wheat, because of its high cost, is used only where corn is in short supply and is then limited to production of grain alcohol for blending or in production of liqueurs. Potatoes have been used in distilled spirits production primarily in central Europe; in the tropics, other starchy roots are employed.

Preparing the mash. *Milling and pressing.* The purpose of milling and pressing operations is to make the starch or sugar more available for enzyme action. Crushing and pressing (grapes and other fruits), milling (cereal grains), or a combination of milling and pressing (sugarcane) are used.

In milling, grains are reduced to a meal to allow wetting of their starch cells. Various types of mills are used. Roller mills, where the grain passes through a series of corrugated rollers, was long the most common type. The grinding action of the rollers is mainly a shearing action. More efficient and economical impact-type mills (such as hammer mills) are now gaining in importance.

After the Industrial Revolution, steam replaced water as the power source for milling. Steam-powered mills were a maze of belts and pulleys. Since the mid-20th century, electricity has been almost the exclusive power source in milling.

Mashing. The purpose of the mashing operation is to (1) mix the proper proportions of grains, (2) increase the availability of the starch for enzyme action, and (3) convert the starches into fermentable sugars.

Mashing is done in a vessel called a mash tub, which is equipped with a means of agitation for mixing and is either jacketed or contains coils for heating and cooling. In mashing, the starch cells of the grain, enclosed in their own protective coatings, are broken to allow wetting and liquefaction of the entire starch mass. The process usually begins with the grain most difficult to treat. When corn is used, the ground meal is wetted at a temperature of approximately 66° C (150° F), and the temperature is then raised to boiling or sometimes higher while under pressure. Temperature is reduced when the starch cells are broken. The grain ranking second in cell resistance (usually rye) is added next. Other starchy substances, such as potatoes, are usually crushed and heated, exploding the starch cells. Temperature of the mash is reduced before ground malt meal, either in dry form or as a water slurry (insoluble mixture), is added. The amylase enzymes in the malt then produce a mixture in which the starches have been converted to fermentable sugars, suitable for utilization by the yeast. The sugars, principally dextrose and maltose, vary in concentration among producers but, generally, are sufficiently concentrated to make a final product ranging from 7 to 9 percent alcohol.

Use of malt

Any germinating cereal grain can be used for malt. In rare cases rye malt is used in making rye whiskey, but because the enzyme activity of malted barley is the highest, barley is used almost exclusively in the distilling industry. Barley malt contains sufficient enzymes to convert approximately 10 times its weight in other unmalted grains. Of the two enzymes—α-amylase and β-amylase—the former is the more important for conversion of other grains. In addition to converting starches from other carbohydrates to sugars, barley malt contains soluble proteins (amino acids), contributing flavour to the distillate secured from fermentation and distillation of grain-malt mixtures. (For a description of the production of malt from germinated barley, see above *Beer: The brewing process.*)

Fermentation. *Yeast and yeast culture.* As mentioned above, yeasts are found throughout the world; more than 8,000 strains of this vegetative microorganism have been classified. Approximately nine or 10 pure strains, with their subclassifications, are used for fermentation of grain mashes; these all belong to the type *Saccharomyces cerevisiae.* Each strain has its own characteristics, imparting its special properties to the distillate derived from its fermentation. A limited number of yeasts are used in the fermentation of wines, from which brandy is distilled. Strains used in the fermentation of grain mashes are also used in fermentation for rum, tequila, and beer production.

In grain-based products, yeast cells are grown in grain mixtures. The preparation of a cooked mash of rye and barley malt is most common. The mash is sterilized, then inoculated with lactic-acid bacteria to increase acidity. (Yeast is more tolerant of higher acidity than many commonly occurring bacteria.) When the desired acidity is reached, the mixture is again sterilized and a pure yeast culture is added. The yeast is grown under controlled conditions until it reaches the optimum point for mixing with the grain mash. In liquid fermentation, as from fruits and sugarcane, the yeast is generally grown in a mixture similar to the one it will be used to ferment; for example, a yeast culture to be used for molasses fermentation is usually grown in molasses.

Fermenting methods. In the fermentation process, simple sugars, including dextrose and maltose, are converted to ethyl alcohol by the action of yeast enzymes. Several intermediate compounds are formed during this complex chemical process before the final ethyl alcohol is obtained.

Yeast functions best in a slightly acid medium, and the prepared grain mash, fruit juice, molasses, or other mixture must be checked for adequate acidity (pH value). If acidity is insufficient, acid or acid-bearing material is added to achieve the necessary adjustment. The previously prepared yeast is then added, and final dilution of the mixture is made. The final concentration of sugars is adjusted so that the yeast fermentation will produce a finished fermented mixture containing between 7 and 9 percent alcohol.

Equipment for fermentation

Commercial fermentation is carried on in large vats. In the past these were open and made of wood, usually cypress. Most plants now use closed stainless steel vats for easier cleaning, and many are equipped with jackets or cooling coils for better temperature control. The time required for completion of fermentation is mainly dependent upon the temperature of the fermenting mash. Normal yeast is most effective in breaking down all of the fermentable sugars at temperatures ranging from 24° to 29° C (75° to 85° F), and, in this range, completion of fermentation requires from 48 to 96 hours. Fermentation at lower temperatures requires longer periods. The mash is ready for distillation upon completion of fermentation. If fermentation is allowed to continue past this period, it will be adversely affected by bacterial action. The ethyl alcohol content will be reduced, and the flavour and aroma of the finished product will be tainted.

Distillation. As mentioned above, the difference in the boiling points of alcohol and water is utilized in distillation to separate these liquids from each other. Basic distillation apparatus consists of three parts: the still or retort, for heating the liquid; the condenser, for cooling the vapours; and the receiver, for collecting the distillate.

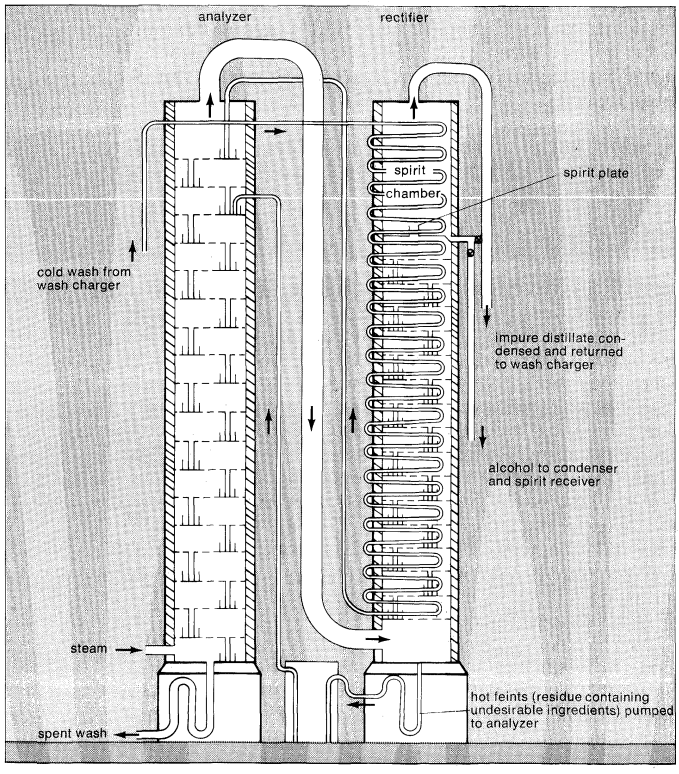
The pot still. The simple pot still is a large enclosed vessel, heated either by direct firing on the bottom or by steam coils within the vessel, with a cylindrical bulb at its top leading to a partially cooled vapour line. The bulb and vapour line separate entrained liquid particles from the vapour on its way to the final condenser. The usual pot-still operation involves a series of two or three pot

stills. Any vapour falling below a predetermined alcoholic content is fed into a second still, and condensed vapour from the second still falling below the required alcoholic content is fed to the third. The condensed vapours of the desired alcoholic content from all three stills are then commingled in a single receiving container.

The pot still, used primarily in Scotland and Ireland for whiskey production and in France for brandies, has had only brief use in distilled spirits production elsewhere and is gradually becoming obsolete. Even in countries in which the pot still has long been used, it has been replaced by continuous distillation for the major portion of alcoholic-liquor production, and its current use is limited to production of flavouring whiskeys and other flavouring ingredients.

The flavour profile of a pot-still product is more complex than that of a continuous-still product of the same alcohol content. This is a result of the different distillation methods. At a given temperature and pressure, vapours over a boiling mixture have a composition that is a function of the vapour pressures of the components of the mixture. In a pot still, the temperature of the fermentation mixture rises as the lower-boiling-temperature alcohol vaporizes. Meanwhile, the alcohol content of the distillate drops as the rising temperature vaporizes more water along with the alcohol. Distillation is allowed to continue until the alcohol content of the distillate falls to a predetemined level. Because of the rising temperature encountered in distilling a single batch, the composition of the first part of the condensate to leave the pot is different from that of the last part. The composition of the final product is the average of the composition of the vapours condensed during the entire run. By contrast, the temperature of the continuous still is held approximately constant throughout the run. This results in a flavour profile that is more uniform.

The continuous still. The continuous still, which came into use in the early 19th century, consists of a tall cylindrical column filled with perforated plates onto which water-rich vapours condense while alcohol-enriched vapours pass through. These plates thus serve as a series of small pot stills, one on top of the other. Live steam, used as the heat source, is fed into the bottom of the still, and the liquid to be distilled is fed near the top. Steam pressure holds the liquid on the plates, and, with any overflow caught by the plate below, the liquid level on each plate is maintained. Use of a sufficient number of plates assures that the concentration of alcohol in the vapour leaving the top of the still will be appropriate for the desired product and that the liquid leaving the bottom has been stripped of any alcohol.



Continuous still for preparing distilled liquors.

Many distillation operations combine column and pot stills. The condensed distillate from the column still is fed to the doubler, a type of pot still heated by closed steam coils, and redistilled.

The rectification still. Rectification is the process of purifying alcohol by repeatedly or fractionally distilling it to remove water and undesirable compounds. As mentioned above, a fermentation mixture primarily contains water and ethyl alcohol and distillation involves increasing the percentage of ethyl alcohol in the mixture. Water vaporizes very easily, however, and, unless care is taken, the distillate of a fermentation mixture will contain unacceptably large quantities of water. The fermentation mixture furthermore contains small quantities of complex constituents that can contribute to the flavour of the product even if they are present only in parts per million. It is important to retain those components that make a positive contribution to the product and to remove those that are unwanted, primarily some organic aldehydes, acids, esters, and higher alcohols. The ones that remain in the product are called congeners, and the congener level is controlled by the particular rectification system and by the system's method of operation.

The multicolumn rectifying system usually consists of three to five columns. The first column is always a preliminary separation column called the beer still, or analyzer. It usually consists of a series of metal plates with holes punched in them and baffles to control the liquid levels on the plates. The product coming from this column is between 55 and 80 percent ethyl alcohol. A 95-percent product can be produced on a two-column system consisting of a beer column and a rectifying column. The bulk of congener removal is accomplished in the rectifier—esters and higher alcohols, for example, being drawn off as side streams. However, a multicolumn system of several specialized rectifiers allows better control of the finished product. An aldehyde column, or purifier, is frequently used to separate these highly volatile low-boiling components, and sometimes ethyl alcohol is recovered in an extractive column and returned to the rectifier.

Three characteristics determine the elimination or retention of flavouring compounds: (1) their boiling points, (2) their solubilities in ethyl alcohol and water, and (3) their specific gravities. Some higher alcohols, for example, are removed on the basis of their solubility and specific gravity. These higher alcohols have limited solubility in water, and their specific gravities are less than that of water. Also, their boiling points are higher than that of ethyl alcohol and lower than that of water. Since they tend to accumulate in the rectifying column at the region where their boiling points cause them to condense, they can be drawn off as a liquid side stream. This side stream also contains a considerable amount of water. The limited solubility in water, plus the lower specific gravities, cause the higher alcohols to float to the top of the alcohol–water mixture, from which they can be removed.

MATURATION, BLENDING, AND PACKAGING

Aging. One method of classifying distilled liquors is as aged or unaged. Vodka, neutral spirits for use in a variety of products, most gins, and some rums and brandies are unaged. Aged products are predominantly whiskeys and most rums and brandies.

The term age refers to the actual duration of storage, while maturity expresses the degree to which chemical changes occur during storage. The maturation of whiskeys falls into two categories, according to whether storage is in new or reused cooperage. New charred, white-oak containers are required by law in the United States for the maturation of products to be called straight bourbon or rye whiskey. These containers, each containing 50 to 55 gallons, are stored in warehouses sometimes having controlled temperature and humidity. Older warehouses are called rick houses because the barrels are stored on sta-

tionary frames called ricks. In many newer houses, barrels are stacked on pallets.

White oak is one of the few woods that can hold liquids while allowing the process of breathing through the pores of the wood. The pore size of the wood is such that small molecules such as water move through the wood more easily than larger molecules such as alcohol. This breathing process is caused by temperature and humidity differences between the liquid in the barrel and the air in the warehouse. Charring the wood makes some of the wood compounds more soluble. As the liquid in the container moves back and forth through the wood, ingredients are extracted and carried back into the container's contents. Maturation also results from the contact of oxygen from the outside air with ingredients in the alcohol mixture. Therefore, maturation during aging consists of the interaction of the original compounds of the distillate, of oxidation reactions, and of the extraction of flavouring compounds from the wood. These factors must be well balanced in the properly matured product. The lower the level of the original congeners, the less wood extract required to achieve a good balance.

Outside the United States, reused cooperage is common. Since used containers have already yielded their initial oak extracts, the resulting product is low in extracted flavouring ingredients, which is desirable in some beverages. This maturation method, typified by Scotch and Irish whiskeys, can be carried on in casks holding up to 132 gallons. These casks have usually had previous use for storage or maturation of other whiskeys or wines and may be reused for many maturation cycles. Maturation in dry warehousing increases the alcoholic content of the liquid in the container, but the more common practice for Scotch and Irish whiskeys of maturation in high humidity warehouses reduces the alcoholic concentration.

The maturation procedure for brandies is similar to that of some whiskeys, but the brandies are usually matured in fairly large casks or oak containers. Most brandies are matured for three to five years, but some remain for as long as 20 to 40 years or even longer.

Rum is usually matured in reused oak containers; high concentrations of oak extracts are not considered desirable. Normal maturation time is two to three years, but rum, generally a blended product, may contain a percentage of older rums.

Most governments specify storage time for various products. The United States requires a two-year storage period for most whiskeys but has no requirement for any pure alcohol or neutral spirits (close to 100 percent alcohol) added to such whiskeys in the production of blended whiskey. Canada requires storage of two years for all distilled spirits. Scotland and England require a three-year storage and Ireland, five years for all products classified as whiskey; there are no requirements for vodka and gin.

Blending. Blending is another method of obtaining a balanced product with precise flavour characteristics. Blended products are composed of one or more highly flavoured components, a high-proof component with a low congener content, a colour adjustment ingredient, and perhaps an additional flavouring material. An example is a blended whiskey, which may contain several whiskeys, a grain spirit distilled at 90 to 95 percent alcohol, caramel colouring, and perhaps a small amount of a flavouring blender (part of which may be sherry or port wine). A blended Scotch consists of several highly flavoured malt whiskeys produced in pot stills and a base whiskey produced from grain in a continuous distillation system.

Packaging. *Bottling.* Distilled spirits react upon exposure to many substances, extracting materials from the container that tend to destroy the liquor aroma and flavour. For this reason, glass, being nonreactive, has been the universal container for packaging alcoholic liquors. (A few products are now packaged in plastic bottles, but these are primarily 50-millilitre miniatures, the light weight of which is particularly suited for use by airlines.) Packaging economics require containers that are standardized in size and shape and that lend themselves to automatic processes.

Early hand methods of filling, labeling, corking, and other operations have been replaced by highly mechanized bottling lines, with bottles cleaned, filled, capped, sealed, labeled, and placed in a shipping container at a rate as high as 400 bottles per minute. This progress became possible with the development of high-strength glass, plastic closures with inert liners, and high-speed machines. Even specialized packaging, long a hand operation, has been replaced by standardization of containers, allowing production on automatic lines.

Mechanized packaging

Designated proof. Spirit strength may be designated in several ways—weight per gallon, percentage by weight, or percentage by volume, all these having reference to absolute (*i.e.,* pure) alcohol and water. There are other standards in common use; *e.g.,* U.S. proof spirit, which is 50 percent by volume alcohol. Each degree of U.S. proof represents 0.5 percent alcohol, so that a liquor having 50 percent alcohol is termed 100 proof. British proof is based on a specific concentration of alcohol, a 50 percent alcoholic content being equivalent to 114.12 U.S. proof. British proof is expressed as degrees over or under proof (that is, over or under 50 percent alcohol), while U.S. proof is expressed in direct proof figures. The metric Gay-Lussac system simply states the percentage by volume of alcohol in a distilled liquor. (F.M.S./A.T.T.)

Soft drinks

Soft drinks are a class of nonalcoholic beverage, usually but not necessarily carbonated, normally containing a natural or artificial sweetening agent, edible acids, natural or artificial flavours, and sometimes juice. Natural flavours are derived from fruits, nuts, berries, roots, herbs, and other plant sources. Coffee, tea, milk, cocoa, and undiluted fruit and vegetable juices are not considered soft drinks.

The term was originated to distinguish the flavoured drinks from hard liquor, or spirits. Soft drinks were recommended as a substitute in the effort to change the hard-drinking habits of early Americans. Indeed, health concerns of modern consumers have led to new categories of soft drinks emphasizing low caloric and sodium content, no caffeine, and "all natural" ingredients.

There are many specialty soft drinks. Mineral waters are very popular in Europe and Latin America. Kava, made from roots of a bushy shrub, *Piper methysticum,* is consumed by the people of Fiji and other Pacific islands. In Cuba people enjoy a carbonated cane juice; its flavour comes from unrefined syrup. In tropical areas, where diets frequently lack sufficient protein, soft drinks containing soybean flour have been marketed. In Egypt carob or locust bean extract is used. In Brazil a soft drink is made using maté as a base. The whey obtained from making buffalo cheese is carbonated and consumed as a soft drink in North Africa. Some eastern Europeans enjoy a drink prepared from fermented stale bread. Honey and orange juice go into a popular drink of Israel.

HISTORY OF SOFT DRINKS

The first marketed soft drinks appeared in the 17th century as a mixture of water and lemon juice sweetened with honey. In 1676 the Compagnie de Limonadiers was formed in Paris and granted a monopoly for the sale of its products. Vendors carried tanks on their backs from which they dispensed cups of lemonade.

Carbonated beverages and waters were developed from European attempts in the 17th century to imitate the popular and naturally effervescent waters of famous springs, with primary interest in their reputed therapeutic values. The effervescent feature of the waters was recognized early as most important. Jan Baptist van Helmont (1577–1644) first used the term gas in his reference to the carbon dioxide content. Gabriel Venel referred to aerated water, confusing the gas with ordinary air. Joseph Black named the gaseous constituent fixed air.

Robert Boyle, the Anglo-Irish scientist who helped found modern chemistry, published his *Short Memoirs for the Natural Experimental History of Mineral Waters* in 1685. It included sections on examining mineral springs, on the properties of the water, on its effects upon human bodies, and, lastly, "of the imitation of natural medicinal waters by chymical and other artificial wayes."

Numerous reports of experiments and investigations were included in the *Philosophical Transactions* of the Royal Society of London in the late 1700s, including the studies of Stephen Hales, Joseph Black, David Macbride, William Brownrigg, Henry Cavendish, Thomas Lane, and others.

First carbonating apparatus

Joseph Priestley is nicknamed "the father of the soft drinks industry" for his experiments on gas obtained from the fermenting vats of a brewery. In 1772 he demonstrated a small carbonating apparatus to the College of Physicians in London, suggesting that, with the aid of a pump, water might be more highly impregnated with fixed air. Antoine Lavoisier in Paris made the same suggestion in 1773.

To Thomas Henry, an apothecary in Manchester, Eng., is attributed the first production of carbonated water, which he made in 12-gallon barrels using an apparatus based on Priestley's. Jacob Schweppe, a jeweler in Geneva, read the papers of Priestley and Lavoisier and determined to make a similar device. By 1794 he was selling his highly carbonated artificial mineral waters to his friends in Geneva; later he started a business in London.

At first, bottled waters were used medicinally, as evidenced in a letter written by English industrialist Matthew Boulton to the philosopher Erasmus Darwin in 1794: "J. Schweppe prepares his mineral waters of three sorts. No. 1 is for common drinking with your dinner. No. 2 is for nephritick patients and No. 3 contains the most alkali given only in more violent cases." By about 1820, improvements in manufacturing processes allowed a much greater output, and bottled water became popular. Mineral salts and flavours were added—ginger in about 1820, lemon in the 1830s, tonic in 1858. In 1886 John Pemberton, a pharmacist in Atlanta, Ga., invented Coca-Cola, the first cola drink.

PRODUCTION

All ingredients used in soft drinks must be of high purity and food grade to obtain a quality beverage. These include the water, carbon dioxide, sugar, acids, juices, and flavours.

Purification of water

Water. Although water is most often taken from a safe municipal supply, it usually is processed further to ensure uniformity of the finished product; the amount of impurities in the municipal supply may vary from time to time. In some bottling plants the water-treatment equipment may simply consist of a sand filter to remove minute solid matter and activated carbon purifier to remove colour, chlorine, and other tastes or odours. In most plants, however, water is treated by a process known as superchlorination and coagulation. There, the water is exposed for two hours to a high concentration of chlorine and to a flocculant, which removes such organisms as plankton (minute plants and animals); it then passes through a sand filter and activated carbon.

Carbon dioxide and carbonation. Carbon dioxide gas gives the beverage its sparkle and tangy taste and prevents spoilage. While it has not been conclusively proved that carbonation offers a direct medical benefit, carbonated beverages are used to alleviate postoperative nausea when no other food can be tolerated, as well as to ensure adequate liquid intake.

Carbon dioxide is supplied to the soft drink manufacturer in either solid form (Dry Ice) or liquid form maintained under approximately 1,200 pounds per square inch (84 kilograms per square centimetre) pressure in heavy steel containers. Lightweight steel containers are used when the liquid carbon dioxide is held under refrigeration. In that case, the internal pressure is about 325 pounds per square inch.

Carbonation (of either the water or the finished beverage mixture) is effected by chilling the liquid and cascading it in thin layers over a series of plates in an enclosure containing carbon dioxide gas under pressure. The amount of gas the water will absorb increases as the pressure is increased and the temperature is decreased.

Flavouring syrup. Flavouring syrup is normally a concentrated solution of a sweetener (sugar or artificial), an acidulant for tartness, flavouring, and a preservative when necessary. The flavouring syrup is made in two steps. First, a "simple syrup" is prepared by making a solution of water and sugar. This simple sugar solution can be treated with carbon and filtered if the sugar quality is poor. All of the other ingredients are then added in a precise order to make up what is called a "finished syrup."

Blending of syrup and water

Finishing. There are two methods for producing a finished product from the flavouring syrup. In the first, the syrup is diluted with water and the product then cooled, carbonated, and bottled. In the second, the maker measures a precise amount of syrup into each bottle, then fills it with carbonated water. In either case, the sugar content (51–60 percent in the syrup) is reduced to 8–13 percent in the finished beverage.

The blending of syrups and mixing with plain or carbonated water, the container washing, and container filling are all done almost entirely by automatic machinery. Returnable bottles are washed in hot alkali solutions for a minimum of five minutes, then rinsed thoroughly. Single-service or "one-trip" containers are generally air-rinsed or rinsed with potable water before filling. Automatic fillers service from 30 to 2,000 containers per minute.

Pasteurizing noncarbonated beverages. Noncarbonated beverages require ingredients and techniques similar to those for carbonated beverages. However, since they lack the protection against spoilage afforded by carbonation, these are usually pasteurized, either in bulk, by continuous flash pasteurization prior to filling, or in the bottle.

Powdered soft drinks. These are made by blending the flavouring material with dry acids, gums, artificial colour, etc. If the sweetener has been included, the consumer need only add the proper amount of plain or carbonated water.

Iced soft drinks. The first iced soft drink consisted of a cup of ice covered with a flavoured syrup. Sophisticated dispensing machines now blend measured quantities of syrup with carbonated or plain water to make the finished beverage. To obtain the soft ice, or slush, the machine reduces the beverage temperature to between −5° and −2° C (22° and 28° F).

PACKAGING AND VENDING

Soft drinks are packaged in glass or plastic bottles, tin-free steel, aluminum, or plastic cans, treated cardboard cartons, foil pouches, or in large stainless steel containers.

Vending of soft drinks had its modest beginning with the use of ice coolers in the early 20th century. Nowadays, most drinks are cooled by electric refrigeration for consumption on the premises. Vending machines dispense soft drinks in cups, cans, or bottles, and restaurants, bars, and hotels use dispensing guns to handle large volume. There are two methods of vending soft drinks in cups. In the "pre-mix" system, the finished beverage is prepared by the soft drink manufacturer and filled into five- or 10-gallon stainless steel tanks. The tanks of beverage are attached to the vending machine where the beverage is cooled and dispensed. In the "post-mix" system the vending machine has its own water and carbon dioxide supply. The water is carbonated as required and is mixed with flavoured syrup as it is dispensed into the cup. (H.E.K./M.J.Pi.)

BIBLIOGRAPHY

Tea: WILLIAM H. UKERS, *All About Tea,* 2 vol. (1935), deals with the historical, technological, scientific, commercial, social, and artistic aspects of tea production and includes extensive bibliographies. C.R. HARLER, *Tea Growing* (1966), *Tea Manufacture* (1963, reprinted 1970), and *The Culture and Marketing of Tea,* 3rd ed. (1964), discuss methods and problems of the tea industry. CLAUD BALD, *Indian Tea: A Textbook on the Culture and Manufacture of Tea,* 7th ed., rev. by C.J. HARRISON (1965), deals with the development of tea production in India, where the modern tea industry originated. T. EDEN, *Tea,* 3rd ed. (1976), discusses tea production in Sri Lanka and the newly developing African tea industry. See also J. WERKHOVEN (comp.), *Tea Processing* (1974), a survey prepared for the Food and Agriculture Organization of the United Nations. (S.Si.)

Coffee: On the history of coffee, see RALPH S. HATTOX, *Coffee and Coffeehouses: The Origins of a Social Beverage in the Medieval Near East* (1985). WILLIAM H. UKERS, *All About Coffee,* 2nd ed. (1935, reissued 1976), offers an excellent view of coffee technology and production. Later sources include M.N. CLIFFORD and K.C. WILLSON (eds.), *Coffee: Botany, Biochemistry, and Production of Beans and Beverage* (1985); R.J. CLARKE and

R. MACRAE (eds.), *Coffee,* 2 vol. (1985–87), on chemistry and technology; MICHAEL SIVETZ and NORMAN W. DESROSIER, *Coffee Technology* (1979), a comprehensive survey of roasted, soluble, and extracted coffees; and C.F. MARSHALL, *The World Coffee Trade: A Guide to the Production, Trading, and Consumption of Coffee* (1983).

(Ed.)

Beer: H.S. CORRAN, *A History of Brewing* (1975), is a well-researched work with references to brewing in Europe (mainly Britain) and in the United States. FRITZ SCHOELLHORN, *Bibliographie des Brauwesens* (1928), continued in F. KUTTER, *Bibliographie des Brauwesens* (1954), list technical references to Latin literature, from the 15th century onward to many German sources, and to similar scientific papers written in English, French, Swedish, Danish, Czech, Russian, Dutch, Norwegian, Italian, and Hungarian. Development of brewing materials and processes to the first quarter of the 20th century is covered in H. LLOYD HIND, *Brewing: Science and Practice,* 2 vol. (1938–40). Later sources include J. DE CLERCK, *A Textbook of Brewing,* 2 vol. (1957–58; originally published in French, 1948), a discussion of worldwide brewing practices and beer analysis; D.E. BRIGGS, *Barley* (1978); and ARTHUR H. COOK (ed.), *Barley and Malt* (1962), studies of the breeding of varieties of barley and of the technical and scientific aspects of malt and malting. A.H. BURGESS, *Hops* (1964), provides the same coverage of hop cultivation and the historical, scientific, and technical aspects of the use of hops in brewing. J.S. HOUGH, *The Biotechnology of Malting and Brewing* (1985), is an introductory account. D.E. BRIGGS *et al., Malting and Brewing Science,* 2nd ed., 2 vol. (1981–82), is a comprehensive text on brewing practice throughout the world and its underlying principles, with an extensive bibliography. J.R.A. POLLOCK (ed.), *Brewing Science,* 3 vol. (1979–87), is a collection of articles by practitioners of brewing, covering brewing science and related technology; and HAROLD M. BRODERICK (ed.), *The Practical Brewer: A Manual for the Brewing Industry,* 2nd ed. (1977), offers information on the American brewing industry and its history.

(T.W.Y.)

Wine: Encyclopaedias and dictionaries on wines and wine production include *Frank Schoonmaker's Encyclopedia of Wine,* rev. and expanded ed., rev. by JULIUS WILE (1978); HUGH JOHNSON, *The World Atlas of Wine: A Complete Guide to the Wines and Spirits of the World,* 3rd rev. ed. (1985), and *Hugh Johnson's Modern Encyclopedia of Wine,* 2nd ed. (1987); and TED GRUDZINSKI, *Winequest, the Wine Dictionary* (1985). Standard works on wine making include M.A. AMERINE *et al., The Technology of Wine Making,* 4th ed. (1980), a treatise on the making of all types of wine; and JEAN RIBERÉAU-GAYON *et al., Sciences et techniques du vin,* 4 vol. (1972–77). Viticulture is discussed in A.J. WINKLER, *General Viticulture,* rev. and enl. ed. (1974), a standard treatise on grape growing.

French wines are discussed in ALEXIS LICHINE and SAMUEL PERKINS, *Alexis Lichine's Guide to the Wines and Vineyards of France,* 3rd ed. (1986); STEVEN SPURRIER, *The Académie du vin Guide to French Wines* (1986); DAVID PEPPERCORN, *Bordeaux* (1982, reprinted 1986); and HUBRECHT DUIJKER, *The Wines of the Loire, Alsace, and Champagne* (1983). For German wines, see GERHARD TROOST, *Technologie des Weines,* 5th rev. ed. (1980); and FRANK SCHOONMAKER, *The Wines of Germany,* 2nd rev. ed., edited by PETER M.F. SICHEL (1983). Wines of other countries are discussed in BURTON ANDERSON, *The Simon and Schuster Pocket Guide to Italian Wines* (1987); JAN READ, MAITE MANJÓN, and HUGH JOHNSON, *The Wine and Food of Spain* (1987); ZOLTÁN HALÁSZ, *The Book of Hungarian Wines* (1981; originally published in Hungarian, 1981); LEN EVANS, *Complete Book of Australian Wine,* 3rd ed. (1978); ANTHONY DIAS BLUE, *American Wine: A Comprehensive Guide* (1985), on the United States; WILLIAM I. KAUFMAN, *Encyclopedia of American Wine, Including Mexico and Canada* (1984); and LEON D. ADAMS, *The Wines of America,* 3rd ed. (1985).

(M.A.A.)

Distilled spirits: Several comprehensive guides include a wealth of information on both wines and distilled spirits: HAROLD J. GROSSMAN, *Grossman's Guide to Wines, Beers, & Spirits,* 7th rev. ed., revised by HARRIET LEMBECK (1983), a popular work prepared for both the industry and the consumer, including an excellent section on both distilled spirits in general and on specific types; and ALEXIS LICHINE *et al., Alexis Lichine's New Encyclopedia of Wines & Spirits,* 5th rev. ed. (1987), a comprehensive encyclopaedia of alcoholic-beverage terminology, with introductory chapters treating the history of distilled spirits and their development and production. For the history of the process, see R.J. FORBES, *A Short History of the Art of Distillation: From the Beginnings Up to the Death of Cellier Blumenthal* (1970). G.G. BIRCH and M.G. LINDLEY (eds.), *Alcoholic Beverages* (1985), is a comprehensive book on production methods.

(F.M.S./A.T.T.)

Soft drinks: JOHN J. RILEY, *A History of the American Soft Drink Industry: Bottled Carbonated Beverages, 1807–1957* (1958, reprinted 1972), studies the evolution of the American flavoured soft drink, European development of simulated effervescent waters in the early 1800s, and early development of the flavoured carbonated beverage in the United States. Later developments can be traced through the history of specific companies: ANNE HOY, *Coca-Cola: The First Hundred Years* (1986); and DOUGLAS A. SIMMONS, *Schweppes, the First 200 Years* (1983). M.B. JACOBS, *Manufacture and Analysis of Carbonated Beverages* (1959), is a detailed treatment. See also L.F. GREEN and H.W. HOUGHTON, *Developments in Soft Drink Technology,* 3 vol. (1978–84).

(H.E.K./M.J.Pi.)

Biblical Literature
and Its Critical Interpretation

Biblical literature, as it is treated in this article, consists of four bodies of written works: the Old Testament writings according to the Hebrew canon; intertestamental works, including the Old Testament Apocrypha; the New Testament writings; and the New Testament Apocrypha.

The Old Testament is a collection of writings that was first compiled and preserved as the sacred books of the ancient Hebrew people. As the Bible of the Hebrews and their Jewish descendants down to the present, these books have been perhaps the most decisive single factor in the preservation of the Jews as a cultural entity and Judaism as a religion. The Old Testament and the New Testament—a body of writings that chronicle the origin and early dissemination of Christianity—constitute the Bible of the Christians.

The literature of the Bible, encompassing the Old and New Testaments and various noncanonical works, has played a special role in the history and culture of the Western world and has itself become the subject of intensive critical study. This field of scholarship, including exegesis (critical interpretation) and hermeneutics (the science of interpretive principles), has assumed an important place in the theologies of Judaism and Christianity. The methods and purposes of exegesis and hermeneutics are treated below. For the cultural and historical contexts in which this literature developed, see JUDAISM; CHRISTIANITY.

The article is divided into the following sections:

Influence and significance

HISTORICAL AND CULTURAL IMPORTANCE

Expression of the Jewish religion in the Old Testament

In Judaism. After the kingdoms of Israel and Judah had fallen, in 722 BCE (before the Common Era, equivalent to BC) and 587/586 BCE, respectively, the Hebrew people outlived defeat, captivity, and the loss of their national independence, largely because they possessed writings that preserved their history and traditions. Many of them did not return to Palestine after their exile. Those who did return did so to rebuild a temple and reconstruct a society that was more nearly a religious community than an independent nation. The religion found expression in the books of the Old Testament: books of the Law (Torah), history, prophecy, and poetry. The survival of the Jewish religion and its subsequent incalculable influence in the history of Western culture are difficult to explain without acknowledgment of the importance of the biblical writings.

When the Temple in Jerusalem was destroyed in 70 CE (Common Era, equivalent to AD), the historical, priestly sacrificial worship centred in it came to an end and was never resumed. But the religion of the Jewish people had by then gone with them into many lands, where it retained its character and vitality because it still drew its nurture from biblical literature. The Bible was with them in their synagogues, where it was read, prayed, and taught. It preserved their identity as a people, inspired their worship, arranged their calendar, permeated their family lives; it shaped their ideals, sustained them in persecution, and touched their intellects. Whatever Jewish talent and genius have contributed to Western civilization is due in no small degree to the influence of the Bible.

In Christianity. The Hebrew Bible is as basic to Christianity as it is to Judaism. Without the Old Testament, the New Testament could not have been written and there could have been no man like Jesus; Christianity could not have been what it became. This has to do with cultural values, basic human values, as much as with religious beliefs. The Genesis stories of prehistoric events and people are a conspicuous example. The Hebrew myths of creation have superseded the racial mythologies of Latin, Germanic, Slavonic, and all other Western peoples. This is

not because they contain historically factual information or scientifically adequate accounts of the universe, the beginning of life, or any other subject of knowledge, but because they furnish a profoundly theological interpretation of the universe and human existence, an intellectual framework of reality large enough to make room for developing philosophies and sciences.

Centrality of the one and only God

This biblical structure of ideas is shared by Jews and Christians. It centres in the one and only God, the Creator of all that exists. All things have their place in this structure of ideas. All mankind is viewed as a unity, with no race existing for itself alone. The Covenant people (*i.e.*, the Hebrews in the Old Testament and Christians in the New Testament) are chosen not to enjoy special privileges but to serve God's will toward all nations. The individual's sacred rights condemn his abuse, exploitation, or neglect by the rich and powerful or by society itself. Widows, orphans, the stranger, the friendless, and the helpless have a special claim. God's will and purpose are viewed as just, loving, and ultimately prevailing. The future is God's, when his rule will be fully established.

The Bible went with the Christian Church into every land in Europe, bearing its witness to God. The church, driven in part by the power of biblical themes, called men to ethical and social responsibility, to a life answerable to God, to love for all men, to sonship in the family of God, and to citizenship in a kingdom yet to be revealed. The Bible thus points to a way of life never yet perfectly embodied in any society in history. Weighing every existing kingdom, government, church, party, and organization, it finds them wanting in that justice, mercy, and love for which they were intended.

MAJOR THEMES AND CHARACTERISTICS

The Bible is the literature of faith, not of scientific observation or historical demonstration. God's existence as a speculative problem has no interest for the biblical writers. What is problematical for them is the human condition and destiny before God.

The great biblical themes are about God, his revealed works of creation, provision, judgment, deliverance, his covenant, and his promises. The Bible sees what happens to mankind in the light of God's nature, righteousness, faithfulness, mercy, and love. The major themes about mankind relate to man's rebellion, his estrangement and perversion. Man's redemption, forgiveness, reconciliation, the gifts of grace, the new life, the coming kingdom, and the final consummation of man's hope are all viewed as the gracious works of God.

The Old Testament contains several types of literature: there are narratives combined with rules and instructions (Torah, or Pentateuch) and anecdotes of Hebrew persons, prophets, priests, kings, and their women (Former Prophets). There is an antiracist love story (Ruth), the story of a woman playing a dangerous game (Esther), and one of a preacher who succeeded too well (Jonah). There is a collection of epigrams and prudential wisdom (Proverbs) and a philosophic view of existence with pessimism and poise (Ecclesiastes). There is poetry of the first rank, devotional poetry in the Psalms, and erotic poetry in the Song of Songs. Lamentations is a poetic elegy, mourning over fallen Jerusalem. Job is dramatic theological dialogue. The books of the great prophets consist mainly of oral addresses in poetic form.

The New Testament also consists of different literary forms. Acts is historical narrative, actually a second volume following Luke. A Gospel is not a history in the ordinary sense but an arrangement of remembered acts and sayings of Jesus retold to win faith in him. There is one apocalypse, Revelation (a work describing the intervention of God in history). But the largest class of New Testament writings is epistolary, the letters of Paul and other Apostles. Originally written to local groups of Christians, they were preserved in the New Testament and given the status of doctrinal and ethical treatises.

INFLUENCES

On Western civilization. The Bible brought its view of God, the universe, and mankind into all the leading Western languages and thus into the intellectual processes of Western man. The Greek translation of the Old Testament made it accessible in the Hellenistic period (*c.* 300 BCE–*c.* 300 CE) and provided a language for the New Testament and for the Christian liturgy and theology of the first three centuries. The Bible in Latin shaped the thought and life of Western people for a thousand years. Bible translation led to the study and literary development of many languages. Luther's translation of the Bible in the 16th century has been called the beginning of modern German. The Authorized Version (English) of 1611 (King James Version) and the others that preceded it caught the English language at the blooming of its first maturity. Since the invention of printing (mid-15th century), the Bible has become more than the translation of an ancient Oriental literature. It has not seemed a foreign book, and it has been the most available, familiar, and dependable source and arbiter of intellectual, moral, and spiritual ideals in the West.

Influence on Western thought and language

Millions of modern people who do not think of themselves as religious live nevertheless with basic presuppositions that underlie the biblical literature. It would be impossible to calculate the effect of such presuppositions on the changing ideas and attitudes of Western people with regard to the nature and purpose of government, social institutions, and economic theories. Theories and ideals usually rest on prior moral assumptions—*i.e.*, on basic judgments of value. In theory, the West has moved from the divine right of kings to the divinely given rights of every citizen, from slavery through serfdom to the intrinsic worth of every person, from freedom to own property to freedom for everyone from the penalties of hopeless poverty. Though there is a wide difference between the ideal and the actual, biblical literature continues to pronounce its judgment and assert that what ought to be can still be.

On the modern secular age. The assumption of many people is that the Bible has lost much of its importance in a secularized world; that is implied whenever the modern period is called the post-Judeo-Christian era. In most ways the label fits. The modern period seems to be a time in which unprecedented numbers of people have discarded traditional beliefs and practices of both Judaism and Christianity. But the influence of biblical literature neither began nor ended with doctrinal propositions or codes of behaviour. Its importance lies not merely in its overtly religious influence but also, and perhaps more decisively, in its pervasive effect on the thinking and feeling processes, the attitudes and sense of values that, whether recognized as biblical or not, still help to make people what they are.

Influence on the arts

The deepest influence of biblical literature may be found in the arts of Western people, their music and, especially, in their best poetry, drama, and creative fiction. Many of the most moving and illuminating interpretations of biblical material—stories, themes, and characters—are made today by novelists, playwrights, and poets who write simply as human beings, not as adherents of any religion. There are two views of the human condition that scholars have attributed to biblical influence and that have become dominant in Western literature.

The first of these is the view that the mystery of existence and destiny is implicit in every man and woman. In contrast to the canons of classical tragedy, a person of any rank or station may experience the extremes of happiness or misery, exaltation or tragedy. An aged Jew of Rembrandt's paintings or an illiterate black woman of Faulkner's novels can reach the height of human dignity. The arts also put down the mighty from their seats and exalt those of low degree. Any man may be Everyman, the symbol of all human possibility.

The second view of the human condition is that the time of encountering all reality is now, and the place is here, in man's workaday activities and contingencies, whatever they may be. To be human is to know one short life in mortal flesh, in which the past and future are dimensions of the present. It is now or never that the choice is made, the offer of the gift of life accepted or declined. Any kingdom there is must be entered at once or lost forever. It is here in the actual situation of work and play, of love and need, and not in some far-off better time and place, that

the crisis is reached and passed, the issue settled, and the record closed.

These views, though here stated in language that has theological overtones, are not confined to adherents of Judaism or Christianity. They are characteristically Western views of the human condition. That they can be put in words reminiscent of the Bible indicates that the representation of man in Western literature is indeed conditioned by biblical literature. (H.G.D./Ed.)

Old Testament canon, texts, and versions

THE CANON

The term canon, from a Hebrew-Greek word meaning a cane or measuring rod, passed into Christian usage as a norm or a rule of faith. The Church Fathers of the 4th century CE first employed it in reference to the definitive, authoritative nature of the body of sacred Scripture.

The Hebrew canon. The Hebrew Bible is often known among Jews as TaNaKh, an acronym derived from the names of its three divisions: Torah (Instruction, or Law, also called the Pentateuch), Nevi'im (Prophets), and Ketuvim (Writings).

Divisions of the Hebrew Bible

The Torah contains five books: Genesis, Exodus, Leviticus, Numbers, and Deuteronomy. The Nevi'im comprise eight books subdivided into the Former Prophets, containing the four historical works, Joshua, Judges, Samuel, and Kings, and the Latter Prophets, the oracular discourses of Isaiah, Jeremiah, Ezekiel, and the Twelve (Minor—*i.e.,* smaller) Prophets—Hosea, Joel, Amos, Obadiah, Jonah, Micah, Nahum, Habakkuk, Zephaniah, Haggai, Zechariah, and Malachi. The Twelve were all formerly written on a single scroll and thus reckoned as one book. The Ketuvim consist of religious poetry and wisdom literature—Psalms, Proverbs, and Job, a collection known as the "Five Megillot" ("scrolls"; *i.e.,* Song of Songs, Ruth, Lamentations, Ecclesiastes, and Esther, which have been grouped together according to the annual cycle of their public reading in the synagogue)—and the books of Daniel, Ezra and Nehemiah, and Chronicles.

The number of books. The number of books in the Hebrew canon is thus 24, referring to the sum of the separate scrolls on which these works were traditionally written in ancient times. This figure is first cited in II Esdras in a passage usually dated *c.* 100 CE and is frequently mentioned in rabbinic (postbiblical) literature, but no authentic tradition exists to explain it. Josephus, a 1st century CE Jewish historian, and some of the Church Fathers, such as Origen (the great 3rd-century Alexandrian theologian), appear to have had a 22-book canon.

English Bibles list 39 books for the Old Testament because of the practice of bisecting Samuel, Kings, and Chronicles, and of counting Ezra, Nehemiah, and the 12 Minor Prophets as separate books.

The tripartite canon. The threefold nature of the Hebrew Bible (the Law, the Prophets, and the Writings) is reflected in the literature of the period of the Second Temple (6th–1st centuries BCE) and soon after it. The earliest reference is that of the Jewish wisdom writer Ben Sira (flourished 180–175 BCE), who speaks of "the law of the Most High . . . the wisdom of all the ancients and . . . prophecies." His grandson (*c.* 132 BCE) in the prologue to Ben Sira's work mentions "the law and the prophets and the others that followed them," the latter also called "the other books of our fathers." The same tripartite division finds expression in II Maccabees, the writings of Philo, a Hellenistic Jewish philosopher, and Josephus, a Hellenistic Jewish historian, as well as in the Gospel According to Luke. The tripartite canon represents the three historic stages in the growth of the canon.

The history of canonization. Because no explicit or reliable traditions concerning the criteria of canonicity, the canonizing authorities, the periods in which they lived, or the procedure adopted have been preserved, no more than a plausible reconstruction of the successive stages involved can be provided. First, it must be observed that sanctity and canonization are not synonymous terms. The first condition must have existed before the second could have been formally conferred. Next, the collection and organization of a number of sacred texts into a canonized corpus (body of writings) is quite a different problem from that of the growth and formation of the individual books themselves.

Criteria of canonicity

No longer are there compelling reasons to assume that the history of the canon must have commenced very late in Israel's history, as was once accepted. The emergence in Mesopotamia, already in the second half of the 2nd millennium BCE, of a standardized body of literature arranged in a more or less fixed order and with some kind of official text, expresses the notion of a canon in its secular sense. Because Babylonian and Assyrian patterns frequently served as the models for imitation throughout the Near East, sacred documents in Israel may well have been carefully stored in temples and palaces, particularly if they were used in connection with the cult or studied in the priestly or wisdom schools. The injunction to deposit the two tables of the Decalogue (Ten Commandments) inside the ark of the covenant and the book of the Torah beside it and the chance find of a book of the Torah in the Temple in 622 BCE tend to confirm the existence of such a practice in Israel.

The Torah. The history of the canonization of the Torah as a book must be distinguished from the process by which the heterogeneous components of the literature as such developed and were accepted as sacred.

The Book of the Chronicles, composed *c.* 400 BCE, frequently refers to the "Torah of Moses" and exhibits a familiarity with all the five books of the Pentateuch. The earliest record of the reading of a "Torah book" is provided by the narrative describing the reformation instituted by King Josiah of Judah in 622 BCE following the fortuitous discovery of a "book of the Torah" during the renovation of the Temple. The reading of the book (probably Deuteronomy), followed by a national covenant ceremony, is generally interpreted as having constituted a formal act of canonization.

Between this date and 400 BCE the only other ceremony of Torah reading is that described in Nehemiah as having taken place on the autumnal New Year festival. The "book of the Torah of Moses" is mentioned and the emphasis is on its instruction and exposition. The Samaritans, the descendants of Israelites intermarried with foreigners in the old northern kingdom that fell in 722 BCE, became hostile to the Judaeans in the time of Ezra and Nehemiah (6th–5th centuries BCE). They would not likely have accepted the Torah, which they did, along with the tradition of its Mosaic origin, if it had only recently been canonized under the authority of their arch-enemies. The final redaction and canonization of the Torah book, therefore, most likely took place during the Babylonian Exile (6th–5th centuries BCE).

The Nevi'im. The model of the Pentateuch probably encouraged the assemblage and ordering of the literature of the prophets. The Exile of the Jews to Bablylonia in 587/586 and the restoration half a century later enhanced the prestige of the prophets as national figures and aroused interest in the written records of their teachings. The canonization of the Nevi'im could not have taken place before the Samaritan schism that occurred during the time of Ezra and Nehemiah, since nothing of the prophetical literature was known to the Samaritans. On the other hand, the prophetic canon must have been closed by the time the Greeks had displaced the Persians as the rulers of Palestine in the late 4th century BCE. The exclusion of Daniel would otherwise be inexplicable, as would also the omission of Chronicles and Ezra–Nehemiah, even though they supplement and continue the narrative of the Former Prophets. Furthermore, the books of the Latter Prophets contain no hint of the downfall of the Persian Empire and the rise of the Greeks, even though the succession of great powers in the East plays a major role in their theological interpretation of history. Their language, too, is entirely free of Grecisms.

Canonization of prophetic writings

These phenomena accord with the traditions of Josephus and rabbinic sources limiting the activities of the literary prophets to the Persian era.

The Ketuvim. That the formation of the Ketuvim as a corpus was not completed until a very late date is ev-

idenced by the absence of a fixed name, or indeed any real name, for the third division of Scripture. Ben Sira refers to "the other books of our fathers," "the rest of the books"; Philo speaks simply of "other writings" and Josephus of "the remaining books." A widespread practice of entitling the entire Scriptures "the Torah and the Prophets" indicates a considerable hiatus between the canonization of the Prophets and the Ketuvim. Greek words are to be found in the Song of Songs and in Daniel, which also refers to the disintegration of the Greek Empire. Ben Sira omits mention of Daniel and Esther. No fragments of Esther have turned up among the biblical scrolls (*e.g.*, the Dead Sea Scrolls) from the Judaean Desert. Rabbinic sources betray some hesitation about Esther and a decided ambivalence about the book of Ben Sira. A third generation Babylonian *amora* (rabbinical interpretive scholar; pl. *amoraim*) actually cites it as "Ketuvim," as opposed to Torah and Prophets, and in the mid-2nd century CE, the need to deny its canonicity and prohibit its reading was still felt. Differences of opinion also are recorded among the *tannaim* (rabbinical scholars of tradition who compiled the Mishna, or Oral Law) and *amoraim* (who created the Talmud, or Gemara) about the canonical status of Proverbs, Song of Songs, Ecclesiastes, and Esther.

All this indicates a prolonged state of fluidity in respect of the canonization of the Ketuvim. A synod at Jabneh (*c.* 100 CE) seems to have ruled on the matter, but it took a generation or two before their decisions came to be unanimously accepted and the Ketuvim regarded as being definitively closed. The destruction of the Jewish state in 70 CE, the breakdown of central authority, and the ever widening Diaspora (collectively, Jews dispersed to foreign lands) all contributed to the urgent necessity of providing a closed and authoritative corpus of sacred Scriptures.

The Samaritan canon. As has been mentioned, the Samaritans accepted the Pentateuch from the Jews. They know of no other section of the Bible, however, and did not expand their Pentateuchal canon even by the inclusion of any strictly Samaritan compositions.

The Alexandrian canon. The Old Testament as it has come down in Greek translation from the Jews of Alexandria via the Christian Church differs in many respects from the Hebrew Scriptures. The books of the second and third divisions have been redistributed and arranged according to categories of literature—history, poetry, wisdom, and prophecy. Esther and Daniel contain supplementary materials, and many noncanonical books, whether of Hebrew or Greek origin, have been interspersed with the canonical works. These extracanonical writings comprise I Esdras, the Wisdom of Solomon, Ecclesiasticus (Ben Sira), Additions to Esther, Judith, Tobit, Baruch, the Epistle of Jeremiah, and additions to Daniel, as listed in the manuscript known as Codex Vaticanus (*c.* 350 CE). The sequence of the books varies, however, in the manuscripts and in the patristic and synodic lists of the Eastern and Western churches, some of which include other books as well, such as I and II Maccabees.

It should be noted that the contents and form of the inferred original Alexandrian Jewish canon cannot be ascertained with certainty because all extant Greek Bibles are of Christian origin. The Jews of Alexandria may themselves have extended the canon they received from Palestine, or they may have inherited their traditions from Palestinian circles in which the additional books had already been regarded as canonical. It is equally possible that the additions to the Hebrew Scriptures in the Greek Bible are of Christian origin.

The canon at Qumrān. In the collection of manuscripts from the Judaean Desert—discovered from the 1940s on—there are no lists of canonical works and no codices (manuscript volumes), only individual scrolls. For these reasons nothing can be known with certainty about the contents and sequence of the canon of the Qumrān sectarians. Since fragments of all the books of the Hebrew Bible (except Esther) have been found, it may be assumed that this reflects the minimum extent of its canon. The situation is complicated by the presence in Qumrān of extracanonical works—some already known from the Apocrypha (so-called hidden books not accepted as canonical by Judaism and the church) and pseudepigrapha (books falsely ascribed to biblical authors) or from the Cairo Geniza (synagogue storeroom), and others entirely new. Some or all of these additional works may have been considered canonical by the members of the sect. It is significant, however, that so far *pesharim* (interpretations) have been found only on books of the traditional Hebrew canon. Still, the great Psalms scroll departs from the received Hebrew text in both sequence and contents. If the Psalms scroll were a canonical Psalter and not a liturgy, then evidence would indeed be forthcoming for the existence of a rival canon at Qumrān.

Canonical and extracanonical writings

The Christian canon. The Christian Church received its Bible from Greek-speaking Jews and found the majority of its early converts in the Hellenistic world. The Greek Bible of Alexandria thus became the official Bible of the Christian community, and the overwhelming number of quotations from the Hebrew Scriptures in the New Testament are derived from it. Whatever the origin of the Apocryphal books in the canon of Alexandria, these became part of the Christian Scriptures, but there seems to have been no unanimity as to their exact canonical status. The New Testament itself does not cite the Apocryphal books directly, but occasional traces of a knowledge of them are to be found. The Apostolic Fathers (late 1st–early 2nd centuries) show extensive familiarity with this literature, but a list of the Old Testament books by Melito, bishop of Sardis in Asia Minor (2nd century), does not include the additional writings of the Greek Bible, and Origen (*c.* 185–*c.* 254) explicitly describes the Old Testament canon as comprising only 22 books.

From the time of Origen on, the Church Fathers who were familiar with Hebrew differentiated, theoretically at least, the Apocryphal books from those of the Old Testament, though they used them freely. In the Syrian East, until the 7th century the Church had only the books of the Hebrew canon with the addition of Ecclesiasticus, or the Wisdom of Jesus the son of Sira (but without Chronicles, Ezra, and Nehemiah). It also incorporated the Wisdom of Solomon, Baruch, the Letter of Jeremiah, and the additions to Daniel. The 6th-century manuscript of the Peshitta (Syriac version) known as Codex Ambrosianus also has III and IV Maccabees, II (sometimes IV) Esdras, and Josephus' *Wars* VII.

Early councils of the African Church held at Hippo (393) and Carthage (397, 419) affirmed the use of the Apocryphal books as Scripture. In the 4th century also, Athanasius, chief theologian of Christian orthodoxy, differentiated "canonical books" from both "those that are read" by Christians only and the "Apocryphal books" rejected alike by Jews and Christians. In the preparation of a standard Latin version, the biblical scholar Jerome (*c.* 347–419/420) separated "canonical books" from "ecclesiastical books" (*i.e.*, the Apocryphal writings), which he regarded as good for spiritual edification but not authoritative Scripture. A contrary view of Augustine (354–430), one of the greatest Western theologians, prevailed, however, and the works remained in the Latin Vulgate version. The *Decretum Gelasianum*, a Latin document of uncertain authorship but recognized as reflecting the views of the Roman Church at the beginning of the 6th century, includes Tobit, Judith, the Wisdom of Solomon, Ecclesiasticus, and I and II Maccabees as biblical.

Use of Apocryphal works in the Middle Ages

Throughout the Middle Ages, the Apocryphal books were generally regarded as Holy Scripture in the Roman and Greek churches, although theoretical doubts were raised from time to time. Thus, in 1333 Nicholas of Lyra, a French Franciscan theologian, had discussed the differences between the Latin Vulgate and the "Hebrew truth." Christian-Jewish polemics, the increasing attention to Hebrew studies, and, finally, the Reformation kept the issue of the Christian canon alive. Protestants denied canonical status to all books not in the Hebrew Bible. The first modern vernacular Bible to segregate the disputed writings was a Dutch version by Jacob van Liesveldt (Antwerp, 1526). Luther's German edition of 1534 did the same thing and entitled them "Apocrypha" for the first time, noting that while they were not in equal esteem with sacred Scriptures they were edifying.

In response to Protestant views, the Roman Catholic Church made its position clear at the Council of Trent (1546) when it dogmatically affirmed that the entire Latin Vulgate enjoyed equal canonical status. This doctrine was confirmed by the Vatican Council of 1870. In the Greek Church, the Synod of Jerusalem (1672) had expressly designated as canonical several Apocryphal works. In the 19th century, however, Russian Orthodox theologians agreed to exclude these works from the Holy Scriptures.

The history of the Old Testament canon in the English Church has generally reflected a more restrictive viewpoint. Even though the Wycliffite Bible (14th century) included the Apocrypha, its preface made it clear that it accepted Jerome's judgment. The translation made by the English bishop Miles Coverdale (1535) was the first English version to segregate these books, but it did place Baruch after Jeremiah. Article VI of the Thirty-nine Articles of religion of the Church of England (1562) explicitly denied their value for the establishment of doctrine, although it admitted that they should be read for their didactic worth. The first Bible in English to exclude the Apocrypha was the Geneva Bible of 1599. The King James Version of 1611 placed it between the Old and New Testaments. In 1615 Archbishop George Abbot forbade the issuance of Bibles without the Apocrypha, but editions of the King James Version from 1630 on often omitted it from the bound copies. The Geneva Bible edition of 1640 was probably the first to be intentionally printed in England without the Apocrypha, followed in 1642 by the King James Version. In 1644 the Long Parliament actually forbade the public reading of these books, and three years later the Westminster Confession of the Presbyterians decreed them to be no part of the canon. The British and Foreign Bible Society in 1827 resolved never to print or circulate copies containing the Apocrypha. Most English Protestant Bibles in the 20th century have omitted the disputed books or have them as a separate volume, except in library editions, in which they are included with the Old and New Testaments.

TEXTS AND VERSIONS

Textual criticism: manuscript problems. The text of the Hebrew printed Bible consists of consonants, vowel signs, and cantillation (musical or tonal) marks. The two latter components are the product of the school of Masoretes (Traditionalists) that flourished in Tiberias (in Palestine) between the 7th and 9th centuries CE. The history of the bare consonantal text stretches back into hoary antiquity and can be only partially traced.

Masoretic signs and marks

The earliest printed editions of the Hebrew Bible derive from the last quarter of the 15th century and the first quarter of the 16th century. The oldest Masoretic codices stem from the end of the 9th century and the beginning of the 10th. A comparison of the two shows that no textual developments took place during the intervening 600 years. A single standardized recension enjoyed an absolute monopoly and was transmitted by the scribes with amazing fidelity. Not one of the medieval Hebrew manuscripts and none of the thousands of fragments preserved in the Cairo Geniza (synagogue storeroom) contains departures of any real significance from the received text.

This situation, however, was a relatively late development; there is much evidence for the existence of a period when more than one Hebrew text-form of a given book was current. In fact, both the variety of witnesses and the degree of textual divergence between them increase in proportion to their antiquity.

No single explanation can satisfactorily account for this phenomenon. In the case of some biblical literature, there exists the real possibility, though it cannot be proven, that it must have endured a long period of oral transmission before its committal to writing. In the interval, the material might well have undergone abridgement, amplification, and alteration at the hands of transmitters so that not only would the original have been transformed, but the process of transmission would have engendered more than one recension from the very beginning of its written, literary career.

The problem is complicated further by the great difference in time between the autograph (original writing) of a biblical work, even when it assumed written form from its inception, and its oldest extant exemplars. In some instances, this may amount to well over a thousand years of scribal activity. Whatever the interval, the possibility of inadvertent and deliberate change, something that affects all manuscript copying, was always present.

The evidence that such, indeed, took place is rich and varied. First there are numerous divergences between the many passages duplicated within the Hebrew Bible itself—*e.g.*, the parallels between Samuel–Kings and Chronicles. Then there are the citations of the Old Testament to be found in the books of the Apocrypha and apocalyptic literature (works describing the intervention of God in history in cryptic terms), in the works of Philo and Josephus, in the New Testament, and in rabbinic and patristic (early Church Fathers) literature. There are also rabbinic traditions about the text-critical activities of the scribes (*soferim*) in Second Temple times. These tell of divergent readings in Temple scrolls of the Pentateuch, of official "book correctors" in Jerusalem, of textual emendations on the part of scribes, and of the utilization of sigla (signs or abbreviations) for marking suspect readings and disarranged verses. The Samaritan Pentateuch and the pre-Masoretic versions of the Old Testament made directly from Hebrew originals are all replete with divergences from current Masoretic Bibles. Finally, the scrolls from the Judaean Desert, especially those from the caves of Qumrān, have provided, at least, illustrations of many of the scribal processes by which deviant texts came into being. The variants and their respective causes may be classified as follows: aurally conditioned, visual in origin, exegetical, and deliberate.

Types and causes of variants

Problems resulting from aural conditioning. Aural conditioning would result from a mishearing of similar sounding consonants when a text is dictated to the copyist. A negative particle *lo'*, for example, could be confused with the prepositional *lo,* "to him," or a guttural *het* with spirant *kaf* so that *aḥ* "brother" might be written for *akh* "surely."

Problems visual in origin. The confusion of graphically similar letters, whether in the paleo-Hebrew or Aramaic script, is another cause for variations. Thus, the prepositions *bet* ("in") and *kaf* ("like") are interchanged in the Masoretic and Dead Sea Scroll texts of Isaiah.

The order of letters also might be inverted. Such metathesis, as it is called, appears in Psalms, in which *qirbam* ("their inward thoughts") stands for *qibram* ("their grave").

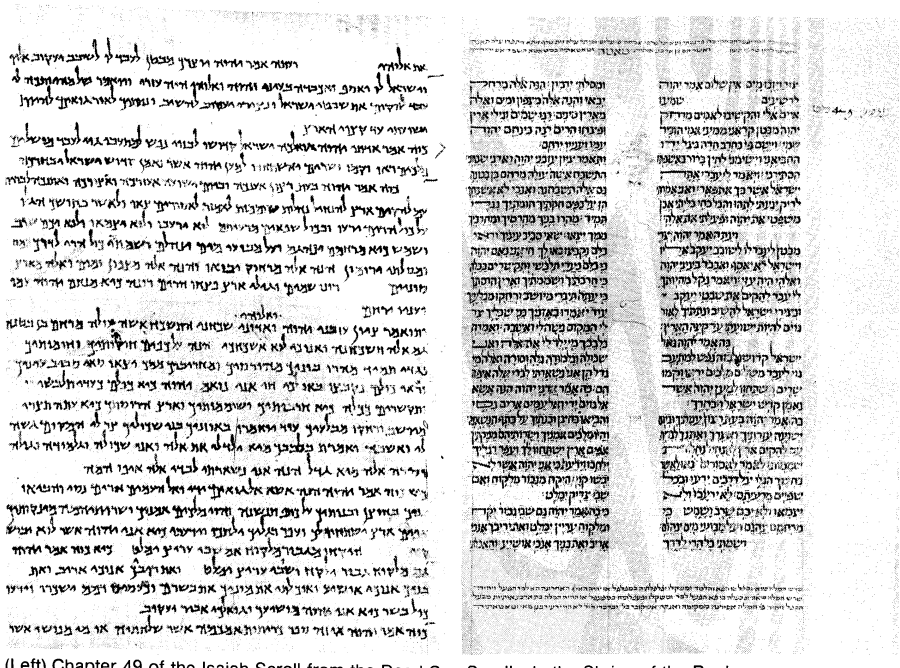
Dittography, or the inadvertent duplication of one or more letters or words, also occurs, as, for example, in the Dead Sea Scroll text of Isaiah and in the Masoretic text of Ezekiel.

Haplography, or the accidental omission of a letter or word that occurs twice in close proximity, can be found, for example, in the Dead Sea Scroll text of Isaiah.

Homoeoteleuton occurs when two separate phrases or lines have identical endings and the copyist's eye slips from one to the other and omits the intervening words. A comparison of the Masoretic text I Samuel, chapter 14 verse 41, with the Septuagint and the Vulgate versions clearly identifies such an aberration.

Exegetical problems. This third category does not involve any consonantal alteration but results solely from the different possibilities inherent in the consonantal spelling. Thus, the lack of vowel signs may permit the word *DBR* to be read as a verb *DiBeR* ("he spoke," as in the Masoretic text of Hosea) or as a noun *DeBaR* ("the word of," as in the Septuagint). The absence of word dividers could lead to different divisions of the consonants. Thus, *BBQRYM* in Amos could be understood as either *BaBeQaRYM* ("with oxen," as in the Masoretic text) or as *BaBaQaR YaM* ("the sea with an ox"). The incorrect solution by later copyists of abbreviations is another source of error. That such occurred is proved by a comparison of the Hebrew text with the Septuagint version in, for example, II Samuel, chapter 1 verse 12; Ezekiel, chapter 12 verse 23; and Amos, chapter 3 verse 9.

Deliberate changes. Apart from mechanical alterations of a text, many variants must have been consciously in-



(Left) Chapter 49 of the Isaiah Scroll from the Dead Sea Scrolls. In the Shrine of the Book, D. Samuel and Jeane H. Gottesman Centre for Biblical Manuscripts, the Israel Museum, Jerusalem. (Right) Hebrew Masoretic text of chapter 49 from the Book of Isaiah, 1341. In the Jewish National and University Library, Jerusalem.

By courtesy of (left) the Shrine of the Book, the Israel Museum, Jerusalem, (right) Jewish National and University Library, Jerusalem

troduced by scribes, some by way of glossing—*i.e.,* the insertion of a more common word to explain a rare one—and others by explanatory comments incorporated into the text. Furthermore, a scribe who had before him two manuscripts of a single work containing variant readings, and unable to decide between them, might incorporate both readings into his scroll and thus create a conflate text.

Textual criticism: scholarly problems. The situation so far described poses two major scholarly problems. The first involves the history of the Hebrew text, the second deals with attempts to reconstruct its "original" form.

As to when and how a single text type gained hegemony and then displaced all others, it is clear that the early and widespread public reading of the Scriptures in the synagogues of Palestine, Alexandria, and Babylon was bound to lead to a heightened sensitivity of the idea of a "correct" text and to give prestige to the particular text form selected for reading. Also, the natural conservatism of ritual would tend to perpetuate the form of such a text. The *Letter of Aristeas,* a document derived from the middle of the 2nd century BCE that describes the origin of the Septuagint, recognizes the distinction between carelessly copied scrolls of the Pentateuch and an authoritative Temple scroll in the hands of the high priest in Jerusalem. The Rabbinic traditions (see above) about the textual criticism of Temple-based scribes actually reflect a movement towards the final stabilization of the text in the Second Temple period. Josephus, writing not long after 70 CE, boasts of the existence of a long-standing fixed text of the Jewish Scriptures. The loss of national independence and the destruction of the spiritual centre of Jewry in 70, accompanied by an ever-widening Diaspora and the Christian schism within Judaism, all made the exclusive dissemination of a single authoritative text a vitally needed cohesive force. The text type later known as Masoretic is already well represented at pre-Christian Qumrān. Scrolls from Wādī al-Murabba'at, Naḥal Ze'elim, and Masada from the 2nd century CE are practically identical with the received text that by then had gained victory over all its rivals.

Reconstruction of an original text

In regard to an attempt to recover the original text of a biblical passage—especially an unintelligible one—in the light of variants among different versions and manuscripts and known causes of corruption, it should be understood that all reconstruction must necessarily be conjectural and perforce tentative because of the irretrievable loss of the original edition. But not all textual difficulties need presuppose underlying mutilation. The Hebrew Bible represents but a small portion of the literature of ancient Israel and, hence, a limited segment of the language. A textual problem may be the product of present limited knowledge of ancient Hebrew, because scholars might be dealing with dialectic phenomena or foreign loan-words. Comparative Semitic linguistic studies have yielded hitherto unrecognized features of grammar, syntax, and lexicography that have often eliminated the need for emendation. Furthermore, each version, indeed each biblical book within it, has its own history, and the translation techniques and stylistic characteristics must be examined and taken into account. Finally, the number of manuscripts that attest to a certain reading is of less importance than the weight given to a specific manuscript.

None of this means that a Hebrew manuscript, an ancient version, or a conjectural emendation cannot yield a reading superior to that in the received Hebrew text. It does mean, however, that these tools have to be employed with great caution and proper methodology.

Texts and manuscripts. *Sources of the Septuagint.* A Greek translation of the Old Testament, known as the Septuagint because there allegedly were 70 or 72 translators, six from each of the 12 tribes of Israel, and designated LXX, is a composite of the work of many translators labouring for well over 100 years. It was made directly from Hebrew originals that frequently differed considerably from the present Masoretic text. Apart from other limitations attendant upon the use of a translation for such purposes, the identification of the parent text used by the Greek translators is still an unsettled question. The Pentateuch of the Septuagint manifests a basic coincidence with the Masoretic text. The Qumrān scrolls have now proven that the Septuagint book of Samuel–Kings goes back to an old Palestinian text tradition that must be earlier than the 4th century BCE, and from the same source comes a short Hebrew recension of Jeremiah that probably underlies the Greek.

The Samaritan Pentateuch. The importance of the recension known as the Samaritan Pentateuch lies in the fact that it constitutes an independent Hebrew witness to the text written in a late and developed form of the paleo-Hebrew script. Some of the Exodus fragments from Qumrān demonstrate that it has close affinities with a pre-Christian Palestinian text type and testify to the faithfulness with which it has been preserved. It contains about

6,000 variants from the Masoretic text, of which nearly a third agree with the Septuagint. Only a minority, however, are genuine variants, most being dogmatic, exegetical, grammatical, or merely orthographic in character.

The Samaritan Pentateuch first became known in the West through a manuscript secured in Damascus in 1616 by Pietro della Valle, an Italian traveler. It was published in the Paris (1628–45) and London Polyglots (1654–57), written in several languages in comparative columns. Many manuscripts of the Samaritan Pentateuch are now available. The Avisha' Scroll, the sacred copy of the Samaritans, has recently been photographed and critically examined. Only Numbers chapter 35 to Deuteronomy chapter 34 appears to be very old, the rest stemming from the 14th century. A new, definitive edition of the Samaritan Pentateuch is being prepared in Madrid by F. Pérez Castro.

The Qumrān texts and other scrolls. Until the discovery of the Judaean Desert scrolls, the only pre-medieval fragment of the Hebrew Bible known to scholars was the Nash Papyrus (*c.* 150 BCE) from Egypt containing the Decalogue and Deuteronomy. Now, however, fragments of about 180 different manuscripts of biblical books are available. Their dates vary between the 3rd century BCE and the 2nd century CE, and all but 10 stem from the caves of Qumrān. All are written on either leather or papyrus in columns and on one side only.

Textual significance of the Qumrān scrolls

The most important manuscripts from what is now identified as Cave 1 of Qumrān are a practically complete Isaiah scroll (1QIsa[a]), dated *c.* 100–75 BCE, and another very fragmentary manuscript (1QIsa[b]) of the same book. The first contains many variants from the Masoretic text in both orthography and text; the second is very close to the Masoretic type and contains few genuine variants. The richest hoard comes from Cave 4 and includes fragments of five copies of Genesis, eight of Exodus, one of Leviticus, 14 of Deuteronomy, two of Joshua, three of Samuel, 12 of Isaiah, four of Jeremiah, eight of the Minor Prophets, one of Proverbs, and three of Daniel. Cave 11 yielded a Psalter containing the last third of the book in a form different from that of the Masoretic text, as well as a manuscript of Leviticus.

The importance of the Qumrān scrolls cannot be exaggerated. Their great antiquity brings them close to the Old Testament period itself—from as early as 250–200 BCE. For the first time, Hebrew variant texts are extant and all known major text types are present. Some are close to the Septuagint, others to the Samaritan. On the other hand, many of the scrolls are practically identical with the Masoretic text, which thus takes this recension back in history to pre-Christian times. Several texts in the paleo-Hebrew script show that this script continued to be used side by side with the Aramaic script for a long time.

Of quite a different order are scrolls from other areas of the Judaean Desert. All of these are practically identical with the received text. This applies to fragments of Leviticus, Deuteronomy, Ezekiel, and Psalms discovered at Masada (the Jewish fortress destroyed by the Romans in CE 73), as well as to the finds at Wādī al-Murabba'at, the latest date of which is CE 135. Here were found fragments of Genesis, Exodus, Leviticus, and Isaiah in addition to the substantially preserved Minor Prophets scroll. Variants from the Masoretic text are negligible. The same phenomenon characterizes the fragments of Numbers found at Naḥal Ḥever.

Masoretic texts. No biblical manuscripts have survived from the six centuries that separate the latest of the Judaean Desert scrolls from the earliest of the Masoretic period. A "Codex Mugah," frequently referred to as an authority in the early 10th century, and the "Codex Hilleli," said to have been written *c.* 600 by Rabbi Hillel ben Moses ben Hillel, have both vanished.

The earliest extant Hebrew Bible codex is the Cairo Prophets written and punctuated by Moses ben Asher in Tiberias (in Palestine) in 895. Next in age is the Leningrad Codex of the Latter Prophets dated to 916, which was not originally the work of Ben Asher, but its Babylonian pointing—*i.e.*, vowel signs used for pronunciation purposes—was brought into line with the Tiberian Masoretic system.

Production of the Aleppo Codex

The outstanding event in the history of that system was the production of the model so-called Aleppo Codex, now in Jerusalem. Written by Solomon ben Buya'a, it was corrected, punctuated, and furnished with a Masoretic apparatus by Aaron ben Moses ben Asher *c.* 930. Originally containing the entire Old Testament in about 380 folios, of which 294 are extant, the Aleppo Codex remains the only known true representative of Aaron ben Asher's text and the most important witness to that particular Masoretic tradition that achieved hegemony throughout Jewry.

Two other notable manuscripts based on Aaron's system are the manuscript designated as BM or. 4445, which contains most of the Pentateuch and which utilized a Masora (text tradition) *c.* 950, and the Leningrad complete Old Testament designated MSB 19a of 1008. Codex Reuchliana of the Prophets, written in 1105, now in Karlsruhe (Germany), represents the system of Moses ben David ben Naphtali, which was more faithful to that of Moses ben Asher.

Collations of the Masoretic materials. The earliest extant attempt at collating the differences between the Ben Asher and Ben Naphtali Masoretic traditions was made by Mishael ben Uzziel in his *Kitāb al-Ḥulaf* (before 1050). A vast amount of Masoretic information, drawn chiefly from Spanish manuscripts, is to be found in the text-critical commentary known as *Minhath Shai,* by Solomon Jedidiah Norzi, completed in 1626 and printed in the Mantua Bible of 1742. Benjamin Kennicott collected the variants of 615 manuscripts and 52 printed editions (2 vol., 1776–80, Oxford). Giovanni Bernado De Rossi published his additional collections of 731 manuscripts and 300 prints (4 vol., 1784–88, Parma), and C.D. Ginsburg did the same for 70 manuscripts, largely from the British Museum, and 17 early printed editions (3 vol. in 4, 1908–26, London).

Printed editions. Until 1488, only separate parts of the Hebrew Bible had been printed, all with rabbinic commentaries. The earliest was the Psalms (1477), followed by the Pentateuch (1482), the Prophets (1485/86), and the Hagiographa (1486/87), all printed in Italy.

The first edition of the entire Hebrew Bible was printed at Soncino (in Italy) in 1488 with punctuation and accents, but without any commentary. The second complete Bible was printed in Naples in 1491/93 and the third in Brescia in 1494. All these editions were the work of Jews. The first Christian production was a magnificent Complutensian Polyglot (under the direction of Cardinal Francisco Jiménez of Spain) in six volumes, four of which contained the Hebrew Bible and Greek and Latin translations together with the Aramaic rendering (Targum) of the Pentateuch that has been ascribed to Onkelos. Printed at Alcala (1514–17) and circulated about 1522, this Bible proved to be a turning point in the study of the Hebrew text in western Europe.

The first rabbinic Bible—*i.e.*, the Hebrew text furnished with full vowel points and accents, accompanied by the Aramaic Targums and the major medieval Jewish commentaries—was edited by Felix Pratensis and published by Daniel Bomberg (Venice, 1516/17). The second edition, edited by Jacob ben Hayyim ibn Adonijah and issued by Bomberg in four volumes (Venice, 1524/25), became the prototype of future Hebrew Bibles down to the 20th century. It contained a vast text-critical apparatus of Masoretic notes never since equalled in any edition. Unfortunately, Ben Hayyim had made use of late manuscripts and the text and notes are eclectic.

In London, Christian David Ginsburg, an emigrant Polish Jew and Christian convert, produced a critical edition of the complete Hebrew Bible (1894, 1908, 1926) revised according to the Masora and early prints with variant readings from manuscripts and ancient versions. It was soon displaced by the *Biblica Hebraica* (1906, 1912) by Rudolf Kittel and Paul Kahle, two German biblical scholars. The third edition of this work, completed by Albrecht Alt and Otto Eissfeldt (Stuttgart, 1937), finally abandoned Ben Hayyim's text, substituting that of the Leningrad Codex (B 19a). It has a dual critical apparatus with textual emendations separated from the manuscript and versional variants. Since 1957 variants from the so-called Judaean Desert scrolls have been included. In progress at the He-

brew University of Jerusalem in the early 1970s was the preparation of a new text of the entire Hebrew Bible based on the Aleppo Codex to include all its own Masoretic notes together with textual differences found in all pertinent sources. A sample edition of the Book of Isaiah appeared in 1965.

Early versions. *The Aramaic Targums.* In the course of the 5th and 6th centuries BCE, Aramaic became the official language of the Persian Empire. In the succeeding centuries it was used as the vernacular over a wide area and was increasingly spoken by the postexilic Jewish communities of Palestine and elsewhere in the Diaspora. In response to liturgical needs, the institution of a *turgeman* (or *meturgeman,* "translator"), arose in the synagogues. These men translated the Torah and prophetic lectionaries into Aramaic. The rendering remained for long solely an oral, impromptu exercise, but gradually, by dint of repetition, certain verbal forms and phrases became fixed and eventually committed to writing.

The Babylonian and Palestinian Targums

There are several Targums (translations) of the Pentateuch. The Babylonian Targum is known as "Onkelos," named after its reputed author. The Targum is Palestinian in origin, but it was early transferred to Babylon where it was revised and achieved great authority. At a later date, probably not before the 9th century CE, it was re-exported to Palestine to displace other, local, Targums. On the whole, Onkelos is quite literal, but it shows a tendency to obscure expressions attributing human form and feelings to God. It also usually faithfully reflects rabbinic exegesis.

The most famous of the Palestinian Targums is that popularly known as "Jonathan," a name derived from a 14th-century scribal mistake that solved a manuscript abbreviation "TJ" as "Targum Jonathan" instead of "Targum Jerusalem." In contrast with two other Targums, which are highly fragmentary (Jerusalem II and III), Pseudo-Jonathan (or Jerusalem I) is virtually complete. It is a composite of the Old Palestinian Targum and an early version of Onkelos with an admixture of material from diverse periods. It contains much rabbinic material as well as homiletic and didactic amplifications. There is evidence of great antiquity, but also much late material, indicating that Pseudo-Jonathan could not have received its present form before the Islāmic period.

Another extant Aramaic version is the Targum to the Samaritan Pentateuch. It is less literal than the Jewish Targums and its text was never officially fixed.

The Targum to the Prophets also originated in Palestine and received its final editing in Babylonia. It is ascribed to Jonathan ben Uzziel, a pupil of Hillel, the famous 1st century BCE–1st century CE rabbinic sage, though it is in fact a composite work of varying ages. In its present form it discloses a dependence on Onkelos, though it is less literal.

The Aramaic renderings of the Hagiographa are relatively late productions, none of them antedating the 5th century CE.

The Septuagint (LXX). The story of the Greek translation of the Pentateuch is told in the *Letter of Aristeas,* which purports to be a contemporary document written by Aristeas, a Greek official at the Egyptian court of Ptolemy II Philadelphus (285–246 BCE). It recounts how the law of the Jews was translated into Greek by Jewish scholars sent from Jerusalem at the request of the king.

This narrative, repeated in one form or another by Philo and rabbinic sources, is full of inaccuracies that prove that the author was an Alexandrian Jew writing well after the events he described had taken place. The Septuagint Pentateuch, which is all that is discussed, does, however, constitute an independent corpus within the Greek Bible, and it was probably first translated as a unit by a company of scholars in Alexandria about the middle of the 3rd century BCE.

Influence of the Septuagint

The Septuagint, as the entire Greek Bible came to be called, has a long and complex history and took well over a century to be completed. It is for this reason not a unified or consistent translation. The Septuagint became the instrument whereby the basic teachings of Judaism were mediated to the pagan world and it became an indispensable factor in the spread of Christianity.

The adoption of the Septuagint as the Bible of the Christians naturally engendered suspicion on the part of Jews. In addition, the emergence of a single authoritative text type after the destruction of the Temple made the great differences between it and the Septuagint increasingly intolerable, and the need for a Greek translation based upon the current Hebrew text in circulation was felt.

The version of Aquila. About 130 CE, Aquila, a convert to Judaism from Pontus in Asia Minor, translated the Hebrew Bible into Greek under the supervision of Rabbi Akiba. Executed with slavish literalness, it attempted to reproduce the most minute detail of the original, even to the extent of coining derivations from Greek roots to correspond to Hebrew usage. Little of it has survived, however, except in quotations, fragments of the Hexapla (see *Origen's Hexapla,* below), and palimpsests (parchments erased and used again) from the Cairo Geniza.

The revision of Theodotion. A second revision of the Greek text was made by Theodotion (of unknown origins) late in the 2nd century, though it is not entirely clear whether it was the Septuagint or some other Greek version that underlay his revision. The new rendering was characterized by a tendency toward verbal consistency and much transliteration of Hebrew words.

The translation of Symmachus. Still another Greek translation was made toward the end of the same century by Symmachus, an otherwise unknown scholar, who made use of his predecessors. His influence was small despite the superior elegance of his work. Jerome did utilize Symmachus for his Vulgate, but other than that, his translation is known largely through fragments of the Hexapla.

Purpose of the Hexapla

Origen's Hexapla. The multiplication of versions doubtless proved to be a source of increasing confusion in the 3rd century. This situation the Alexandrian theologian Origen, working at Caesarea between 230 and 240 CE, sought to remedy. In his Hexapla ("six-fold") he presented, in parallel vertical columns, the Hebrew text, the same in Greek letters, and the versions of Aquila, Symmachus, the Septuagint, and Theodotion, in that order. In the case of some books, Psalms for instance, three more columns were added. The Hexapla serves as an important guide to Palestinian pre-Masoretic pronunciation of the language. The main interest of Origen lay in the fifth column, the Septuagint, which he edited on the basis of the Hebrew. He used the obels (— or ÷) and asterisk (*) to mark respectively words found in the Greek text but not in the Hebrew and vice versa.

The Hexapla was a work of such magnitude that it is unlikely to have been copied as a whole. Origen himself produced an abbreviated edition, the Tetrapla, containing only the last four columns. The original manuscript of the Hexapla is known to have been extant as late as *c.* 600 CE. Today it survives only in fragments.

Manuscripts and printed editions of the Septuagint. The manuscripts are conveniently classified by papyri uncials (capital letters) and minuscules (cursive script). The papyri fragments run into the hundreds, of varying sizes and importance, ranging from the formative period of the Septuagint through the middle of the 7th century. Two pre-Christian fragments of Deuteronomy from Egypt are of outstanding significance. Although not written on papyrus but on parchment or leather, the fragments from Qumrān of Exodus, Leviticus, and Numbers, and the leather scroll of the Minor Prophets from Naḥal Ḥever from the first pre-Christian and post-Christian centuries, deserve special mention among the earliest extant. The most important papyri are those of the Chester Beatty collection, which contains parts of 11 codices preserving fragments of nine Old Testament books. Their dates vary between the 2nd and 4th centuries. During the next 300 years papyri texts multiplied rapidly, and remnants of about 200 are known.

Uncial and minuscule codices

The uncials are all codices written on vellum between the 4th and 10th centuries. The most outstanding are Vaticanus, which is an almost complete 4th-century Old Testament, Sinaiticus, of the same period but less complete, and the practically complete 5th-century Alexandrinus. These three originally contained both Testaments. Many others were partial manuscripts from the beginning. One of the most valuable of these is the Codex Marchalianus of the Prophets written in the 6th century.

(Left) Coptic papyrus of the Gospel According to John, 4th century. In the library of the British and Foreign Bible Society, London. (Right) Illustrated text from the earliest known Ethiopic Bible. In the Bibliothèque Nationale, Paris.

By courtesy of the (left) British and Foreign Bible Society, London, (right) Bibliotheque Nationale, Paris

The minuscule codices begin to appear in the 9th century. From the 11th to the 16th century they are the only ones found, and nearly 1,500 have been recorded.

The first printed Septuagint was that of the Complutensian Polyglot (1514–17). Since it was not released until 1522, however, the 1518 Aldine Venice edition actually was available first. The standard edition until modern times was that of Pope Sixtus V, 1587. In the 19th and 20th centuries several critical editions have been printed.

Coptic versions. The spread of Christianity among the non-Greek speaking peasant communities of Egypt necessitated the translation of the Scriptures into the native tongue (Coptic). These versions may be considered to be wholly Christian in origin and largely based on the Greek Bible. They also display certain affinities with the Old Latin. Nothing certain is known about the Coptic translations except that they probably antedate the earliest known manuscripts from the end of the 3rd and the beginning of the 4th centuries CE.

The Armenian version. The Armenian version is an expression of a nationalist movement that brought about a separation from the rest of the Church (mid-5th century), the discontinuance of Syriac in Greek worship, and the invention of a national alphabet by St. Mesrob, also called Mashtots (*c.* 361–439/440). According to tradition, St. Mesrob first translated Proverbs from the Syriac. Existing manuscripts of the official Armenian recension, however, are based on the Hexaplaric Septuagint, though they show some Peshitta (Syriac version) influence. The Armenian Bible is noted for its beauty and accuracy.

The Georgian version. According to Armenian tradition, the Georgian version was also the work of Mesrob, but the Psalter, the oldest part of the Georgian Old Testament, is probably not earlier than the 5th century. Some manuscripts were based upon Greek versions, others upon the Armenian.

The Ethiopic version. The Ethiopic version poses special problems. The earliest Bible probably was based on Greek versions, after Ethiopia had been converted to Christianity during the 4th and 5th centuries. The earliest existing manuscripts, however, belong to the 13th century. Most manuscripts from the 14th century on seem to reflect Arabic or Coptic influence, and it is not certain whether these represent the original translation or later ones. Many readings agree with the Hebrew against the Septuagint, which may have been caused by a Hexaplaric influence.

Special problems of the Ethiopic version

The Gothic version. The Gothic version was produced in the mid-4th century by Ulfilas, a Christian missionary who also invented the Gothic alphabet. It constitutes practically all that is left of Gothic literature. The translation of the Old Testament has entirely disappeared except for fragments of Ezra and Nehemiah. Though a Greek base is certain, some scholars deny the attribution of these remnants to Ulfilas.

The Old Latin version. The existence of a Latin translation can be attested in North Africa and southern Gaul as early as the second half of the 2nd century CE, and in Rome at the beginning of the following century. Its origins may possibly be attributed to a Christian adoption of biblical versions made by Jews in the Roman province of Africa, where the vernacular was exclusively Latin. Only portions or quotations from it, however, have been preserved, and from these it can be assumed that the translation was made not from Hebrew but from Greek. For this reason, the Old Latin version is especially valuable because it reflects the state of the Septuagint before Origen's revision. By the 3rd century, several Latin versions circulated, and African and European recensions can be differentiated. Whether they all diverged from an original single translation or existed from the beginning independently cannot be determined. The textual confusion and the vulgar and colloquial nature of the Old Latin recension had become intolerable to the church authorities by the last decade of the 4th century, and *c.* 382 Pope Damasus decided to remedy the situation.

The Vulgate. The task of revision fell to Eusebius Hieronymus, generally known as St. Jerome (died 419/420), whose knowledge of Latin, Greek, and Hebrew made him the outstanding Christian biblical scholar of his time.

Jerome produced three revisions of the Psalms, all extant. The first was based on the Septuagint and is known as the Roman Psalter because it was incorporated into the liturgy at Rome. The second, produced in Palestine from the Hexaplaric Septuagint, tended to bring the Latin closer to the Hebrew. Its popularity in Gaul was such that it came to be known as the Gallican Psalter. This version was later adopted into the Vulgate. The third revision, actually a fresh translation, was made directly from the Hebrew, but it never enjoyed wide circulation. In the course of preparing the latter, Jerome realized the futility of revising the Old Latin solely on the basis of the Greek and apparently left that task unfinished. By the end of 405 he had executed his own Latin translation of the entire Old Testament based on the "Hebrew truth" (*Hebraica veritas*).

Acceptance of the Vulgate within the Western Church

Because of the canonical status of the Greek version within the church, Jerome's version was received at first with much suspicion, for it seemed to cast doubt on the authenticity of the Septuagint and exhibited divergences from the Old Latin that sounded discordant to those familiar with the traditional renderings. Augustine feared a consequent split between the Greek and Latin churches. The innate superiority of Jerome's version, however, assured its ultimate victory, and by the 8th century it had become the Latin Vulgate ("the common version") throughout the churches of Western Christendom, where it remained the chief Bible until the Reformation.

In the course of centuries of rival coexistence, the Old Latin and Jerome's Vulgate tended to react upon each other so that the Vulgate text became a composite. Other corruptions—noted in over 8,000 surviving manuscripts—crept in as a result of scribal transmission. Several medieval attempts were made to purify the Vulgate, but with little success. In 1546 the reforming Council of Trent accorded this version "authentic" status, and the need for a corrected text became immediate, especially because printing (introduced in the mid-15th century) could ensure, at last, a stabilized text. Because the Sixtine edition of Pope Sixtus V (1590) did not receive widespread support, Pope Clement VIII produced a fresh revision in 1592. This Clementine text remained the official edition of the Roman Church. Since 1907, the Benedictine Order, on the initiative of Pope Pius X, has been preparing a comprehensive edition. By 1969 only the Prophets still awaited publication to complete the Old Testament. A year later, a papal commission under Cardinal Augustinus Bea of Germany was charged with the task of preparing a new "revision of the Vulgate," taking the Benedictine edition as its working base.

Syriac versions. The Bible of the Syriac Churches is known as the Peshitta ("simple" translation). Though neither the reason for the title nor the origins of the versions are known, the earliest translations most likely served the needs of the Jewish communities in the region of Adiabene (in Mesopotamia), which are known to have existed as early as the 1st century CE. This probably explains the archaic stratum unquestionably present in the Pentateuch, Prophets, and Psalms of the Peshitta, as well as the undoubtedly Jewish influences generally, though Jewish-Christians also may have been involved in the rendering.

The Peshitta displays great variety in its style and in the translation techniques adopted. The Pentateuch is closest to the Masoretic text, but elsewhere there is much affinity with the Septuagint. This latter phenomenon might have resulted from later Christian revision.

Following the split in the Syriac Church in the 5th century into Nestorian (East Syrian) and Jacobite (West Syrian) traditions, the textual history of the Peshitta became bifurcated. Because the Nestorian Church was relatively isolated, its manuscripts are considered to be superior.

A revision of the Syriac translation was made in the early 6th century by Philoxenos, bishop of Mabbug, based on the Lucianic recension of the Septuagint. Another (the Syro-Hexaplaric version) was made by Bishop Paul of Tella in 617 from the Hexaplaric text of the Septuagint. A Palestinian Syriac version, extant in fragments, is known to go back to at least 700, and a fresh recension was made by Jacob of Edessa (died 708).

There are many manuscripts of the Peshitta, of which the oldest bears the date 442. Only four complete codices are extant from between the 5th and 12th centuries. No critical edition yet exists, but one is being prepared by the Peshitta Commission of the International Organization for the Study of the Old Testament.

Arabic versions. There is no reliable evidence of any pre-Islāmic Arabic translation. Only when large Jewish and Christian communities found themselves under Muslim rule after the Arab conquests of the 7th century did the need for an Arabic vernacular Scripture arise. The first and most important was that of Sa'adia ben Joseph (892–942), made directly from Hebrew and written in Hebrew script, which became the standard version for all Jews in Muslim countries. The version also exercised its influence upon Egyptian Christians and its rendering of the Pentateuch was adapted by Abū al-Ḥasan to the Samaritan Torah in the 11th–12th centuries. Another Samaritan Arabic version of the Pentateuch was made by Abū Sa'īd (Abū al-Barakāt) in the 13th century. Among other translations from the Hebrew, that of the 10th-century Karaite Yāphith ibn 'Alī is the most noteworthy.

Influence of Arabic versions

In 946 a Spanish Christian of Córdova, Isaac son of Velásquez, made a version of the Gospels from Latin. Manuscripts of 16th-century Arabic translations of both testaments exist in Leningrad, and both the Paris and London polyglots of the 17th century included Arabic versions. In general, the Arabic manuscripts reveal a bewildering variety of renderings dependent on Hebrew, Greek, Samaritan, Syriac, Coptic, and Latin translations. As such they have no value for critical studies. Several modern Arabic translations by both Protestants and Catholics were made in the 19th and 20th centuries.

Later and modern versions: English. Knowledge of the pre-Wycliffite English renditions stems from the many actual manuscripts that have survived and from secondary literature, such as booklists, wills, citations by later authors, and references in polemical works that have preserved the memory of many a translation effort.

Anglo-Saxon versions. For about seven centuries after the conversion of England to Christianity (beginning in the 3rd century), the common man had no direct access to the text of the Scriptures. Ignorant of Latin, his knowledge was derived principally from sermons and metrical prose paraphrases and summaries. The earliest poetic rendering of any part of the Bible is credited to Caedmon (flourished 658–680), but only the opening lines of his poem on the Creation in the Northumbrian dialect have been preserved.

An actual translation of the Psalter into Anglo-Saxon is ascribed to Aldhelm, bishop of Sherborne (died 709), but nothing has survived by which its true character, if it actually existed, might be determined. Linguistic considerations alone rule out the possibility that the prose translation of Psalms 1–50 extant in the Bibliothèque Nationale at Paris is a 7th-century production. In the next century, Bede (died 735) is said to have translated parts of the Gospels, and, though he knew Greek and possibly even some Hebrew, he does not appear to have applied himself to the Old Testament.

The outstanding name of the 9th century is that of King Alfred the Great. He appended to his laws a free translation of the Ten Commandments and an abridgment of the enactments of Exodus 21–23. These actually constitute the earliest surviving examples of a portion of the Old Testament in Anglo-Saxon prose.

An important step towards the emergence of a true English translation was the development of the interlinear gloss, a valuable pedagogic device for the introduction of youthful members of monastic schools to the study of the Bible. The Vespasian Psalter is the outstanding surviving example of the technique from the 9th century. In the next century the Lindisfarne Gospels, written in Latin *c.* 700, were glossed in Anglo-Saxon *c.* 950.

The last significant figure associated with the vernacular

Bible before the Norman Conquest was the so-called Aelfric the Grammarian (*c.* 955–1020). Though he claimed to have rendered several books into English, his work is more a paraphrase and abridgment than a continuous translation.

Anglo-Norman versions. The displacement of the English upper class, with the consequent decline of the Anglo-Saxon tradition attendant upon the Norman invasion, arrested for a while the movement toward the production of the English Bible. Within about 50 years (*c.* 1120) of the Conquest, Eadwine's *Psalterium triplex,* which contained the Latin version accompanied by Anglo-Norman and Anglo-Saxon renderings, appeared. The contemporary Oxford Psalter achieved such influence that it became the basis of all subsequent Anglo-Norman versions. By 1361 a prose translation of most of Scripture in this dialect had been executed.

The Wycliffite versions. By the middle of the 13th century the English component in the Anglo-Norman amalgam had begun to assert itself and the close of the century witnessed a Northumbrian version of the Psalter made directly from Latin, which, because it survived in several manuscripts, must have achieved relatively wide circulation. By the next century, English had gradually superseded French among the upper classes. When the first complete translation of the Bible into English emerged, it became the object of violent controversy because it was inspired by the heretical teachings of John Wycliffe. Intended for the common man, it became the instrument of opposition to ecclesiastical authority.

Effect of the Wycliffite Bible

The exact degree of Wycliffe's personal involvement in the Scriptures that came to bear his name is not clear. Because a note containing the words "Here ends the translation of Nicholas of Hereford" is found in a manuscript copy of the original (and incomplete) translation, it may be presumed that, though there must have been other assistants, Hereford can be credited with overall responsibility for most of the translation and that his summons before a synod in London and his subsequent departure for Rome in 1382 terminated his participation in the work. Who completed it is uncertain.

The Wycliffite translations encountered increasing ecclesiastical opposition. In 1408 a synod of clergy summoned to Oxford by Archbishop Arundel forbade the translation and use of Scripture in the vernacular. The proscription was rigorously enforced, but remained ineffectual. In the course of the next century the Wycliffite Bible, the only existing English version, achieved wide popularity as is evidenced by the nearly 200 manuscripts extant, most of them copied between 1420 and 1450.

The translation of William Tyndale. Because of the influence of printing and a demand for scriptures in the vernacular, William Tyndale began working on a New Testament translation directly from the Greek in 1523. The work could not be continued in England because of political and ecclesiastical pressures, and the printing of his translation began in Cologne (in Germany) in 1525. Again under pressure, this time from the city authorities, Tyndale had to flee to Worms, where two complete editions were published in 1525. Copies were smuggled into England where they were at once proscribed. Of 18,000 copies printed (1525–28), two complete volumes and a fragment are all that remain.

When the New Testament was finished Tyndale began work on the Old Testament. The Pentateuch was issued in Marburg in 1530, each of the five books being separately published and circulated. Tyndale's greatest achievement was the ability to strike a felicitous balance between the needs of scholarship, simplicity of expression, and literary gracefulness, all in a uniform dialect. The effect was the creation of an English style of Bible translation, tinged with Hebraisms, that was to serve as the model for all future English versions for nearly 400 years.

Influence of Tyndale's Bible

The translation of Miles Coverdale. A change in atmosphere in England found expression in a translation that, for all its great significance, turned out to be a retrograde step in the manner of its execution, although it proved to be a vindication of Tyndale's work. On October 4, 1535, the first complete English Bible, the work of Miles Coverdale, came off the press either in Zürich or in Cologne. The edition was soon exhausted. A second impression appeared in the same year and a third in 1536. A new edition, "overseen and corrected," was published in England by James Nycholson in Southwark in 1537. Another edition of the same year bore the announcement, "set forth with the king's most gracious license." In 1538 a revised edition of Coverdale's New Testament printed with the Latin Vulgate in parallel columns issued in England was so full of errors that Coverdale promptly arranged for a rival corrected version to appear in Paris.

The Thomas Matthew version. In the same year that Coverdale's authorized version appeared, another English Bible was issued under royal license and with the encouragement of ecclesiastical and political power. It appeared (Antwerp?) under the name of Thomas Matthew, but it is certainly the work of John Rogers, a close friend of Tyndale. Although the version claimed to be "truly and purely translated into English," it was in reality a combination of the labours of Tyndale and Coverdale. Rogers used the former's Pentateuch and 1535 revision of the New Testament and the latter's translation from Ezra to Malachi and his Apocrypha. Rogers' own contribution was primarily editorial.

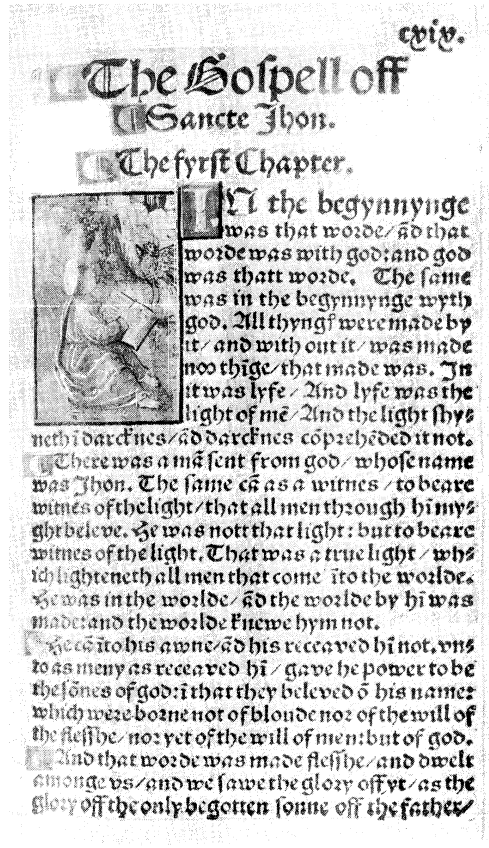
The Great Bible. In an injunction of 1538, Henry VIII commanded the clergy to install in a convenient place in every parish church, "one book of the whole Bible of the largest volume in English." The order seems to refer to an anticipated revision of the Matthew Bible. The first edition was printed in Paris and appeared in London in April 1539 in 2,500 copies. The huge page size earned it the sobriquet the Great Bible. It was received with immediate and wholehearted enthusiasm.

Henry VIII's support of the Great Bible

The first printing was exhausted within a short while, and it went through six subsequent editions between 1540 and 1541. "Editions" is preferred to "impressions" here since the six successive issues were not identical.

The Geneva Bible. The brief efflorescence of the Protestant movement during the short reign of Edward VI

By courtesy of the Baptist College, Bristol, England


cxiv.

The Gospell off Sancte Jhon.

The fyrſt Chapter.

IN the begynnynge was that worde/ & that worde was with god: and god was thatt worde. The ſame was in the begynnynge wyth god. All thyngſ were made by it/ and with out it/ was made noo thīge/ that made was. In it was lyfe/ And lyfe was the light of mē/ And the light ſhyneth i darcknes/ & darcknes cōprehēded it not.

There was a mā ſent from god/ whoſe name was Jhon. The ſame cā as a witnes/ to beare witnes of the light/ that all men thrōugh hī myght believe. He was nott that light: but to beare witnes of the light. That was a true light/ whych lightneth all men that come into the worlde. He was in the worlde/ & the worlde by hī was made: and the worlde knewe hym not.

He cā to his awne/ & his receaved hī not. Vnto as many as receaved hī/ gave he power to be the ſōnes of god: i that they believed ō his name: which were borne not of blonde nor of the will of the fleſſhe/ nor yet of the will of man: but of god.

And that worde was made fleſſhe/ and dwelt amonge vs/ and we ſawe the glory off yt/ as the glory off the only begotten ſonne off the father/

The opening page of chapter 1 of the Gospel According to John from Tyndale's Bible, 1525–26. In the library of the Baptist College, Bristol, England.

(1547–53) saw the reissue of the Scriptures, but no fresh attempts at revision. The repressive rule of Edward's successor, Mary, a Roman Catholic, put an end to the printing of Bibles in England for several years. Their public reading was proscribed and their presence in the churches discontinued.

The persecutions of Protestants caused the focus of English biblical scholarship to be shifted abroad where it flourished in greater freedom. A colony of Protestant exiles, led by Coverdale and John Knox (the Scottish Reformer), and under the influence of John Calvin, published the New Testament in 1557.

The editors of the Geneva Bible (or "Breeches Bible," so-named because of its rendering of the first garments made for Adam and Eve in chapter three, verse seven of Genesis)—published in 1560—may almost certainly be identified as William Whittingham, the brother-in-law of Calvin's wife, and his assistants Anthony Gilby and Thomas Sampson. The Geneva Bible was not printed in England until 1576, but it was allowed to be imported without hindrance. The accession of Elizabeth in 1558 put an end to the persecutions and the Great Bible was soon reinstated in the churches. The Geneva Bible, however, gained instantaneous and lasting popularity over against its rival, the Great Bible. Its technical innovations contributed not a little to its becoming for a long time the family Bible of England, which, next to Tyndale, exercised the greatest influence upon the King James Version.

The Bishops' Bible. The failure of the Great Bible to win popular acceptance against the obvious superiority of its Geneva rival and the objectionable partisan flavour of the latter's marginal annotations made a new revision a necessity. By about 1563–64 Archbishop Matthew Parker of Canterbury had determined upon its execution and the work was apportioned among many scholars, most of them bishops, from which the popular name was derived.

The Bishops' Bible came off the press in 1568 as a handsome folio volume, the most impressive of all 16th-century English Bibles in respect of the quality of paper, typography, and illustrations. A portrait of the Queen adorned the engraved title page, but it contained no dedication. For some reason Queen Elizabeth never officially authorized the work, but sanction for its public use came from the Convocation (church synod or assembly) of 1571 and it thereby became in effect the second authorized version.

Need for a Roman Catholic Version

The Douai–Reims Bible. The Roman Catholics addressed themselves affirmatively to the same problem faced by the Anglican Church: a Bible in the vernacular. The initiator of the first such attempt was Cardinal Allen of Reims (in France), although the burden of the work fell to Gregory Martin, professor of Hebrew at Douai. The New Testament appeared in 1582, but the Old Testament, delayed by lack of funds, did not appear until 1609, when it was finally published at Douai under the editorship of Thomas Worthington. In the intervening period it had been brought into line with the new text of the Vulgate authorized by Clement VIII in 1592.

The King James (Authorized) Version. Because of changing conditions, another official revision of the Protestant Bible in English was needed. The reign of Queen Elizabeth had succeeded in imposing a high degree of uniformity upon the church. The failure of the Bishops' Bible to supplant its Geneva rival made for a discordant note in the quest for unity.

A conference of churchmen in 1604 became noteworthy for its request that the English Bible be revised because existing translations "were corrupt and not answerable to the truth of the original." King James I was quick to appreciate the broader value of the proposal and at once made the project his own.

By June 30, 1604, King James had approved a list of 54 revisers, although extant records show that 47 scholars actually participated. They were organized into six companies, two each working separately at Westminster, Oxford, and Cambridge on sections of the Bible assigned to them. It was finally published in 1611.

Significance of the King James Version

Not since the Septuagint had a translation of the Bible been undertaken under royal sponsorship as a cooperative venture on so grandiose a scale. An elaborate set of rules was contrived to curb individual proclivities and to ensure its scholarly and nonpartisan character. In contrast to earlier practice, the new version was to preserve vulgarly used forms of proper names in keeping with its aim to make the Scriptures popular and familiar.

The impact of Jewish sources upon the King James Version is one of its noteworthy features. The wealth of scholarly tools available to the translators made their final choice of rendering an exercise in originality and independent judgment. For this reason, the new version was more faithful to the original languages of the Bible and more scholarly than any of its predecessors. The impact of the Hebrew upon the revisers was so pronounced that they seem to have made a conscious effort to imitate its rhythm and style in the Old Testament. The English of the New Testament actually turned out to be superior to its Greek original.

Two editions were actually printed in 1611, later distinguished as the "He" and "She" Bibles because of the variant reading "he" and "she" in the final clause of chapter 3, verse 15 of Ruth: "and he went into the city." Both printings contained errors. Some errors in subsequent editions have become famous: The so-called Wicked Bible (1631) derives from the omission of "not" in chapter 20 verse 14 of Exodus, "Thou shalt commit adultery," for which the printers were fined £300; the "Vinegar Bible" (1717) stems from a misprinting of "vineyard" in the heading of Luke, chapter 20.

The English Revised Version. The remarkable and total victory of the King James Version could not entirely obscure those inherent weaknesses that were independent of its typographical errors. The manner of its execution had resulted in a certain unequalness and lack of consistency. The translators' understanding of the Hebrew tense system was often limited so that their version contains inaccurate and infelicitous renderings. In particular, the Greek text of the New Testament, which they used as their base, was a poor one. The great early Greek codices were not then known or available, and Hellenistic papyri, which were to shed light on the common Greek dialect, had not yet been discovered.

A committee established by the Convocation of Canterbury in February 1870 reported favourably three months later on the idea of revising the King James Version: two companies were formed, one each for the Old and New Testaments. A novel development was the inclusion of scholars representative of the major Christian denominations, except the Roman Catholics (who declined the invitation to participate). Another innovation was the formation of parallel companies in the United States to whom the work of the English scholars was submitted and who, in turn, sent back their reactions. The instructions to the committees made clear that only a revision and not a new translation was contemplated.

Cooperation in the publication of the Revised Version

The New Testament was published in England on May 17, 1881, and three days later in the United States, after 11 years of labour. Over 30,000 changes were made, of which more than 5,000 represent differences in the Greek text from that used as the basis of the King James Version. Most of the others were made in the interests of consistency or modernization.

The publication of the Old Testament in 1885 stirred far less excitement, partly because it was less well known than the New Testament and partly because fewer changes were involved. The poetical and prophetical books, especially Job, Ecclesiastes, and Isaiah, benefitted greatly.

The revision of the Apocrypha, not originally contemplated, came to be included only because of copyright arrangements made with the university presses of Oxford and Cambridge and was first published in 1895.

The American Standard Version. According to the original agreement, the preferred readings and renderings of the American revisers, which their British counterparts had declined to accept, were published in an appendix to the Revised Version. In 1900 the American edition of the New Testament, which incorporated the American scholars' preferences into the body of the text, was produced. A year later the Old Testament was added, but not the Apocrypha. The alterations covered a large number of

obsolete words and expressions and replaced Anglicisms by the diction then in vogue in the United States.

The Revised Standard Version. The American Standard Version had been an expression of sensitivity to the needs of the American public. At the same time, several individual and unofficial translations into modern speech made from 1885 on had gained popularity, their appeal reinforced by the discovery that the Greek of the New Testament used the common nonliterary variety of the language spoken throughout the Roman Empire when Christianity was in its formative stage. The notion that a nonliterary modern rendering of the New Testament best expressed the form and spirit of the original was hard to refute. This, plus a new maturity of classical, Hebraic, and theological scholarship in the United States, led to a desire to produce a native American version of the English Bible.

In 1928 the copyright of the American Standard Version was acquired by the International Council of Religious Education and thereby passed into the ownership of churches representing 40 major denominations in the United States and Canada. A two-year study by a special committee recommended a thorough revision, and in 1937 the council gave its authorization to the proposal. Not until 1946, however, did the revision of the New Testament appear in print, and another six years elapsed before the complete Revised Standard Version (RSV) was published, the work of 32 scholars, one of them Jewish, drawn from the faculties of 20 universities and theological seminaries. A decision to translate the Apocrypha was not made until 1952 and the revision appeared in 1957. Insofar as the RSV was the first to make use of the Dead Sea Scroll of Isaiah, it was revolutionary.

Modernization in the Revised Standard Version

The Revised Standard Version was essentially not a new translation into modern speech, but a revision. It did engage in a good deal of modernization, however. It dispensed with archaic pronouns, retaining "thou" only for the Deity. But its basic conservatism was displayed in the retention of forms or expressions in passages that have special devotional or literary associations even where this practice makes for inconsistency. The primary aim was to produce a version for use in private and public worship.

Jewish versions. Though Jews in English-speaking lands generally utilized the King James Version and the Revised Version, the English versions have presented great difficulties. They contain departures from the traditional Hebrew text; they sometimes embody Christological interpretations; the headings were often doctrinally objectionable and the renderings in the legal portions of the Pentateuch frequently diverged from traditional Jewish exegesis. In addition, where the meaning of the original was obscure, Jewish readers preferred to use the well-known medieval Jewish commentators. Finally, the order of the Jewish canon differs from Christian practice and the liturgical needs of Jews make a version that does not mark the scriptural readings for Sabbaths and festivals inconvenient.

Until 1917 all Jewish translations were the efforts of individuals. Planned in 1892, the project of the Jewish Publication Society of America was the first translation for which a group representing Jewish learning among English-speaking Jews assumed joint responsibility.

This version essentially retained the Elizabethan diction. It stuck unswervingly to the received Hebrew text that it interpreted in accordance with Jewish tradition and the best scholarship of the day. For over half a century it remained authoritative, even though it laid no claim to any official ecclesiastical sanction.

With an increasingly felt need for modernization, a committee of translators was established composed of three professional biblical and Semitic scholars and three rabbis. It began its work in 1955 and the Pentateuch was issued in 1962. The Song of Songs, Ruth, Lamentations, Ecclesiastes, Esther, and Jonah, all in a single volume for the convenience of synagogue use, followed in 1969; and Isaiah and Psalms appeared in 1973. A second committee had been set up in 1955 to work separately on the rest of the Hagiographa (Ketuvim).

The New English Bible. The idea of a completely new translation into British English was first broached in 1946. Under a joint committee, representative of the major Protestant churches of the British Isles, with Roman Catholics appointed as observers, the New Testament was published in 1961 and a second edition appeared in 1970. The Old Testament and Apocrypha were also published in 1970.

Success of the New English Bible

The New English Bible proved to be an instant commercial success, selling at a rate of 33,000 copies a week in 1970. The translation differed from the English mainstream Bible in that it was not a revision but a completely fresh version from the original tongues. It abandoned the tradition of "biblical English" and, except for the retention of "thou" and "thy" in addressing God, freed itself of all archaisms. It endeavoured to render the original into the idiom of contemporary English and to avoid ephemeral modernisms.

Catholic versions. With the exception of a version by Irish-American archbishop Francis Patrick Kenrick (1849–60), all translations up to the 20th century were merely versions of the Douai–Reims Bible. A celebrated translation was that of Ronald Knox (New Testament, 1945; Old Testament, 1949; complete edition with Old Testament revised, 1955).

The most significant development in modern Catholic translations was initiated by the Confraternity of Christian Doctrine in 1936. A New Testament version of the Latin Clementine Vulgate (1941), intended as a revision, in effect was a new translation into clear and simple English. The Old Testament revision remained unfinished, the work having been interrupted by a decision inspired by the Pontifical Biblical Commission in 1943 to encourage modern vernacular translations from the original languages instead of from the Latin Vulgate. Accordingly, both the Old and New Testaments were respectively retranslated into modern English from the Hebrew and Greek originals. The resultant Confraternity Version (1952–61) was later issued as the New American Bible (1970). Another modern version, more colloquial, is the Jerusalem Bible (1966), translated from the French Catholic Bible de Jérusalem (one-volume edition, 1961).

Later and modern versions: continental. *Dutch versions.* Until the Reformation, Dutch Bible translations were largely free adaptations, paraphrases, or rhymed verse renderings of single books or parts thereof. A popular religious revival at the end of the 12th century accelerated the demand for the vernacular Scriptures, and one of the earliest extant examples is the Liège manuscript (*c.* 1270) translation of the *Diatessaron* (a composite rendering of the four Gospels) by Tatian, a 2nd century Syrian Christian heretical scholar; it is believed to derive from a lost Old Latin original. Best known of all the rhymed versions is the *Rijmbijbel* of Jacob van Maerlant (1271) based on Peter Comestar's *Historia scholastica.* Despite the poor quality of Johan Schutken's translation of the New Testament and Psalms (1384), it became the most widely used of medieval Dutch versions.

Early vernacular versions

With the Reformation came a renewed interest in the study of the Scriptures. Luther's Bible (see *German versions,* below) was repeatedly rendered into Dutch, the most important version being that of Jacob van Liesveldt (1526). It was mainly to counter the popularity of this edition that Roman Catholics produced their own Dutch Bible, executed by Nicolaas van Winghe (Louvain, 1548). A revision printed by Jan Moerentorf (Moretus, 1599) became the standard version until it was superseded by that of the Peter Canisius Association (1929–39), now in general use. A fresh translation of the New Testament in modern Dutch appeared in 1961.

French versions. The deep conflicts that characterized the history of Christianity in France made it difficult for one authoritative version to emerge.

The first complete Bible was produced in the 13th century at the University of Paris and toward the end of that century Guyart des Moulins executed his *Bible Historiale.* Both works served as the basis of future redactions of which the Bible printed in Paris (date given variously as 1487, 1496, 1498) by order of King Charles VIII, is a good example.

The real history of the French Bible began in Paris, in 1523, with the publication of the New Testament, almost certainly the work of the Reformer Jacques Lefèvre

d'Étaples (Faber Stapulensis). The Old Testament appeared in Antwerp in 1528 and the two together in 1530 as the Antwerp Bible. The first true Protestant version came out in Serrières, near Neuchâtel, five years later, the work of Pierre Robert, called Olivétan. This version was frequently revised throughout the 16th century, the most celebrated editions being Calvin's of 1546 and that of Robert Estienne (Stephanus) of 1553. The Roman Catholics produced a new version, the Louvain Bible of 1550, based on both Lefèvre and Olivétan. Modernizations of Olivétan appeared in succeeding centuries. The most important French version of the 20th century is the Jerusalem Bible prepared by professors at the Dominican École Biblique de Jérusalem (Paris, 1949–54, complete, 1956).

Gothic and pre-Reformation Bibles

German versions. The early Old Testament in Gothic has already been described. The New Testament remains are far more extensive and are preserved mainly in the Codex Argenteus (*c.* 525) and Codex Gissensis. The translation, essentially based on a Byzantine text, is exceedingly literal and not homogeneous. It is difficult to determine the degree of contamination that the original Gospels translation of Ulfilas had undergone by the time it appeared in these codices.

Nothing is known of the vernacular Scriptures in Germany prior to the 8th century when an idiomatic translation of Matthew from Latin into the Bavarian dialect was made. From Fulda (in Germany) *c.* 830 came a more literal East Franconian German translation of the Gospel story. In the same period was produced the *Heliand* ("Saviour"), a versified version of the Gospels. Such poetic renderings cannot, strictly speaking, be regarded as translations. There is evidence, however, for the existence of German Psalters from the 9th century on. By the 13th century, the different sects and movements that characterized the religious situation in Germany had stimulated a demand for popular Bible reading. Since all the early printed Bibles derived from a single family of late 14th-century manuscripts, German translations must have gained wide popularity. Another impetus towards the use of the German Scriptures in this period can be traced to mystics of the Upper Rhine. A complete New Testament, the Augsburg Bible, can be dated to 1350, and another from Bohemia, Codex Teplensis (*c.* 1400), has also survived.

The Wenzel Bible, an Old Testament made between 1389 and 1400, is said to have been ordered by King Wenceslas, and large numbers of 15th-century manuscripts have been preserved.

The first printed Bible (the Mentel Bible) appeared at Strassburg no later than 1466 and ran through 18 editions before 1522. Despite some evidence that ecclesiastical authority did not entirely look with favour upon this vernacular development, the printed Bible appeared in Germany earlier, and in more editions and in greater quantity than anywhere else.

A new era opened up with the work of Martin Luther, to whom a translation from the original languages was a necessary and logical conclusion of his doctrine of justification by faith—to which the Scriptures provided the only true key. His New Testament (Wittenberg, 1522) was made from the second edition of Erasmus' Greek Testament. The Old Testament followed in successive parts, based on the Brescia Hebrew Bible (1494). Luther's knowledge of Hebrew and Aramaic was limited, but his rendering shows much influence of Rashi, the great 11th–12th-century French rabbinical scholar and commentator, through the use of the notes of Nicholas of Lyra. The complete Lutheran Bible emerged from the press in 1534. Luther was constantly revising his work with the assistance of other scholars, and between 1534 and his death in 1546, 11 editions were printed, the last posthumously. His Bible truly fulfilled Luther's objective of serving the needs of the common man, and it, in turn, formed the basis of the first translations in those lands to which Lutheranism spread. It proved to be a landmark in German prose literature and contributed greatly to the development of the modern language.

Influence of Luther's Bible

The phenomenal success of Luther's Bible and the failure of attempts to repress it led to the creation of German Catholic versions, largely adaptations of Luther. Hieronymus Emser's edition simply brought the latter into line with the Vulgate. Johann Dietenberger issued a revision of Emser (Mainz, 1534) and used Luther's Old Testament in conjunction with an Anabaptist (radical Protestant group) version and the Zürich (Switzerland) version of 1529. It became the standard Catholic version. Of the 20th-century translations, the Grünewald Bible, which reached a seventh edition in 1956, is one of the most noteworthy.

Yiddish versions

German glosses in Hebrew script attached to Hebrew Bibles in the 12th and 13th centuries constitute the earliest Jewish attempts to render the Scriptures into that German dialect current among the Jews of middle Europe, the dialect that developed into Judeo-German, or Yiddish. The first translation proper has been partially preserved in a manuscript from Mantua dated 1421. The earliest printed translation is that of the Scriptural dictionaries prepared by a baptized Jew, Michael Adam (Constance, 1543–44; Basel, 1583, 1607). The version of Jacob ben Isaac Ashkenazi of Janów, known as the *Tz'enah u-Re'na* (Lublin, 1616), became one of the most popular and widely diffused works of its kind.

The first Jewish translation into pure High German, though in Hebrew characters (1780–83), made by Moses Mendelssohn, opened a new epoch in German-Jewish life. The first Jewish rendering of the entire Hebrew Bible in German characters was made by Gotthold Salomon (Altona, 1837). An attempt to preserve the quality of the Hebrew style in German garb was the joint translation of two Jewish religious philosophers, Martin Buber and Franz Rosenzweig (15 vol., Berlin, 1925–37; revised ed. Cologne, 4 vol., 1954–62).

Greek versions. A 13th-century manuscript of Jonah by a Jew is the earliest known post-Hellenistic Greek biblical work. A rendering of Psalms was published by a Cretan monk Agapiou in 1563. A version in Hebrew characters (a large part of the Old Testament) appeared in the Constantinople Polyglot Pentateuch in 1547.

The first New Testament was done by Maximus of Gallipoli in 1638 (at Geneva?). The British and Foreign Bible Society published the Old Testament in 1840 (London) and the New Testament in 1848 (Athens). Between 1900 and 1924, however, the use of a modern Greek version was prohibited. The theological faculty of the University of Athens is now preparing a fresh translation.

Hungarian versions. The spread of Lutheranism in the Reformation period gave rise to several vernacular versions. János Sylvester (Erdősi) produced the first New Testament made from the Greek (Sárvár, 1541). The Turkish occupation of much of Hungary and the measures of the Counter-Reformation arrested further printing of the vernacular Bible, except in the semi-independent principality of Transylvania. The first complete Hungarian Bible, issued at Vizsoly in 1590, became the Protestant Church Bible.

In the 20th century, a new standard edition for Protestants was published, the New Testament appearing in 1956 and the Old Testament (Genesis to Job) in 1951 and following. A new modernized Catholic edition of the New Testament from the Greek appeared in Rome in 1957.

Italian versions. The vernacular Scriptures made a relatively late appearance in Italy. Existing manuscripts of individual books derive from the 13th century and mainly consist of the Gospels and the Psalms.

These medieval versions were never made from the original languages. They were influenced by French and Provençal renderings as well as by the form of the Latin Vulgate current in the 12th and 13th centuries in southern France. There is evidence for a Jewish translation made directly from the Hebrew as early as the 13th century.

The first printed Italian Bible appeared in Venice in 1471, translated from the Latin Vulgate by Niccolò Malermi. In 1559 Paul IV proscribed all printing and reading of the vernacular Scriptures except by permission of the church. This move, reaffirmed by Pius IV in 1564, effectively stopped further Catholic translation work for the next 200 years.

The first Protestant Bible (Geneva, 1607, revised 1641) was the work of Giovanni Diodati, a Hebrew and Greek scholar. Frequently reprinted, it became the standard Prot-

estant version until the 20th century. Catholic activity was renewed after a modification of the ban by Pope Benedict XIV in 1757. A complete Bible in translation made directly from the Hebrew and Greek has been in progress under the sponsorship of the Pontifical Biblical Institute since the 1920s.

Portuguese versions. The first Portuguese New Testament (Amsterdam), the work of João Ferreira d'Almeida, did not appear until 1681. The first complete Bible (2 vol., 1748–53) was printed in Batavia (in Holland). Not until late in the 18th century did the first locally published vernacular Scriptures appear in Portugal. A revision of d'Almeida was issued in Rio de Janeiro (in Brazil), the New Testament in 1910 and the complete Bible in 1914 and 1926; an authorized edition in modernized orthography was published by the Bible Society of Brazil (New Testament, 1951; Old Testament, 1958). A new translation of the New Testament from Greek by José Falcão came out in Lisbon (1956–65).

Scandinavian versions. In pre-Reformation times, only partial translations were made, all on the basis of the Latin Vulgate and all somewhat free. The earliest and most celebrated is that of Genesis–Kings in the so-called *Stjórn* ("Guidance"; *i.e.,* of God) manuscript in the Old Norwegian language, probably to be dated about 1300. Swedish versions of the Pentateuch and of Acts have survived from the 14th century and a manuscript of Joshua–Judges by Nicholaus Ragnvaldi of Vadstena from *c.* 1500. The oldest Danish version covering Genesis–Kings derives from 1470.

Danish, Norwegian, Icelandic, and Swedish versions

Within two years of publication, Luther's New Testament had already influenced a Danish translation made at the request of the exiled king Christian II by Christiern Vinter and Hans Mikkelsen (Wittenberg, 1524). In 1550 Denmark received a complete Bible commissioned by royal command (the Christian III Bible, Copenhagen). A revision appeared in 1589 (the Frederick II Bible) and another in 1633 (the Christian IV Bible).

A rendering by Hans Paulsen Resen (1605–07) was distinguished by its accuracy and learning and was the first made directly from Hebrew and Greek, but its style was not felicitous and a revision was undertaken by Hans Svane (1647). Nearly 200 years later (1819), a combination of the Svaning Old Testament and the Resen–Svane New Testament was published. In 1931 a royal commission produced a new translation of the Old Testament with the New Testament following in 1948 and the Apocrypha in 1957.

The separation of Norway from Denmark in 1814 stimulated the revival of literature in the native language. The Old Testament of 1842–87 (revised, 1891) and New Testament of 1870–1904 were still intelligible to Danish readers, but the version of E. Blix (New Testament, 1889; complete Bible, 1921) is in New Norwegian. A revised Bible in this standardized form of the language, executed by R. Indrebö, was published by the Norwegian Bible Society in 1938.

The first Icelandic New Testament was the work of Oddur Gottskálksson (Roskilde, Denmark, 1540), based on the Latin Vulgate and Luther. It was not until 1584 that the complete Icelandic Scriptures were printed (at Hólar), mainly executed by Gudbrandur Thorláksson. It was very successful and became the Church Bible until displaced by the revision of Thorlákur Skúlason (1627–55), based apparently on Resen's Danish translation. In 1827 the Icelandic Bible Society published a new New Testament and a complete Bible in 1841 (Videyjar; 1859, Reykjavík), revised and reprinted at Oxford in 1866. A completely new edition (Reykjavík, 1912) became the official Church Bible.

Soon after Sweden achieved independence from Denmark in the early 16th century, it acquired its own version of the New Testament published by the royal press (Stockholm, 1526). Luther's New Testament of 1522 served as its foundation, but the Latin Vulgate and Erasmus' Greek were also consulted. The first official complete Bible and the first such in any Scandinavian country was the Gustav Vasa Bible (Uppsala; 1541), named for the Swedish king under whose reign it was printed. It utilized earlier Swedish translations as well as Luther's. A corrected version (the Gustavus Adolphus Bible, named for the reigning Swedish king) was issued in 1618, and another with minor alterations by Eric Benzelius in 1703. The altered Bible was called the Charles XII Bible, because it was printed during the reign of Charles XII. In 1917 the church diet of the Lutheran Church published a completely fresh translation directly from modern critical editions of the Hebrew and Greek originals and it received the authorization of Gustaf V to become the Swedish Church Bible.

South, East, and West Slavic versions

Slavic versions. The earliest Old Church Slavonic translations are connected with the arrival of the brothers Cyril and Methodius in Moravia in 863, and resulted from the desire to provide vernacular renderings of those parts of the Bible used liturgically. The oldest manuscripts derive from the 11th and 12th centuries. The earliest complete Bible manuscript, dated 1499, was used for the first printed edition (Ostrog, 1581). This was revised in Moscow in 1633 and again in 1712. The standard Slavonic edition is the St. Petersburg revision of 1751, known as the Bible of Elizabeth.

The printing of parts of the Bulgarian Bible did not begin until the mid-19th century. A fresh vernacular version of the whole Bible was published at Sofia in 1925, having been commissioned by the Synod of the Bulgarian Orthodox Church.

The Serbian and Croatian literary languages are identical; they differ only in the alphabet they use. To further the dissemination of Protestantism among the southern Slavs, Count Jan Ungnad set up a press in 1560 at Urach that issued a translation of the New Testament, in both Glagolitic (1562–63) and Cyrillic (1563) characters. The efforts of the Serbian leader Vuk Karadžić to establish the Serbo-Croatian vernacular on a literary basis resulted in a new translation of the New Testament (Vienna, 1847) that went through many revisions.

The spread of the Lutheran Reformation to the Slovene-speaking provinces of Austria stimulated the need for vernacular translations. The first complete Slovene Bible, translated from the original languages but with close reference to Luther's German, was made by Jurij Dalmatin (Wittenberg, 1584). Not until two centuries later did a Slovene Roman Catholic version, rendered from the Latin Vulgate, appear (Laibach, 1784–1802).

Between the 9th and 17th centuries the literary and ecclesiastical language of Russia was Old Slavonic. A vernacular Scriptures was thus late in developing. An incomplete translation into the Belorussian dialect was prepared by Franciscus Skorina (Prague, 1517–19) from the Latin Vulgate and Slavonic and Bohemian versions, but not until 1821 did the first New Testament appear in Russian, an official version printed together with the Slavonic. With the more liberal rule of Alexander II, the Holy Synod sponsored a fresh version of the Gospels in 1860. The Old Testament was issued at St. Petersburg in 1875. A Jewish rendering was undertaken by Leon Mandelstamm, who published the Pentateuch in 1862 (2nd ed., 1871) and the Psalter in 1864. Prohibited in Russia, it was first printed in Berlin. A complete Bible was published in Washington in 1952.

No manuscript in the Czech vernacular translation is known to predate the 14th century, but at least 50 complete or fragmentary Bibles have survived from the 15th. The first complete Bible was published in Prague in 1488 in a text based on earlier, unknown translations connected with the heretical Hussite movement. The most important production of the century, however, was that associated principally with Jan Blahoslav. Based on the original languages, it appeared at Kralice in six volumes (1579–93). The Kralice Bible is regarded as the finest extant specimen of classical Czech and became the standard Protestant version.

Closely allied to the Czech language, but not identical with it, Slovakian became a literary language only in the 18th century. A Roman Catholic Bible made from the Latin Vulgate by Jiří Palkovič was printed in the Gothic script (2 vol. Gran, 1829, 1832) and another, associated with Richard Osvald, appeared at Trnava in 1928. A Protestant New Testament version of Josef Rohaček was

published at Budapest in 1913 and his completed Bible at Prague in 1936. A new Slovakian version by Stefan Zlatoš and Anton Jan Surjanský was issued at Trnava in 1946.

A manuscript of a late 14th-century Psalter is the earliest extant example of the Polish vernacular Scriptures, and several books of the Old Testament have survived from the translation made from the Czech version for Queen Sofia (Sârospatak Bible, 1455). Otherwise, post-Reformation Poland supplied the stimulus for biblical scholarship. The New Testament first appeared in a two-volume rendering from the Greek by the Lutheran Jan Seklucjan (Königsberg, 1553). The "Brest Bible" of 1563, sponsored by Prince Radziwiłł, was a Protestant production made from the original languages. A version of this edition for the use of Socinians (Unitarians) was prepared by the Hebraist Szymon Budny (Nieswicz, 1570–82), and another revision, primarily executed by Daniel Mikołajewski and Jan Turnowski (the "Danzig Bible") in 1632, became the official version of all Evangelical churches in Poland. This edition was burnt by the Catholics and had to be subsequently printed in Germany. The standard Roman Catholic version (1593, 1599) was prepared by Jakób Wujek whose work, revised by the Jesuits, received the approval of the Synod of Piotrkow in 1607. A revised edition was put out in 1935.

Spanish versions. The history of the Spanish Scriptures is unusual in that many of the translations were based, not on the Latin Vulgate, but on the Hebrew, a phenomenon that is to be attributed to the unusual role played by Jews in the vernacular movement.

Nothing is known from earlier than the 13th century when James I of Aragon in 1233 proscribed the possession of the Bible in "romance" (the Spanish vernacular) and ordered such to be burnt. Several partial Old Testament translations by Jews as well as a New Testament from a Visigoth Latin text are known from this century. In 1417 the whole Bible was translated into Valencian Catalan, but the entire edition was destroyed by the Inquisition.

Between 1479 and 1504, royal enactments outlawed the vernacular Bible in Castile, Leon, and Aragon, and the expulsion of the Jews from Spain in 1492 transferred the centre of Spanish translation activity to other lands. In 1557, the first printed *Index of Forbidden Books* of the Spanish Inquisition prohibited the "Bible in Castilian romance or any other vulgar tongue," a ban that was repeated in 1559 and remained in force until the 18th century. In 1916 the Hispano-Americana New Testament appeared in Madrid as an attempt to achieve a common translation for the entire Spanish-speaking world. The first Roman Catholic vernacular Bible from the original languages was made under the direction of the Pontifical University of Salamanca (Madrid, 1944, 9th ed. 1959).

Swiss versions. Four parts of Luther's version were reprinted in the Swyzerdeutsch dialect in Zürich in 1524–25. The Prophets and Apocrypha appeared in 1529. A year later, the first Swiss Bible was issued with the Prophets and Apocrypha independently translated. The Swiss Bible underwent frequent revision between 1660 and 1882. A fresh translation from the original languages was made between 1907 and 1931.

Non-European versions. Translations of parts of the Bible are known to have existed in only seven Asian and four African languages before the 15th century. In the 17th century Dutch merchants began to interest themselves in the missionary enterprise among non-Europeans. A pioneer was Albert Cornelius Ruyl, who is credited with having translated Matthew into High Malay in 1629, with Mark following later. Jan van Hasel translated the two other Gospels in 1646 and added Psalms and Acts in 1652. Other traders began translations into Formosan Chinese (1661) and Sinhalese (1739).

A complete printed Japanese New Testament reputedly existed in Miyako in 1613, the work of Jesuits. The first known printed New Testament in Asia appeared in 1715 in the Tamil language done by Bartholomäus Ziegenbalg, a Lutheran missionary. A complete Bible followed in 1727. Six years later the first Bible in High Malay came out.

The distinction of having produced the first New Testament in any language of the Americas belongs to John Eliot, a Puritan missionary, who made it accessible to the Massachusetts Indians in 1661. Two years later he brought out the Massachusetts Indian Bible, the first Bible to be printed on the American continent.

By 1800 the number of non-European versions did not exceed 13 Asian, four African, three American, and one Oceanian. With the founding of missionary societies after 1800, however, new translations were viewed as essential to the evangelical effort. First came renderings in those languages that already possessed a written literature. A group at Serāmpore (in India) headed by William Carey, a Baptist missionary, produced 28 versions in Indian languages. Robert Morrison, the first Protestant missionary to China, translated the New Testament into Chinese in 1814 and completed the Bible by 1823. Adoniram Judson, an American missionary, rendered the Bible into Burmese in 1834.

With European exploration of the African continent often came the need to invent an alphabet, and in many instances the translated Scriptures constituted the first piece of written literature. In the 19th century the Bible was translated into Amharic, Malagasy, Tswana, Xosa, and Ga.

American Indian versions

In the Americas, James Evans invented a syllabary for the use of Cree Indians, in whose language the Bible was available in 1862, the work of W. Mason, also a Wesleyan missionary. The New Testament appeared in Ojibwa in 1833, and the whole Bible was translated for the Dakota Indians in 1879. The Labrador Eskimos had a New Testament in 1826 and a complete Bible in 1871.

In Oceania, the New Testament was rendered into Tahitian and Javanese in 1829 and into Hawaiian and Low Malay in 1835. By 1854 the whole Bible had appeared in all but the last of these languages as well as in Rarotonga (1851).

In the 20th century the trend toward the development of non-European Bible translations was characterized by an attempt to produce "union" or "standard" versions in the common language underlying different dialects. One such is the Swahili translation (1950) that makes the Scriptures accessible to most of East Africa. Within the realm of non-European translation there has also been a movement for the updating of versions to bring them in line with the spoken language, especially through the use of native Christian scholars. The first example of this was the colloquial Japanese version of 1955.

By 1970 some part, if not the entire Bible, had been translated into more than 100 languages or dialects spoken in India and over 300 in Africa. (N.M.Sa.)

Old Testament history

History is a central element of the Old Testament. It is the subject of narration in the specifically historical books and of celebration, commemoration, and remonstration in all of the books. History in the Old Testament is not history in the modern sense; it is the story of events seen as revealing the divine presence and power. Nevertheless, it is the account of an actual people in an actual geographical area at certain specified historical times and in contact with other particular peoples and empires known from other sources. Hence, far more than with other great religious scriptures, a knowledge of the historical background is conducive, if not essential, to an adequate understanding of a major portion of the Old Testament. Recent archaeological discoveries as well as comparative historical research and philological studies, collated with an analysis and interpretation of the Old Testament text (still the major source of information), have made possible a fuller and more reliable picture of biblical history than in previous eras. For another presentation of Old Testament history, see JUDAISM.

EARLY DEVELOPMENTS

Background and beginnings. The geographical theatre of the Old Testament is the ancient Near East, particularly the Fertile Crescent region, running from the Tigris and Euphrates rivers up to Syria and down through Palestine to the Nile Delta. In this area great civilizations and em-

pires developed and seminomadic ethnic groups, such as the Hebrews, were involved in the mixture of peoples and cultures. The exact origin of the Hebrews is not known with certainty, but the biblical tradition of their origin in a clan that migrated from Mesopotamia to Canaan (Palestine) early in the 2nd millennium BCE has analogues in what is known of the movements of other groups in that area and period. There are, moreover, obvious Mesopotamian motifs in biblical cosmogony and primeval history in the early part of the Bible, and Mesopotamian placenames are the obvious bases of some of the personal names of the clan's forebears. Canaanite influences are evident in the Hebrew alphabet, poetry, and certain mythological themes. Linguistic and other similarities with neighbouring Semitic peoples, such as the Amorites and Moabites, are also evident.

Exodus and conquest. According to biblical tradition, the clan migrated to Egypt because of a famine in the land of Canaan, were later enslaved and oppressed, and finally escaped from Egypt to the desert east of the Isthmus of Suez under a remarkable leader, Moses. The account—a proclamation, celebration, and commemoration of the event—is replete with legendary elements, but present-day scholars tend to believe that behind the legends there is a solid core of fact; namely, that Hebrew slaves who built the fortified cities of Pithom and Rameses somehow fled from Egypt, probably in the 13th century BCE, under a great leader (see also MOSES). A stele (inscribed stone pillar) of the pharaoh Merneptah of that time in which he claims to have destroyed Israel is the first known nonbiblical reference to the people by name. Whether the destruction was in the intervening desert or in Canaan (and whether a true or a false claim) is not clear. The tradition ascribes to Moses the basic features of Israel's faith: a single God, called YHWH, who cannot be represented iconically, bound in a covenant relationship with his special people Israel, to whom he has promised possession of (not, as with their forefathers, mere residence in) the land of Canaan. There is some dispute among scholars as to when such features as the Mosaic Covenant actually emerged and as to which of the traditional 12 tribes of Israel entered Canaan at the end of the period of wandering in the desert.

The work and faith of Moses

The biblical account of the conquest of Canaan is again, from the point of view of historical scholarship, full of legendary elements that express and commemorate the elation and wonder of the Israelites at these events. The conquest of Canaan—according to tradition, a united national undertaking led by Moses' successor, Joshua—was a rather drawn out and complicated matter. Archaeological evidence tends to refute some of the elements of the biblical account, confirm others, and leave some open. According to the tradition, after an initial unified assault that broke the main Canaanite resistance, the tribes engaged in individual mopping-up operations. Scholars believe that Hebrews who had remained resident in Canaan joined forces with the invading tribes, that the other Canaanite groups continued to exist, and that many of them later were assimilated by the Israelites.

The tribal league. The invading tribes who became masters of parts of Canaan, although effectively autonomous and lacking a central authority, considered themselves a league of 12 tribes, although the number 12 seems to have been more canonical or symbolical than historical. Some scholars, on the analogy of Greek leagues of six or 12 tribes or cities with a common sanctuary, speak of the Israelite league as an "amphictyony," the Greek term for such an association; but others hold that there is no evidence that the Israelites maintained a common shrine.

The period of the judges

Certain leaders arose, called judges, who might rule over several tribes, but this arrangement was usually of a local or regional character. However, the stories about such "judges" (who were frequently local champions or heroes, such as Gideon, Jephthah, and Samson), though encrusted with legend, are now thought to be substantially historical. The period from about 1200 to 1020 is called, after them, the period of the judges. It was during this period that Israelite assimilation of Canaanite cultural and religious ideas and practices began to be an acute problem and that other invaders and settlers became a threat to the security of Israel. One of the chief threats was from the Philistines, an Aegean people who settled (*c.* 12th century BCE) on the coast of what later came to be called, after them, Palestine. Organized in a league of five cities, or principalities, the Philistines, who possessed a monopoly of iron implements and weapons, pushed eastward into the Canaanite hinterland and subjugated Israelite tribes, such as the Judahites and Danites, that stood in their way, even capturing the sacred ark from the famous shrine of Shiloh when it was brought into battle against them. The Philistine threat was probably the decisive factor in the emergence of a permanent political (but at first primarily military) union of all Israel under a king—what historians call the united monarchy (or kingdom).

The united monarchy. The monarchy was initiated during the career of Samuel, a prophet of great influence and authority who was also recognized as a judge and is depicted in varying biblical accounts as either favouring or not favouring the reign of a human king over Israel. In any case, he anointed Saul, a courageous military leader of the tribe of Benjamin, as king (*c.* 1020 BCE). Saul won substantial victories over the Ammonites, Philistines, and Amalekites, leading the tribes in a "holy war," and for a time the Philistine advance was stopped; but Saul and his son Jonathan were killed in a disastrous battle with the Philistines in central Palestine. His successor, David, a former aide (and also his son-in-law) who had fallen out of favour with him, at first took over (*c.* 1010) the rule of Judah in the south and then of all Israel (*c.* 1000). Through his military and administrative abilities and his political acumen, David established a centralized rule in Israel, cleared the territory of foreign invaders, and, in the absence of any aggressive foreign empire in the area, created his own petty empire over neighbouring city-states and peoples. He established his capital in Jerusalem, which until then had maintained its independence as a Canaanite city-state wedged between the territories of Saul's tribe Benjamin and David's tribe Judah, and moved the ark there from the small Israelite town in which it had been stored by the Philistines, establishing it in a tent shrine. This felicitous combination of holy ark, political reign, and central city was to be hailed and proclaimed by future ages. Under David's successor, his son Solomon (reigned *c.* 961–922), Israel became a thriving commercial power; numerous impressive buildings were erected, including the magnificent Temple (a concrete symbol of the religio-political unity of Israel); a large harem of foreign princesses was acquired, sealing relations with other states; the country was divided into 12 districts for administrative, supply, and taxation purposes. Foreign cults set up to serve the King's foreign wives and foreign traders led to charges of idolatry and apostasy by religious conservatives. In the latter years of his reign, Solomon's unpopular policies, such as oppressive forced labour, led to internal discontent and rebellion, while externally the vassal nations of Damascus (Aram) and Edom staged successful revolts against his rule. The central and northern tribes, called Israel in the restricted sense, were especially galled by the oppressive policies, and soon after Solomon's death Israel split off to become a separate kingdom. The united monarchy thus became the divided monarchy of Israel (the northern kingdom) and Judah (the southern kingdom).

The centralization of state and religion in Jerusalem

FROM THE PERIOD OF THE DIVIDED MONARCHY THROUGH THE RESTORATION

The divided monarchy: from Jeroboam I to the Assyrian conquest. Jeroboam I, the first king of the new state of Israel, made his capital first at Shechem, then at Tirzah. Recognizing the need for religious independence from Jerusalem, he set up official sanctuaries at Dan and Bethel, at the two ends of his realm, installing in them golden calves (or bulls), for which he is castigated in the anti-northern account in the First Book of the Kings. Israel engaged in conflicts with Judah and, sometimes jointly with Judah, against foreign powers. At first there was great dynastic instability in the northern kingdom, until the accession of Omri (reigned *c.* 884–*c.* 872), one of its greatest kings, who founded a dynasty that lasted through the reign of his two grandsons (to 842). Under Omri an

impressive building program was initiated at the capital, Moab was subjugated (an event confirmed in an extrabiblical source, the Moabite Stone), and amicable relations were established with Judah. The Phoenician kingdom of Tyre was made an ally through the marriage of his son Ahab to the Tyrian princess Jezebel. Ahab (reigned *c.* 874–853 BCE)—unless the episode recounted in I Kings, chapter 20 actually took place four reigns later—fought off an attempt by Damascus, heading a coalition of kings, to take over Israel. Near the end of his reign, Ahab joined with Damascus and other neighbouring states to fight off the incursions of the great Assyrian Empire in their area. Peaceful relations were cemented with Judah through the marriage of Ahab's daughter (or sister) Athaliah to Jehoram, the son of the king of Judah (not to be confused with Ahab's son, Jehoram of Israel). But the establishment of a pagan Baal temple for Jezebel and her attempt to spread her cult aroused great opposition on the part of the zealous Yahwists among the common people. There was also resentment at the despotic Oriental manner of rule that Ahab, incited by Jezebel, exercised. She and her cult were challenged by Elijah, a prophet whose fierce and righteous character and acts, as illumined by legend, are dramatically depicted in the First Book of the Kings. In the reign of Ahab's son Jehoram, Elijah's disciple Elisha inspired the slaughter of Jezebel and the whole royal family, as well as of all the worshippers of Baal, thus putting a stop to the Baalist threat. Jehu, Jehoram's general who led this massacre, became king and established a dynasty that lasted almost a century (*c.* 842–745), the longest in the history of Israel.

Jezebel and the Baal cult: Elijah and Elisha

Meanwhile, in Judah, the Baal cult introduced by Athaliah, the queen mother and effective ruler for a time, was suppressed after a revolt, led by the chief priests, in which Athaliah was killed and her grandson Joash (Jehoash) was made king. In the ensuing period, down to the final fall of the northern kingdom, Judah and Israel had varying relations of conflict and amity and were involved in the alternative expansion and loss of power in their relations with neighbouring states. Damascus was the main immediate enemy, which annexed much of Israel's territory, exercised suzerainty over the rest, and exacted a heavy tribute from Judah. Under Jeroboam II (783–741) in Israel, and Uzziah (Azariah; 783–742) in Judah, both of whom had long reigns at the same time, the two kingdoms cooperated to achieve a period of prosperity, tranquillity, and imperial sway unequalled since Solomon's reign. The threat of the rising Assyrian Empire under Tiglath-Pileser III soon reversed this situation. When a coalition of anti-Assyrian states, including Israel, marched against Judah to force its participation, the Judahite king Ahaz (*c.* 735–720) called on Assyria for protection; the result was the defeat of Israel, which suffered heavily in captives, money tribute, and lost provinces, while Judah became a vassal state of Assyria. In about 721, after an abortive revolt under King Hoshea, the rump state of Israel was annexed outright by Assyria and became an Assyrian province; its elite cadre, amounting to nearly 30,000 according to Assyrian figures, was deported to Mesopotamia and Media, and settlers were imported from other lands. Thus, the northern kingdom of Israel ceased to exist. Its decline and fall were a major theme in the prophecies of Amos, Hosea, Isaiah, and Micah.

The final period of the kingdom of Judah. Meanwhile, the southern kingdom of Judah was to have another century and a half of existence before a similar and even grimmer fate befell it. Hezekiah (reigned *c.* 715–*c.* 686), who instituted a religious reform to return worship to a pure Yahwist form, also displayed political independence, joining a coalition of Palestinian states against Assyria. But the coalition was soon defeated, and Judah—with Jerusalem besieged—bought off the Assyrians, led by Sennacherib, with tribute. In the reign of Manasseh (692–638) there was a revival of pagan rites, including astral cults in the very forecourts of the temple of YHWH, child sacrifice, and temple prostitution; hence, he is usually portrayed as the most wicked of the kings of Judah. If he had any tendencies toward independence from Assyrian domination, they apparently were suppressed by his being taken in chains to Babylon, where he was molded into proper vassal behaviour, although one edifying and probably unhistorical biblical account reports his repentance and attempt at religious reform after his return to Judah. The great religious reform took place in the reign of his grandson Josiah (640–609) during a period when the Assyrian empire was in decline and was precipitated by the discovery of the Book of the Law during the restoration of the Temple. It was proclaimed by the king to be the Law of the realm, and the people pledged obedience to it. In accordance with its admonitions, the pagan altars and idols in the Temple were removed, rural sanctuaries ("high places") all the way into Samaria were destroyed, and the Jerusalem Temple was made the sole official place of worship. (For an identification of the law book with the legal portion of Deuteronomy, see below *Old Testament literature: Deuteronomy.*) Josiah also made an attempt at political independence and expansion but was defeated and killed in a battle with the Egyptians, the new allies of the fading Assyrian Empire. During the reigns of his sons Jehoiakim (*c.* 609–598) and Zedekiah (597–586), Judah's independence was gradually extinguished by the might of the new dominant Babylonian Empire under Nebuchadrezzar. The end came in 586 with the Babylonian capture of Jerusalem and the destruction of the principal buildings, including the Temple and the fortifications. The first deportation of Judahites to Babylon, during the brief reign of Josiah's grandson Jehoiachin in 597, was followed by the great deportation of 586, which was to be a theme of lament and remembrance for millennia to come. (Numerous Jews also migrated to Egypt during this troubled time.) Exhortations and prophecies on the decline and fall of Judah are to be found in Zephaniah, Nahum, Habakkuk, and Jeremiah (who played a significant role in the events), while the conditions and meaning of the exile are proclaimed by Ezekiel and Deutero-Isaiah (chapters 40–55 of Isaiah).

Josiah's reform and the Book of the Law

The Babylonian Exile and the restoration. The Babylonian Exile (586–538) marks an epochal dividing point in Old Testament history, standing between what were subsequently to be designated the pre-exilic and post-exilic eras. The Judahite community in Babylonia was, on the whole, more Yahwist in religion than ever, following the Mosaic Law, emphasizing and redefining such distinctive elements as circumcision and the sabbath and stressing personal and congregational prayer—the beginnings of synagogal worship. It is possible that they also reached an understanding of historical events (like that taught by the great pre-exilic and exilic prophets)—as the chastening acts of a universal God acting in history through Nebuchadrezzar and other conquerors. To this period is also ascribed the beginning of the compilation of significant portions of the Old Testament and of the organizing view behind it. In any event, it was from this community that the leadership and the cadres for the resurrection of the Judahite nation and faith were to come when Cyrus the Great (labelled "the Lord's anointed" in Deutero-Isaiah) conquered Babylon and made it possible for them to return (538). A contingent of about 50,000 persons, including about 4,000 priests and 7,000 slaves, returned under Sheshbazzar, a prince of Judah.

The first great aim was the rebuilding of the Temple as the centre of worship and thus also of national existence; this was completed in 515 under the administration of Zerubbabel and became the place of uninterrupted sacrificial worship for the next 350 years. The next task was to rebuild the walls of Jerusalem, which was undertaken by Nehemiah, a Babylonian Jew and court butler who was appointed governor of Judah and arrived in 444. Nehemiah also began religious reforms, emphasizing tithing, observance of the sabbath, and the prohibition against intermarriage with "foreign" women. This reform was carried through systematically and zealously by Ezra, a priest and scribe who came from Babylon about 400 BCE, called the people together, and read them the "book of the law of Moses" to bring them back to the strict and proper observance maintained in Babylon: circumcision, sabbath observance, keeping the feasts, and, to seal it all, avoiding intermarriage. (In this presentation, modern criti-

Rebuilding of the Temple and religious reform under Nehemiah and Ezra

cal scholarship is being followed, placing Nehemiah before Ezra instead of the traditional sequence, which reverses the positions.) Haggai, Zechariah, and Malachi are the prophets of this restoration period. Ezra and Nehemiah are its narrators.

It was in this period that enmity between the Jews, or Judaeans, as they came to be called, and the Samaritans, a term applied to the inhabitants of the former northern kingdom (Israel), was exacerbated. It has been surmised that this goes back to the old political rivalry between Israel and Judah or even further back to the conflict between the tribes of Joseph and Judah. Scholars ascribe the exacerbation of enmity in the restoration period variously to the Samaritans' being excluded from participating in the rebuilding of the Temple; to Nehemiah's rebuilding of the walls of Jerusalem (regarded as a threatening act by the Samaritan authorities); or to the proscriptions of intermarriage by Ezra. The animus of the Jews against the Samaritans is frequently expressed in the biblical books dealing with the restoration (expressions perhaps engendered by later events), but the attitude of the Samaritans and a good deal else about them is not evident. At some time they became a distinct religious community, with a temple of their own on Mt. Gerizim and a Scripture that was limited solely to the Pentateuch, excluding the Prophets and Writings.

Old Testament history proper ends with the events described in the books of Ezra and Nehemiah. The books of Chronicles give all the preceding history, from Adam to the Babylonian sack of Jerusalem and the exile. The last two verses of the Second Book of the Chronicles are repeated in the first two verses of Ezra: God inspires Cyrus to send the Jews back to Jerusalem to rebuild the Temple. The Persian period of Jewish history ended with the conquest of Alexander the Great in 323 BCE to begin the Hellenistic era, in which some of the biblical (including apocryphal or deuterocanonical) writings were created (for Hellenistic Judaism, see JUDAISM).

Old Testament literature

THE TORAH (LAW, PENTATEUCH, OR FIVE BOOKS OF MOSES)

Composition and authorship. The Torah, or Pentateuch (Five Scrolls), traditionally the most revered portion of the Hebrew canon, comprises a series of narratives, interspersed with law codes, providing an account of events from the beginning of the world to the death of Moses. Modern critical scholarship tends to hold that there were originally four books (Genesis, Exodus, Leviticus, and Numbers) resulting from the division into manageable scrolls—a so-called Tetrateuch—to which later was added a fifth scroll, or book, Deuteronomy. A theory, once widely held, that the Book of Joshua was originally integral with the first five books to form a Hexateuch (Six Scrolls) is now generally regarded as dubious.

The traditional Jewish and Christian view has been that Moses was the author of the five books, that "of Moses" means "by Moses," citing in support passages in the Pentateuch itself that claim Mosaic authorship. Since these claims, however, are written in the third person, the question still arises as to the authorship of the passages; *e.g.*, in Deuteronomy, chapter 31, verse 9: "And Moses wrote this law, and gave it to the priests . . . and to all the elders of Israel." The last eight verses of Deuteronomy (and of the Pentateuch), describing Moses' death, were a problem even to the rabbis of the 2nd century CE, who held that "this law" in the verse quoted refers to the whole Torah preceding it. There are also other passages that seem to be written from the viewpoint of a much later period than the events they narrate.

The documentary hypothesis. Beyond these obvious discrepancies, modern literary analysis and criticism of the texts has pointed up significant differences in style, vocabulary, and content, apparently indicating a variety of original sources for the first four books, as well as an independent origin for Deuteronomy. According to this view, the Tetrateuch is a redaction primarily of three documents: the Yahwist, or J (after the German spelling of Yahweh); the Elohist, or E; and the Priestly code, or P. They refer, respectively, to passages in which the Hebrew personal name for God, YHWH (commonly transcribed "Yahweh"), is predominantly used, those in which the Hebrew generic term for God, Elohim, is predominantly used, and those (also Elohist) in which the priestly style or interest is predominant. According to this hypothesis, these documents—along with Deuteronomy (labelled D)—constituted the original sources of the Pentateuch. On the basis of internal evidence, it has been inferred that J and E are the oldest sources (perhaps going as far back as the 10th century BCE), probably in that order, and D and P the more recent ones (to about the 5th century BCE). Genesis, Exodus, and Numbers are considered compilations of J, E, and P, with Leviticus assigned to P and Deuteronomy to D.

J, E, P, and D documents

The Yahwist, or J, is the master of narrative in biblical literature, who sketches people by means of stories. He takes his materials wherever he finds them, and if some are crude he does not care, as long as they make a good story. The book of Genesis, for example, contains the story of Abraham's passing off his wife as his sister, so if the king took her as a concubine he would honour her supposed brother instead of having her husband killed, a story told by J without any moralistic homily. Not given to subtle theological speculations, J nearly always refers to the Deity as YHWH, by his specifically Israelite personal name (usually rendered "the Lord" in English translations), though he is not hidebound and also employs the term Elohim ("God"), especially when non-Hebrews are speaking or being addressed. He presents God as one who acts and speaks like human persons, a being with whom they have direct intercourse. The Yahwist, however, has one very definite theological (or theo-political) preoccupation: to establish Israel's divinely bestowed right to the land of Canaan.

More reflective and theological in the apologetic sense is the Elohist, or E. No fragment of E on the primeval history (presented in the first 11 chapters of Genesis) has been preserved, and it is probable that none ever existed but that the Elohist began his account with the patriarchs (presented in the remainder of Genesis, in which the J and E strands are combined). The first passage that can be assigned to E with reasonable certainty is chapter 20 of Genesis, which parallels the two J variants of the "She is my sister" story noted above. Unlike these, it tries to mitigate the offensiveness of the subterfuge: though the patriarch did endanger the honour of his wife to save his life, his statement was not untrue but merely (deliberately) misleading. The Elohist is also distinct from the Yahwist in generally avoiding the presentation of God as being like a human person and treating him instead as a more remote, less directly accessible being. Significantly, E avoids using the term YHWH throughout Genesis (with one apparent exception), and it is only after telling how God revealed his proper name to Moses, in chapter 3 of Exodus, that he refers to God as YHWH regularly, though not exclusively. This account (paralleled in the P strand in chapter 6 of Exodus) is apparently based on a historical recollection of Moses' paramount role in establishing the religion of YHWH among the Israelites (the former Hebrew slaves). Also noteworthy is E's choice of the term prophet for Abraham and his characterization of a prophet as one who is an effective intercessor with God on behalf of others. This is in line with his speculations on the unique character of Moses as the great intercessor as compared with other prophets (and also with Joshua as Moses' attendant).

The Elohistic tone and stress

It is inferred from certain internal evidence that E was produced in the northern kingdom (Israel) in the 8th century BCE and was later combined with J. Because it is not always possible or important to separate J from E, the two together are commonly referred to as JE.

The third major document of the Tetrateuch, the Priestly code, or P, is very different from the other two. Its narrative is frequently interrupted by detailed ritual instructions, by bodies of standing laws of a ritual character, and by dry and exhaustive genealogical lists of the generations. According to one theory, the main author of P seems to

have worked in the 7th century and to have been the editor who combined the J and E narratives; for his own part, he is content to add some brief, drab records—with frequent dates—of births, marriages, and migrations. The P material is to be found not merely in Leviticus but throughout the Tetrateuch, including the early chapters of Genesis and one of the creation accounts and ranging from the primeval history (Adam to Noah) to the Mosaic era. Like the Elohist, P uses the term Elohim for God until the self-naming of God to Moses (Exodus, chapter 3, in the P strand) and shows a non-anthropomorphic transcendent stress.

The Deuteronomist, or D, has a distinctive hortatory style and vocabulary, calling for Israel's conformity with YHWH's covenant laws and stressing his election of Israel as his special people (for a detailed consideration of D, see below *Deuteronomy*). To the Deuteronomist or the Deuteronomic school is also attributed the authorship of the Former Prophets (Joshua, Judges, Samuel, and Kings), which scholars call the "Deuteronomic history."

Other Pentateuchal theories. This documentary theory of the composition of the Pentateuch has been challenged by eminent 20th-century scholars who have offered alternative or additional methods of analysis and interpretation. Form criticism, for example, has stressed particular literary forms and the historical setting out of which they arose: the sagas, laws, legends, and other forms and the particular tribal or cultic context that gives them meaning. Tradition criticism centres on the pre-literary sources; *i.e.,* on the oral traditions and the circles out of which they originated as accounting for the variety of the materials in the Pentateuch. Archaeological criticism has tended to substantiate the reliability of the typical historical details of even the oldest periods and to discount the theory that the Pentateuchal accounts are merely the reflection of a much later period. The new methods of criticism have served to direct attention to the life, experience, and religion out of which the Pentateuchal writings arose and to take a less static and literal view of the constituent documentary sources; yet most scholars still accept the documentary theory, in its basic lines, as the most adequate and comprehensive ordering of the variegated Pentateuchal materials. The following presentation rests mainly on an analysis and interpretation of the literary sources. (See below *The critical study of biblical literature: exegesis and hermeneutics.*)

In any case, the five books that have come down in various texts and versions have been seen as a unit in the religious communities that preserved them. Their basic content may be divided thus: (1) beginnings of the world and man—the primeval history; (2) patriarchal narratives—from Abraham to Joseph; (3) Egyptian slavery and the Exodus; (4) the revelation and Covenant at Sinai; (5) wanderings and guidance in the wilderness (divisible into two separate sub-blocks, before and after Sinai); (6) various legal materials—the Decalogue, Covenant Code, and passages of cultic and Deuteronomic laws—interspersed in the narrative, which take up the greater portion of the Pentateuch.

Genesis. This book is called *Bereshit* in the Hebrew original, after its first word (and the first word of the Bible), meaning "In the beginning." It tells of the beginnings of the world and man and of those acclaimed as ancestors of the Hebrew people—all under the shaping action and purpose of God. The book falls into two main parts: chapters 1–11, dealing with the primeval history, and chapters 12–50, dealing with the patriarchal narratives; the latter section is again divisible into the story of Abraham, Isaac, and Jacob (chapters 12–36) and the story of Joseph (chapters 37–50), which may be treated as a unit of its own.

The primeval history. The Bible begins with the creation of the universe. It tells the story with images borrowed from Babylonian mythology, transformed to express its own distinctive view of God and man. Out of primary chaos, darkness, void, depths, and waters God creates the heaven and the earth and all that dwell therein—a coherent order of things—by his will and word alone. He says, "Let there be . . ." and there is. Actually, there are two creation accounts: the first (1–2:4), ascribed to P, simply gives a terse day-by-day account including the culminating creation of man, in the divine "image and likeness," followed by the primordial sabbath on the seventh day. The other (2:4–25), ascribed to J, starts with an arid wasteland and the creation of man (Adam), described specifically as being formed by God out of dust and made into a living thing by God blowing the breath of life into him. He and the woman (Eve) created for him out of his rib are put into a paradisal garden (Eden), especially created for them to till and to tend and to sustain life. The two are forbidden only to eat of the tree of the knowledge of good and evil on pain of death (there is also a tree of life in the middle of the garden). The cosmic setting and concern of the P account is thus followed by the human setting and concern of the J account. Creation is followed by temptation, disobedience, and fall and all that follows from that for the history of mankind. At the instigation of the serpent, the shrewdest of the beasts, who holds out the possibility of attaining godlike knowledge, the woman eats of the fruit of the tree of knowledge and gives some to her husband to eat also. Their distinction from beasts and children manifests itself immediately by a sense of modesty about exposing their bodies, and loincloths become the first products of the higher knowledge. The primal human couple are punished by God for their disobedience by being driven out of the idyllic garden into the world of pain, toil, and death.

The eating of the forbidden fruit

The reason given by YHWH to the divine beings is: "Behold, the man has become like one of us, knowing good and evil; and now, lest he put forth his hand and take also of the tree of life, and eat, and live for ever." These words apparently point back to the polytheistic mythology (the existence of divine, magical powers; the gods' jealousy of mankind; the tree of eternal life; etc.) from which the Yahwist drew his images and symbols explaining man's suffering, frustration, and limitation. In the biblical framework and rendering (and subsequent interpretation), the archaic stories and images acquire a different meaning, suitable to the idea of a transcendent deity and an imperfect mankind.

With the exile from the garden, human history and culture begins. In the story of Adam's sons, Cain and Abel, man has already become a herdsman and farmer, and also a murderer: again probably a reflection of older mythical material and, again, one that puts an emphasis on human sin and estrangement from God. In the story of the Flood that follows there are evident borrowings from the Mesopotamian stories of a flood sent by the gods to destroy mankind, but in the biblical account it is emphasized that man's extreme wickedness is the cause and that Noah is saved along with his family by God's deliberate choice because he is a righteous man. (In the flood story in the Babylonian Gilgamesh epic, by contrast, there is no apparent moral reason why the gods resolved to destroy mankind, and the only reason why the hero of the Flood and his kin are saved is that he is favoured by one of the gods, who tricks the others, including the chief god.) After the Flood, God blesses Noah and bestows on man the earth and the things on it for sustenance and makes a covenant with Noah and all creatures that he will never again unleash a world-destroying flood. The permanent order of the world is assured, and God's blessing and covenant make their first explicit appearance in the Bible.

The Flood and the Tower of Babel

In the story of the Tower of Babel, the final story in the primeval history, a primal unity of mankind in which there is only one language is shattered when, in their pride, men decide to build a city and a tower that will reach up to the heavens. YHWH again takes steps to check dangerous collaboration: He says (to the celestial council), "Come, let us go down, and there confuse their language, that they may not understand one another's speech," and scatters them over the earth. Again, the Yahwist has apparently used ancient mythological motifs to explain the diversity of mankind; the story may be regarded as simply a direct borrowing from the older traditions, without any monotheistic adaptation; in its textual setting, however, it may also be taken as another instance of the ruin of primal harmony by human willfulness and pride.

The patriarchal narratives. The universal primal history

The fathers of the Hebrew people

of man in the first 11 chapters of Genesis is followed by an account of the fathers of the Hebrew people; *i.e.,* of the origins of a particular group. From a literary point of view, this portion may be divided into the sagas of Abraham, Isaac, and Jacob and the story of Joseph. Although these narratives are not historical in the ordinary sense, they have an evident historical setting and refer to various particulars that fit in with what is generally known of the time and area. They apparently rest on the traditions of particular families, clans, or tribes and were probably passed down orally before they took written form. Theologically, they are an account of a divine promise and Covenant and of man's faith and unfaith in response, with Abraham as the model man of faith.

The Elohist, as well as J and P, tells the remarkable story of how God singled out Abraham (Abram) to migrate from Mesopotamia and sojourn in Canaan, promised him that he would make him the ancestor of great nations and that his posterity would inherit the land of his sojournings, and singled out as the heirs to the latter promise first Isaac, Abraham's son by his chief wife, Sarah, and then Jacob, the younger of Isaac's two sons; how Jacob acquired the additional name of Israel and how the wives, children, and children's children who, in Jacob-Israel's own lifetime, came to constitute a family of 70 souls, became the nucleus of the Israelite people; and how it came about that this ethnic group, prior to becoming, as promised, the masters of the land of their sojournings, first vacated it to sojourn for a time in Egypt. Apart from the low-keyed P strand, it is mostly splendid narrative, including the Elohist's account of the (aborted) sacrifice of Isaac by his father in response to God's command, a terse story packed with meaning, and the Joseph story about the son of Jacob who is sold into slavery by his brothers, rises to a high post in the Egyptian court, and ultimately helps his family to settle in Egypt. The 12 sons of Jacob-Israel are eponymous ancestors of Israelite tribes (ancestors after whom the tribes are named); the actions and fortunes of the eponymous ancestors, including certain blessings and other pronouncements of Jacob-Israel, account for the future positions and fortunes of the particular tribes. Though there is less history and more legend, much of the atmosphere of an older age is preserved, with the patriarchs represented as seminomadic, essentially peaceful and pastoral tent dwellers—alien residents—among the settled Canaanites and as observing customs otherwise only attested in Mesopotamia. Anachronistic features, however, insinuate themselves from time to time

The God of the patriarchs

The God of the patriarchs is presented as Yahweh—explicitly by the Yahwist and implicitly by E and P—*i.e.,* as the same God who would later speak to Moses. God apparently was originally the personal, tutelary deity of each of the patriarchs, called by a variety of names and later unified into the one God of Abraham, Isaac, and Jacob. There are various cult legends in this portion of Genesis, etiological accounts of the origins of various cult sites and practices; though probably of Canaanite origin, these all indicate the places and customs held holy by the Israelites and perhaps also by their claimed Hebrew ancestors. There are direct appearances of God to some of the main figures in the narratives, intimate personal communication between men and God. God's particular blessing upon and Covenant with Abraham is the paradigmatic high point, to be referred back to continually in later biblical and post-biblical traditions.

Exodus. The title (in the Greek, Latin, and English versions) means "a going out," referring to the seminal event of the liberation of Israel from Egyptian bondage through the wondrous acts and power of God. The book celebrates and memorializes this great saving event in song and story and also the awesome revelation and covenant at Mt. Sinai. The contents of the book may be summarized thus: (1) Israel in Egypt, (2) the Exodus and wanderings, (3) the Covenant at Sinai, (4) the apostasy of the people and renewal of the Covenant, and (5) the instructions on building the Tabernacle and their execution.

The God of the Exodus

Redemption and revelation. Significant in the early chapters is God's special concern for the Hebrew slaves, his reference to them as "my people," and his revelation to Moses, the rebel courtier whom he has picked to be their leader, that he is YHWH, the God of their fathers, an abiding presence that will rescue them from their misery and bring them into Canaan, the land of promise. This assurance is repeated at the critical moments that follow (*e.g.,* "And I will take you for my people, and I will be your God"). In the series of frustrations, obstacles, and redeeming events that are narrated, God's special causal power and presence are represented as being at work. God hardens the Pharaoh's heart, sends plagues that afflict the Egyptians but spare the Hebrews, causes the waters to recede in the Sea of Reeds (or Papyrus Marsh) to permit passage to the fleeing Israelites and then to engulf the pursuing Egyptians ("the horse and his rider he has thrown into the sea"), and gives the people guidance in their wandering in the wilderness. The cryptic "name" that God gives to himself in his revelation to Moses (*'ehye 'asher 'ehye*), often translated "I am that I am" or "I will be what I will be," may also be rendered "I will cause to be that which I will cause to be." In either case, it is a play on, and an implied interpretation of, the name YHWH.

The constancy of God's directive power and concern is displayed notably in the period (40 years) of wilderness wandering (on the eastern and southern borders of Canaan), when Israel is tested and tempered not only by hardship but also by rebellious despair that looks back longingly to Egyptian bondage (see also below *Numbers*). God sends the people bread from heaven (manna) and quail for their sustenance (J and P strands) and, through Moses, brings forth hidden sources of water (JE strand). When the Amalekites (a nomadic desert tribe) attack, Moses, stationed on a nearby hill, controls the tide of battle by holding high the rod of God (a symbol of divine power), and when the enemy is routed he builds an altar called "The Lord is my banner" (E strand). Also inserted here is the account (E) of the visit of Moses' father-in-law, Jethro, a priest of another people (Midianite) who, impressed by YHWH's marvellous deliverance of Israel, blesses, extols, and sacrifices to him—under the name Elohim, but in the context the same God is clearly meant.

The Covenant at Mt. Sinai

God's power and presence manifest themselves impressively in the culminating account of the Covenant at Mt. Sinai (or Horeb). The people, forewarned by God through Moses, agree beforehand to carry out the terms of the Covenant that is to be revealed, because God has liberated them from Egypt and promises to make them his special holy people; they purify themselves for the ensuing Covenant ceremony, according to God's instructions. Yahweh appears in fire and smoke, attended by the blare of a ram's horn at the top of the mountain, where he reveals to Moses the terms of the Covenant, which Moses then passes on to the people below. Here follow in the text the Ten Commandments and the so-called Covenant Code (or Book of the Covenant) of lesser, specific ordinances, moral precepts, and cultic regulations, accompanied by a promise to help the people conquer their enemies if they will serve no other gods. After this comes the Covenant ceremony with burnt offerings and the sacrifice of oxen, with the blood of the animals thrown both on the altar and on the people to sacramentally seal the Covenant, followed by a sacral meal of Moses and the elders at the mountaintop, during which they see God. Many modern scholars hold that this is presented as the initial form of a Covenant renewal ceremony that was repeated either annually or every seven years in ancient Israel.

The stone tablets inscribed by God and Moses

There are certain problems and apparent discrepancies in this account that are explained by critical scholarship as deriving from the combination of different sources, mainly J and E, traditions, or emphases. In the opening portion (chapter 19) the people are gathered at the foot of the mountain so as to hear and meet God, and Moses himself brings down to them God's words. In a later portion (24:12–18, also 32:15–20), after the sacral meal, Moses goes up on the mountain to receive "the tables of stone, with the law and commandments," inscribed by God himself, and returns with two stone tablets written on both sides by the hand of God—which he breaks in anger at the people's worship of the molten calf that has developed in his absence. Later (chapter 34), at God's command, Moses

cuts two new stone tablets, upon which after hearing God's various promises and exhortations, he writes "the words of the covenant, the ten commandments"; finally, he brings the new tablets down to the people and tells them what YHWH has commanded. There seem to be two parallel accounts of the same event, woven together by the skillful redactor into a continuing story. There also seem to be two distinct strands in the account of the sealing of the Covenant in the first 11 verses of chapter 24. According to one, the elders are to worship from afar, and only Moses is to come near YHWH; in the other strand, as noted, the elders eat the sacred meal on the mountaintop in the direct presence of God.

Legislation. The book of Exodus includes not only the narrative and celebration of God's redemptive action in the Exodus and wanderings and his revealing presence at Mt. Sinai but also a corpus of legislation, both civil and religious, that is ascribed to God and this revelation event. The Covenant Code, or Book of the Covenant, presented in chapters 20–23, immediately following the Decalogue (Ten Commandments), opens with a short passage on ritual ordinances, followed by social and civil law applying to specific situations (case law), including the treatment of slaves, capital crimes, compensation for personal injuries and property damage, moneylending and interest, precepts on the administration of justice, and further ritual ordinances. Scholars generally date this code in the later agricultural period of the settlement in Canaan, but some hold that it is analogous to more ancient Near Eastern law codes and may go back to Moses or to his time. In any case, it seems to be a compilation from various sources, inserted into and breaking the flow of the narrative.

Instructions on the Tabernacle. Also interspersed in the story (chapters 25–31) are God's detailed instructions to Moses for building and furnishing the Tabernacle, the clothing and ordination of priests, and other liturgical matters. According to this segment (evidently P in inspiration), an elaborate structure is to be set up in the desert, in the centre of the camp, taken apart, transported, and assembled again, like the simple "Tent of Meeting" outside the camp, where Moses received oracular revelations from God. Indeed, the two concepts seem to have fused and the Tabernacle is also called the Tent of Meeting. Its prime function is to serve as a sanctuary in which sacrifices and incense are offered on altars and bread presented on a table; it is also equipped with various other vessels and furnishings, including a wooden ark, or cabinet, to contain the two tablets of the Covenant—the famous ark of the Covenant. It is, moreover, to be the place of God's occasional dwelling and meeting with the people. Scholars believe that the elaborate details and materials described stem from a later, Canaanite, period but that the essential concept of a tent of meeting goes back to an earlier desert time. An account of the execution of the instructions for the building of the Tabernacle is presented in chapters 35–40 (following the apostasy, tablet breaking, and Covenant-renewal episodes), which duplicates to the letter the instructions in chapters 25–31. After the Tabernacle is completed and consecrated, it is occupied by the "glory," or presence, of YHWH, symbolized by a cloud resting upon it. It is on this note that the book of Exodus ends.

The portable "Tent of Meeting" and elaborate sanctuary

Leviticus. The cultic and priestly laws presented in Exodus are expanded to take up virtually the whole of Leviticus, the Latin Vulgate title for the third of the Five Books of Moses, which may be translated the Book (or Manual) of Priests. With one exception (chapters 8–10), the narrative portions are brief connective or introductory devices to give an ostensibly narrative framework for the detailed lists of precepts that provide the book's content. The source of Leviticus, both for the legal and narrative passages, is definitely identified as P; it is the only book in the so-called Tetrateuch to which a single source is attributed. Apparently the book consists of materials from various periods, some of them going back to the time of Moses, which were put together at a later date, possibly during or after the Babylonian Exile. Recent scholarship tends to emphasize the ancient origin of much of the material, as opposed to the previous tendency to ascribe a late, even post-exilic date. Despite its content and its dry, repetitive style, many interpreters caution against taking Leviticus as merely a dull, spiritless manual of priestly ritual, holding that it is strictly inseparable from the ethical emphasis and spiritual fervour of the religion of ancient Israel. It is in Leviticus that the so-called law of love, "You shall love your neighbour as yourself," first appears. The rituals set forth drily here probably presuppose an inward state in offering to God, as well as humanitarian and compassionate ethics.

The book may be divided thus: chapters 1–7, offerings and sacrifices; chapters 8–10, inauguration of priestly worship; chapters 11–16, purification laws; chapters 17–26, holiness code; chapter 27, commutation of vows and tithes.

Offerings, sacrifices, and priestly worship. The first verse attributes these regulations to YHWH, who speaks to Moses from the Tent of Meeting, beginning with the rules for offerings by the individual layman. These include burnt, cereal, peace, sin, and guilt offerings, all described in precise details. The prescription for priestly offerings is about the same, with some slight differences in the order of actions, and is presented much more briefly. In chapters 8–10 the narrative that was interrupted at the end of Exodus is resumed, and the ordination of Aaron and his sons by Moses, before the people assembled at the door of the Tent of Meeting is described, as are various animal sacrifices by Aaron and his sons under Moses' direction and the subsequent appearance of God's "glory" to the people. Aaron's two older sons are burned to death by fire issuing forth from God because they have offered "unholy fire." This story apparently emphasizes the importance of adherence to the precise cultic details, as does also the account (at the end of the chapter) of Moses' anger at Aaron's two remaining sons for not eating the sin offering. These stories were apparently used by the priestly authors to buttress the authority of the Aaronic priesthood.

The Aaronic priesthood

Purification laws. With chapter 11 begin the regulations on ritual cleanness and uncleanness, starting with animals and other living things fit and unfit to eat—the basis of the famous Jewish dietary laws. Then come the uncleanness and required purification of women after childbirth, skin diseases, healed lepers, infected houses, and genital discharges. Chapter 16, which belongs in the narrative flow immediately after chapter 10, describes the priestly actions on the Day of Atonement, the culmination of ritual cleansing in Israel. It is a chapter rich in details on Israelite ritual and bound up with the salient religious theme of atonement.

The Holiness Code. Next (chapters 17–26) comes what has been designated the "Holiness Code," or "Law of Holiness," which scholars regard as a separate, distinctive unit within the P material (designated H). It calls upon the people to be holy as God is holy by carrying out his laws, both ritual and moral, and by avoiding the polluting practices of neighbouring peoples; and it proceeds to lay down laws, interspersed with exhortations, to attain this special holiness. Although many scholars tend to date its compilation in the exilic period, some see evidence that it was compiled in pre-exilic times; in any case, the consensus is that the laws themselves come from a much earlier time.

These—a most miscellaneous collection—begin with injunctions on the proper (kosher) slaughtering of animals for meat; go on to a list of precepts against outlawed sexual relations (incest, homosexuality) and an injunction against defiling the (holy) land; proceed to a list of ethical injunctions, including the law of love and kindness to resident aliens, all interspersed with agronomic instructions and warnings against witchcraft; and then, after an injunction against sacrificing children, return to the listing of illicit sexual relations and the warning that the land will spew the people out if they do not obey the divine norms and laws. There follow special requirements for preserving the special holiness of priests and assuring that only unblemished animals will be used in sacrifices; instructions on the observance of the holy days—the sabbath, feasts, and festivals; commands on the proper making of oil for the holy lamp in the Tent of Meeting and of the sacred shewbread, to which are appended the penalties for blasphemy and other crimes; and finally, rules for observance

of the sabbatical (seventh) and jubilee (50th) years, in which the land is to lie fallow, followed by rules on the redemption of land and the treatment of poor debtors and Hebrew slaves.

Punishment and atonement for iniquity

This miscellany, presented in chapters 17–25, is followed by a final exhortation, in chapter 26, promising the people that if they follow these laws and precepts, all will go well with them but warning that if they fail to do so all kinds of evil will befall them, including exile and the desolation of the Promised Land. Yet, if they confess their iniquity and atone for it, God will not destroy them utterly but will remember his Covenant with their forebears. Such a passage points to a later time but not necessarily to the exilic period, as some commentators have assumed. The chapter concludes: "These are the statutes and ordinances and laws which the Lord made between him and the people of Israel on Mt. Sinai by Moses," connecting these precepts with the primal revelation in Exodus.

Commutation of vows and tithes. In the final chapter of Leviticus (27), the P material is resumed with a presentation of the rules for the commutation of votive gifts and tithes. It provides for the release from vows (of offerings of persons, animals, or lands to God) through specified money payments. Some commentators understand the vow to offer persons to refer originally to human sacrifice, others as pledging their liturgical employment in the sanctuary. Special provisions are made for the poor to relieve them from the stipulated payments. Only grain and fruit tithes, not animal tithes, are redeemable. This chapter and the book of Leviticus end, like chapter 26, with the verse, "These are the commandments which the Lord commanded Moses for the people of Israel on Mount Sinai."

Numbers. In the Hebrew Bible this book is entitled Bemidbar (In the Wilderness) after one of its opening words, while in English versions it is called Numbers, a translation of the Greek Septuagint title Arithmoi. Each of the titles gives an indication of the content of the book: (1) the narrative of "40 Years" of wanderings in the wilderness, or desert, between Sinai and Canaan; and (2) the census of the people and other numerical and statistical matters, preceding and interspersing that account. It is a composite of various sources (J, E, and predominantly P) and traditions, which as a whole continue the story of God's special care and testing of his people in the events of the archaic period that formed them. Numbers continues the account of what many modern scholars call the "salvation history" of Israel, which apprehends and narrates events (or the image and impact of events) as involving divine action and direction.

The wilderness wanderings and the census of the people

The book may be divided into the following sections: (1) the conclusion of the Sinai sojourn (1:1–10:10), covering 20 days; (2) the wanderings in the desert of Paran (10:11–20:13), covering 38–40 years; and (3) the events in Edom and Moab (20:14–36:13), covering five months.

The conclusion of the Sinai sojourn. The book opens with a command from God to Moses, early in the second year after the Exodus, to take a census of the arms-bearing men over 20 in each of the clans of Israel. Moses and Aaron, aided by the clan chiefs, take the count, clan by clan, and reach a total of 603,550 men—according to critical scholars, an unbelievably large total for the time and conditions. The Levites, to whom is entrusted the care of the Tabernacle and its equipment, are exempted from this secular census and are counted in a later census, of males one month and over, along with a census of firstborn males from other tribes. The Lord had required that the latter be consecrated to him when he slew all the firstborn of the Egyptians but spared those of the Israelites; now the bulk of them were released by the Levites being taken in their stead to minister to the priests, while for the excess of firstborn over Levites "redemption" payments were collected. A further census of men 30–50 years old is taken among the Levite clans, so as to assign them their various duties, which are here stipulated. Also specified are the positions of the tribes (separated into four divisions of three tribes each) in the camp and on the march, with an assignment of specific portions of the Tabernacle and its equipment to be carried by the Levite clans. YHWH is to give the signal to break camp by lifting the cloud by day or the fire by night from above the Tabernacle and then to advance it in the direction the people are to march. YHWH's signal is to be followed by a blast by the priests (Aaron's sons) on two specially made silver trumpets.

YHWH's signal: a cloud by day and a fire by night

The above directions are set forth in chapters 1–4 and 9–10 (through verse 10). There are intervening chapters containing various materials: expelling leprous or other unclean persons from the camp, the ordeal for a woman suspected of adultery, regulations for Nazirites (those who take special ascetic vows), the offerings brought at the dedication of the Tabernacle, and the purification of the Levites preparatory to taking up their special sacred functions. The priestly emphasis of the materials in chapters 1–10 is evident, and it is also clear that there are various strands of priestly interpretation involved.

Wanderings in the desert of Paran. This section apparently combines various traditions of how the Israelites came into Palestine, and J, E (or JE), and P sources have been discerned in these chapters. The traditional "40 years" in the wilderness (38 or 39, according to critical calculations) were spent mostly in the wilderness of Paran, with a short stay in the oasis of Kadesh, according to P; while, according to J, they spent most of their time in Kadesh; and chapter 13, verse 26, puts Kadesh in the wilderness of Paran, thus encapsulating both traditions. The discrepancy may stem from two separate traditions of how the tribes entered Canaan: from the south or from the north through Transjordan.

The P narrative begins (chapter 10, verse 11) with the lifting of the cloud from the Tabernacle and the setting out of the Israelites for the Promised Land, with their holy Tabernacle and ark, in the order prescribed in chapter 2. According to the P account (verses 11–28), the cloud settles down over the wilderness of Paran, the signal to make camp; whereas in the JE account (verses 29–36) it is the ark of the Covenant that goes ahead to seek out a stopping place, and where it stops the Israelites rest, the cloud simply accompanying them overhead (perhaps to shield them from the blazing desert sun). Chapters 11–12 (JE) deal with the complaints of the people about their hardships and the rebellion of Miriam and Aaron against their brother Moses. When the people express their longing for the good food they had in Egypt and their disgust with the unvarying manna, God sends them a storm of quail, which remain uneaten because he also sends them a plague. This is a somewhat different account from that in Exodus, but the point is the same: the mighty, infinite power of God (chapter 11, verse 23). (Also inserted here is the story of God visiting his spirit on 70 selected elders so that they may share Moses' burdens.) When Miriam and Aaron question God's speaking only through Moses, God proclaims his unique relation with Moses, who alone receives direct revelations from God, not indirectly through dreams and visions, like the prophets.

Chapters 13–14 tell of the despatch of spies from Paran to reconnoiter Canaan and of the despair, rebellion, and unsuccessful foray of the people in response to the spies' reports. Scholars discern two separate accounts of the spying incident artfully woven together. According to the JE account, the spies go only as far as Hebron in the south and return with a glowing report of a fertile land, which is, however, they warn, too strongly defended to be taken from that quarter: only one spy, Caleb, advocates attacking it. In the P account the spies reconnoiter the whole country and give a pessimistic report of it as a land that "devours its inhabitants," who are, moreover, giants compared to the Israelites. The people cry out in despair at this report and want to go back to Egypt, while Caleb and Joshua (added by P) plead with them to trust in God and go forward to take the land. God, disgusted with the people, condemns them to wander in the wilderness for 40 years and decrees that only their children, along with Caleb and Joshua, shall enter into the land of promise. Ruefully, the people now decide to attack and go forth, against Moses' warning, to a resounding defeat.

The spies' report and the people's rebellion

Chapter 15 is a P document or addition, setting forth various ritual regulations. Chapters 16–18 deal with the comparative rights and duties of priests and Levites. Chapter 16 is a composite document dealing with revolts

against Moses and Aaron by certain Levites who question their special authority in a community where all are holy, as also by certain Reubenites who resent Moses' leadership. The dispute is settled when 250 revolting Levites attempt to offer incense (a priestly Aaronic function) and are consumed by fire sent by God, while the leaders of the revolt are swallowed up in the earth. Yet the stubborn people continue their complaint against Moses and Aaron, bringing forth the Lord's anger and a plague, from which they are saved by Aaron's (proper and effective) offering of incense. This latter incident occurs in chapter 17 in the Hebrew text and Jewish translations but concludes chapter 16 in some Christian versions. Chapter 17 in both arrangements, with its story of Aaron's rod, associates Levitical with Aaronic authority; Aaron's name is inscribed on the staff of Levi, which alone among the staffs of the chiefs of the tribes of Israel blossoms and bears fruit, thus authenticating Aaron's, and thereby the Levites', special claims. The relative functions and payments (tithes) of priests and Levites are prescribed in chapter 18. Chapter 19, inserted here, has to do with purification from uncleanness incurred through touching the dead, accomplished through washing in water mixed with the ashes of a red heifer.

Events in Edom and Moab. Chapter 20, verse 14, resumes the narrative of Israel's onward march, starting with their arrival in the wilderness of Zin and stay at Kadesh, marked by Miriam's death and God's exclusion of Moses and Aaron from entering the Promised Land because of their ascribed lack of confidence in God when Moses drew forth water from a rock in response to still more Israelite complaints, but did so in anger and impatience, striking the rock twice with his rod, instead of telling it to give forth water, as the Lord had instructed (the incident of the waters of Meribah). Refused permission by the King of Edom to pass through that land, over the much-used King's Highway, they proceed from Kadesh to Mt. Hor, where Aaron dies and is succeeded by his son Eleazar, and from which they proceed (chapter 21) to bypass Edom in an attempt to approach Canaan from the east. Arrived at the border of what was geographically part of Moab but politically the Amorite kingdom of Sihon, they are refused passage and proceed to defeat the Amorites and take possession of their land. This is from the JE strand of the composite narrative; the P strand does not recognize the existence of settled and politically organized populations between Kadesh and the plains of Moab.

Balaam, the pagan prophet of YHWH

At this point, in chapters 22–24, apparently a very mixed composite of various J and E strands, is presented the fascinating story (or collection of stories) of the non-Israelite seer, or prophet, Balaam, from the region of the Middle Euphrates. Alarmed at the Israelite host encamped at his border, the King of Moab commissions the seer Balaam to put a curse on them, but Balaam refuses, at the order of YHWH, who is also the God of Balaam. On three occasions at the King's request Balaam seeks an oracle from God against Israel, but each time, to the King's rage, he is told by the Lord that Israel is graced with the divine blessing and cannot be cursed. The seer, who is ordered back to his own country, without payment by the disgruntled King, offers a final, unsolicited oracle prophesying the destruction of Moab and other nations by Israel's might: "I will let you know what this people will do to your people in the latter days."

Chapter 25 (combining JE and P strands) provides a lurid interlude in which the Israelites go whoring after Moabite women and offer sacrifices and worship to their god, Baal of Peor. Phinehas, the son of Eleazar, is so incensed at the sight of an Israelite consorting with a Midianite woman that he kills them both, thus ending a plague that has broken out and earning God's special favour: a covenant of perpetual priesthood with him and his descendants (a forward reference to the Zadokite priesthood of post-exilic times). This account is connected by the last two verses with God's call for Israel to harass and smite the Midianites (see below). After the plague ends, in the account (P) in chapter 26, a second census of arms-bearing men and of the Levites is taken, and again a fantastically large total, 601,730, is given, perhaps referring to a much later time. It is noted at the end that all of the previous 603,730 had died in the wilderness, as prophesied, except for Caleb and Joshua, who have been especially picked out by God. This census, coming at the end of the 40-year period of wilderness wanderings, is for the purpose of allotting lands to the various tribes and families. Hence the logical positioning of the passage (P) in the first 11 verses of chapter 27 assuring that a family may inherit through a daughter when there is no son and through a brother when there are no children and through the closest relative when there are neither.

At this point (chapter 27, verse 12) comes the impressive and poignant passage (also P) in which Moses ascends the heights, at God's bidding, to look over the Promised Land, which he is not to enter, and calls on God to appoint a leader to succeed him. At God's command, Moses selects Joshua, and before the priest Eleazar and the whole community he lays his hands on him and commissions him to lead Israel. It is noteworthy that Joshua is invested only with some of Moses' authority and is to learn God's will through Eleazar and the sacred lot (Urim), not directly, as did Moses.

Sacrificial rules for holy days

Again, the narrative is interrupted by three chapters (P) dealing with various religious regulations. Chapters 28–29 stipulate the sacrifices to be made by the whole community daily, on the sabbath, at the new moon, and on these holidays: the Feast of Unleavened Bread (Passover), the Feast of Weeks (Shavuot), The Feast of Trumpets, *i.e.*, New Year (Rosh Hashana), the Day of Atonement (Yom Kippur), and the Feast of Tabernacles (Sukkot). The last two verses of chapter 29 specify that these public offerings are in addition to individual offerings, such as those specified in chapter 15. Critical scholars hold that these elaborate regulations stem from a much later (post-exilic) period, though they may go back to very ancient practices. Some see them as a liturgical commentary on chapter 23 of Leviticus, which presents the cycle of feasts and festivals (see above *Leviticus*). Chapter 30 gives women special exemption from keeping vows (presumably of offerings or abstinence) when countermanded by a father or husband; only widows or divorcees are bound, like men, unconditionally to keep their vows.

Chapter 31, likewise from P, deals with the annihilation of the Midianites following God's command at the end of chapter 25. The Israelites, a thousand from each tribe, go forth to battle led by the priest Eleazar, who carries the sacred vessels and the trumpets. They kill every man and seize all the movable property but spare the women and children. Moses, however, orders every male child and all nonvirgin women killed. There follow instructions for purification for the stain caused by killing a person or touching a dead body and for the distribution of the booty, which includes sheep, cattle, asses, and 32,000 virgins. The rules are that half of the spoils go to the fighting men, half to the rest of the people; in addition, the Lord's share is allotted thus: one five-hundredth of the fighting men's portion goes to the priest, and one-fiftieth of the people's portion goes to the Levites. Scholars are inclined to treat this chapter as a piece of fiction intended really to set forth the rules for purification and dividing the spoils through an invented story. The seer-diviner Balaam is here (verse 16) blamed for the whoring and apostasy incidents in chapter 25; but texts providing his connection with these events are lacking.

Special settlement of Reuben and Gad east of the Jordan

Chapter 32, dealing with the settlement east of the Jordan, concludes the narrative portion of Numbers and thus of the Tetrateuch (a story that is continued in chapter 34 of Deuteronomy and in the Book of Joshua). This very composite account (JEP) tells how the tribes of Reuben and Gad, after an initial angry remonstrance from Moses, are granted permission to settle in the rich pasturelands east of the Jordan on the assurance that after they erect sheepfolds and fortified towns for their flocks and families, they will provide the shock troops spearheading the advance of the Israelites into Canaan, and will not return to their homes until their brethren hold the land. Thereupon Moses allots the various conquered kingdoms and towns east of Jordan to the Gadites and Reubenites. The various Gadite, Reubenite, and Manassite towns are listed.

The rest of the book of Numbers (P in its final form)

consists of an itemized summary of the route from Egypt to the plains of Moab outside Canaan (chapter 33) and various additional materials (chapters 34–36). Verses 50–56 of chapter 33 present the divine command to dispossess the people of Canaan, destroy their idols and cultic places, and apportion the land to each clan by lot. In chapter 34 the Lord specifies the boundaries of the whole land of Canaan that is to be Israel's inheritance and names the tribal leaders who, along with Eleazar and Joshua, are to oversee the division of the land by lot. In chapter 35, the Lord orders 48 towns with extensive pasturelands to be set aside for the Levites; six of these are to be cities of refuge for manslayers whose guilt of intentional murder has not yet been determined and who are provided sanctuary from the traditional blood vengeance. Although these settlements do not constitute an independent tribal territory but are scattered through the territories of the other tribes, the contradiction with chapter 18, verse 24, of Leviticus, commanding that the Levites are to have no share of the land but are to subsist solely on tithes, is obvious and raises critical questions. Finally, chapter 36 concludes the book of Numbers with a supplement to the law of inheritance through daughters laid down in chapter 27, enjoining daughters from marrying outside the tribe, so that the tribe will hold its portion of the land, which was given from God, in perpetuity. As before, the general injunction is laid down in a story dealing with a particular case (the daughter of Zelophehad).

Deuteronomy. *Special nature and problems.* The English title of this work, meaning "second law," is derived from a faulty Greek translation of chapter 17, verse 18, referring to "a copy of this law": the implication being that the book is a second law or an expanded version of the original law for the new generation of Israelites about to enter Canaan. Hebrew texts take the opening words of the book as title, Ele ha-Devarim (These Are The Words), or simply Devarim (Words). As noted in *Composition and authorship,* above, the book is in a class by itself in the Pentateuch, so much so that modern scholars tend to consider it apart from the other four books, and some see it in style, content, and concerns more closely related to the succeeding books of Joshua, Judges, Samuel, and Kings, constituting a "Deuteronomic history." In spite of its homogeneous style and tone—it is assigned for the most part to a single source, D—the content indicates to critical scholars very composite traditions, ages, and situations behind the finished form. This book has elicited a library of scholarship going back to the early 19th century, not only because of the complicated critical and historical problems calling for solution but also because of its spiritual and theological message, which gives it a special place among Old Testament writings.

In form, the book is ostensibly a discourse by Moses "to all Israel" in the final month in Moab before they go over the Jordan into Canaan. Actually it comprises three separate discourses, a set of laws, two poems, and various other matters, all ascribed to Moses directly—here it is Moses who sets forth the laws, not God through him. These materials are centred on the presentation of the rules of life and worship for the coming stay in the Promised Land, along with exhortations and explanations pointing to YHWH, the marvellous liberator from Egypt and guide in the wilderness, as the divine source and reason for the commands. The traditional view was that, with the possible exception of the account of Moses' death, the whole book was written by Moses, based on the phrase "And Moses wrote this song" in chapter 31, verse 22.

Deuteronomy identified with Josiah's "book of the law"

Some early Church Fathers identified the book with "the book of the law" (II Kings, chapter 22, verse 8), found in the 18th year of King Josiah's reign (*c.* 621 BCE), and made the basis of his great religious reform the following year. Wilhelm M.L. de Wette, a German biblical scholar, in 1805 established the predominant modern view that Deuteronomy (or its nucleus, or main portion) was found in Josiah's time and was a distinctive book, separate from the Tetrateuch. He also held that it was composed shortly before its discovery; other, more recent, scholars would put it as much as a century earlier and connect it with earlier reforms, while some associate it with the writings and teachings of the 8th-century-BCE prophet Hosea and with the E source. Furthermore, the references to localities near Shechem as cultic places, taken with certain passages in Joshua, indicate a northern provenance for the book and not the southern source connected with a cultic centre at Jerusalem, as had been previously supposed from the associated material in II Kings. Some scholars see the form and occasion of Deuteronomy as a Covenant renewal ceremony in which the whole law is read, as in Joshua, chapter 8, verses 30–35, and thus view it as a liturgical document, as well as a lawbook. In any case, the tendency is to see various layers of materials and lines of transmission, perhaps going back to quite early preliterary sources, before its final formation in the 8th or 7th century BCE.

The book may be divided as follows: (1) introductory discourse to the whole book (chapter 1 to chapter 4, verse 43); (2) introductory discourse to the lawbook (chapter 4, verse 44, through chapter 11); (3) the lawbook (chapters 12–28); (4) concluding exhortation and traditions about the last days and death of Moses (chapters 29–34).

First introductory discourse of Moses. The first introductory discourse, spoken by Moses, traces the journey of the Israelites from Mt. Horeb to Moab, with some noticeable differences in detail from the account in Exodus and Numbers and an emphasis on Moses being banned from entrance into the Promised Land because the Lord was angry at the Israelites. To this historical retrospect is appended an exhortation to the people to obey God's laws and norms, recalling the imageless God of the revelation and Covenant at Horeb as a warning against making images and serving man-made gods. The uniqueness and soleness of the God of the Exodus and Covenant, his power and presence in his marvellous acts of redemption and revelation, and his gracious selection of Israel are proclaimed in rhetorical questions; moreover, it is emphasized that the God of Israel ("YHWH your God") "is God in heaven above and on the earth beneath; there is no other." The injunctions against idolatry appear to come from later experience and religious crisis in Canaan. The fact that other nations have their own gods and objects of worship is recognized elsewhere in Deuteronomy.

The Shema: "Hear, O Israel"

Second introductory discourse. The second discourse, also ascribed to Moses, again refers to the Covenant at Horeb and sets forth the Ten Commandments, which the people are admonished to obey rigorously, emphasizing the mediating function of Moses at Horeb between the awesome divine presence and the awestruck people. Israel is further admonished to obey the law through wholehearted love of God, expressed in what became the central liturgical expression of Israel's faith, beginning, "Hear, O Israel: The Lord is our God, the Lord Alone. You must love the Lord your God with all your heart and with all your soul and with all your might." If they obey God's laws, avoid other gods, and do what is right and good, they will possess the land promised by God—him who rescued them from Egypt and has brought them thus far. They are to avoid marriage and all other intercourse with the peoples of the land, utterly destroying them and their idolatrous altars and cultic places, for they are a special, holy people chosen by God out of all the peoples because of his love, not because of their greatness or power. This marvellous love will continue to be exercised, and the people will be blessed with all good things—prosperity, fertility, health, and success in battle—if they obey God's ordinances. They are urged to remember the 40-year period of wilderness wandering, in which they were tested (disciplined) by God through hardship and hunger (to find out whether or not they would keep his commands) and saved by him: man does not live by bread alone but, rather, by whatever God provides (*e.g.,* manna from heaven). Another time of testing will come when they live in the rich, fertile land of Canaan and eat their fill and perhaps forget the Lord and his laws, ascribing their wealth to their own power and might and even venturing into idolatrous worship of the gods of the land. If they do so they shall perish, just as the idolatrous nations of the land shall.

A long list of the apostasies of Israel is presented in chapter 9 to demonstrate the point that Israel is going in to possess the land of Canaan not through any virtue of

their own but because of God's promise to the patriarchs. This is followed in chapter 10 by a moving declaration of what God requires of Israel—fear (reverence), walking in his ways, love, wholehearted service, and keeping his commandments—and an extolling of the wondrous, unique, powerful God who liberated them from Egypt. Chapter 11 extols the richness of the land of Canaan and describes how it will bloom for them if they are observant of God's commandments and promises that they will hold the territory from the wilderness to Lebanon and from the Euphrates to the western sea (Mediterranean). It closes with the choice set before them by Moses of "a blessing and a curse"—the former if they obey the commandments, the latter if they do not. This choice is posed to them immediately before the presentation of the laws and norms beginning in chapter 12.

The lawbook. The laws are the central core and purport of the book of Deuteronomy. They are couched in a hortatory, sermonic style that has led to their being categorized as preached law. Emphatic statements of what must or must not be done are connected with exhortations to fulfill these injunctions, pointing to the motivations and spirit in which they should be carried out. There is a wide variety of laws here—ritual, criminal, social—but they are all set within this preaching context and aimed at the service of God. This is no dry legal code but, rather, a book written in fluent and moving prose. Scholars have seen duplications and parallels between the laws presented here and those in the Covenant Code in chapters 21–23 of Exodus; but to this a common source may be ascribed, and Deuteronomy may be considered a work in its own right and not a mere expansion of the Covenant Code.

The basic injunction: a central sanctuary

The lawbook comprises chapters 12–26, supplemented by chapters 27–28. After an initial order to destroy the pagan cultic places and idols, the lawbook goes to its basic injunction: to set up a single central sanctuary in Canaan, where all Israel is to make their offerings, as distinct from the present unregulated practice, "every man doing whatever is right in his own eyes." The spot is designated only "the place which the Lord your God will choose," which some interpreters, following King Josiah, have understood to be Jerusalem and which others understand to be Shechem. (The blessing and curse passage immediately preceding in chapter 11 specifies Mts. Gerizim and Ebal, on either side of Shechem, as the places of blessing and curse, respectively; and an even more elaborate ritual is prescribed for the same locality in chapter 27.) Instructions are given for the proper killing of animals for food, previously connected with the sacrificial cult, and the people are admonished when they settle in Canaan not to inquire about how other nations serve their gods, possibly to follow their abominable practices. Inserted at this point is the striking exhortation, "Everything that I command you you shall be careful to do; you shall not add to it or take from it."

Chapter 13 warns the people to beware of the temptations to apostasy arising from the urging or example of prophet-diviners, kinfolk or friends, or a whole town; they are to kill the tempters and destroy the towns. Chapter 14 is devoted mainly to a list of living things that may or may not be eaten, the "clean" and "unclean," similar to the list in Leviticus, chapter 11; and to laws for tithes and first fruits to be brought annually to the central sanctuary and triennially to the Levites in the towns, who are specified as having no "portion" of their own (two years to the centre, the third year to the town Levites). Chapter 15 deals mainly with the releases to be granted every seventh year to debtors of their debts and Hebrew slaves of their bondage; lenders are exhorted and commanded not to refuse loans to the poor in the sabbatical year of release, and God's redemption of Israel from Egypt is given as the reason for freeing one's Hebrew slaves in the sabbatical release. The first section of chapter 16, verses 1–17, gives the rules for celebrating the three main festivals of the religious year: Unleavened Bread, Weeks, and Booths, which are to be observed at the central sanctuary (hence later called the three pilgrim festivals).

Beginning with verse 18 of chapter 16 there is a discussion of the appointment and character of judges, and of judicial procedures and punishments for apostasy, homicide, and other crimes; similarly, beginning with verse 14 of chapter 17 there are rules on the selection of a king and for his conduct, and the injunction that he read from "a copy of this law," so that he may be edified and chastened. The first portion of chapter 18 deals with the office and support of priests, referred to here as "the Levitical priests . . . all the tribe of Levi," not distinguishing the Aaronic priests from the lesser Levites. This is followed—after a passage inveighing against abominable cultic and divinatory practices of the nations of the land—by a promise that God will raise up prophets among the people and instructions on how to tell true from false prophets. Thus the offices of judge, king, priest, and prophet are considered in chapters 16–18.

Chapter 19 deals again with crime and punishment. It distinguishes between unintentional manslaughter and murder, setting up cities of refuge for the manslayer and ordering the murderer to be killed by the blood avengers. It also lays down the rules for witnesses and the punishment for perjury. It closes with the famous lex talionis: "Life for life, eye for eye, tooth for tooth, hand for hand, foot for foot," which in context may spell out what is to happen to the false witness and even could be interpreted as a moderating, rather than an inhumane, precept (no more than an eye for an eye, etc.). Chapter 20 gives the rules for holy war, listing the situations that exempt men from military service (*e.g.*, a newly married man) and distinguishing the treatment of non-Canaanite and Canaanite cities; the latter are to be utterly destroyed, yet it is forbidden to destroy fruit-bearing trees. There are also rules on holy war in 21:10–14; 23:9–14; 24:5; and 25:17–19. Chapters 20–25 contain a great variety of laws; the just treatment of women captives, sexual offenses, exclusions from the religious community, public hygiene in campgrounds, and many other things.

The lex talionis

The last of the laws are set forth in chapter 26, dealing with the first fruits offering and tithes. At the annual offering (or soon after entering Canaan), in the central sanctuary, the worshipper is to recite a piece beginning, "A wandering Aramaean was my father," affirming his link with the patriarchs and extolling God's wondrous deeds on behalf of Israel. And every third year he is to set aside his tithe "to the Levite, the sojourner, the fatherless, and the widow" and make an affirmation "before the Lord" that he has complied and avoided any ritual stain.

The final passage in chapter 26 proclaims that "this day" God has proclaimed his law, Israel has affirmed its commitment to God and his law, and God has affirmed his choice of Israel as his special, holy people, to be set up high above all the nations. This is the hortatory conclusion to chapters 12–26 and to the "second law," or Covenant, contained therein.

Covenant ratification and renewal ceremonies

The emphasis on the laws given on "this day" is continued in the supplementary chapters 27–28, which deal with Covenant ratification and renewal ceremonies, apparently a reference to an original ceremony in Moab, one in Canaan on the first day in the land, and subsequent, possibly annual, renewal ceremonies. Blessings and curses are to be pronounced from Mts. Gerizim and Ebal for respectively fulfilling or disobeying the Covenant: all good things or all bad things will befall the people, as they keep or fail to keep the Covenant. Some of the curse consequences in chapter 28, referring to siege, subjugation, and exile, are believed by some scholars to reflect late pre-exilic or exilic situations. The curse consequences fill up the bulk of these chapters and are recounted in powerful, moving language, ending with a threat to return the people to Egypt.

Concluding exhortation and traditions about the last days of Moses. Chapters 29–31 comprise the third and last address of Moses to the people of Israel. They are preceded by an introductory verse referring to "these words" as a covenant made in Moab, in addition to the one made at Horeb (Sinai). After reminding them of all that God has done for them, Moses calls on the whole people to enter into the sworn Covenant made this day that they may be his people and he may be their God, warning the secret apostate of the calamities that will befall him. Yet the possibility of a return to God and the land is held out to

those who will suffer exile and persecution as punishment for their apostasy, again presumably a reflection of the exilic situation (chapter 30 verses 1–10 seems clearly to be an interpolation inspired by the actual experience of exile). This law, it is emphasized, is no recondite, remote thing up in the sky but is, rather, very close to men, "in your mouth and in your heart"; what is revealed is made plain, it is not the secret things of God. Moses sets before them the classic Deuteronomic choice: "life and good" over "death and evil." The people are given that choice and told the consequences of loving the Lord and keeping the Covenant or of going the other way.

The last words and acts of Moses

The final chapters are concerned with the last words and acts of Moses: directing Joshua to lead Israel after his death, writing down "this law," calling for a sabbatical renewal ceremony of it on the Feast of Booths, ordering that it be put beside the ark of the Covenant, and uttering two poems. The first, "The Song of Moses" (chapter 32), praises the faithfulness and power of the Lord, decries the faithlessness and wickedness of Israel, and predicts the consequent divine punishment; it adds, however, that in the end the Lord will relent and will vindicate his people. The second poem, "The Blessing of Moses" (chapter 33), blesses each of the tribes of Israel, one by one, and the blessings are associated with God's love, the law commanded by Moses, and the kingship of God over his people. There are indications in both poems of a considerably later date (after Joshua's time, perhaps in the period of the Judges); Moses is spoken of in the third person in "The Blessing" poem.

The narrative of Deuteronomy, and thus of the Pentateuch, ends with Moses' ascent to the top of Mt. Pisgah, his being shown the Promised Land by God, and his death there in the land of Moab, buried by God in an unknown grave. It is emphasized in the closing words that Moses was a unique prophet "whom the Lord knew face to face" and through whom the Lord wrought unique "signs and wonders" and "great and terrible deeds." Thus end the Five Books of Moses. (S.C.)

THE NEVI'IM (THE PROPHETS)

The canon of the Prophets. The Hebrew canon of the section of the Old Testament known as the Nevi'im, or the Prophets, is divided into two sections: the Former Prophets and the Latter Prophets. The Former Prophets contains four historical books—Joshua, Judges, Samuel, and Kings; the Latter Prophets includes four prophetic works—the books of Isaiah, Jeremiah, Ezekiel, and the Twelve (Minor) Prophets. The Twelve Prophets, formerly written on a single scroll, include the books of Hosea, Joel, Amos, Obadiah, Jonah, Micah, Nahum, Habakkuk, Zephaniah, Haggai, Zechariah, and Malachi. Thus, in the Hebrew canon of the Prophets there are, in effect, eight books.

Variations in the canon of the Prophets

The Christian canon of the Prophets does not include the Former Prophets section in its division of the Prophets; instead, it calls the books in this section Historical Books. In addition to Isaiah, Jeremiah, and Ezekiel, the Christian canon of the Prophets includes two works from the division of the Hebrew canon known as the Ketuvim (the Writings): the Lamentations of Jeremiah and the Book of Daniel. The Twelve (Minor) Prophets are separated into individual books. The number of works in the Christian canon, however, varies. The Protestant canon contains all the books of the Latter Prophets and the two books from the Ketuvim, thus listing 17 works among the prophetic writings. The Roman Catholic canon accepts one other book as a canonical prophetic work, namely, Baruch (including the Letter of Jeremiah); the number of prophetic writings in the Roman Catholic canon is, therefore, 18. The Greek Orthodox Synod of Jerusalem in 1672 did not accept Baruch as canonical.

As far as the Former Prophets is concerned, the Protestant canon, following the Septuagint, separates Samuel and Kings into two sections each: I and II Samuel, and I and II Kings. The Roman Catholic and Orthodox churches in the past divided these two works into I, II, III, and IV Kings, but most Roman Catholic translations now follow the listing as it is in the Septuagint.

Hebrew prophecy. Hebrew prophecy was rooted in the prophetic activities of various individuals and groups from the nations and peoples of the ancient Near East. Though prophecy among ancient Egyptians, Mesopotamians, and Canaanites—as well as among the peoples of the Aegean civilization—generally was connected with "foretelling" (or predicting) the future, the Hebrew view of prophecy centred on "forthtelling" (or proclaiming), though it included predictive aspects. Thus, in Hebrew prophecy the phrase "Thus says the Lord" is repeated constantly to emphasize the "forthtelling" motif. The Hebrew prophets were very conscious of the absolute holiness (separateness) of God and the purpose of God for his chosen people, Israel. Because of this consciousness, they developed an acute awareness of sin and its effects on man and society and, in consequence of such an awareness, a radical ethical outlook that applied to both the individual and the community.

Emphasis on "forthtelling"

The Hebrew term for prophet (*navi'*) is probably related etymologically to the Akkadian verb *nabū*, meaning "to call" or "to name." The Hebrew prophet may thus be viewed as a "caller," or spokesman, for God. Other designations for prophet in the Old Testament are *ro'e*, or "seer," and *ḥoze*, or "visionary," the two latter terms indicating that the predictive element was operative in Hebrew prophecy. The distinctive element of Hebrew prophecy, however, was the relationship of the prophet to God, the Lord of the Covenant, and to Israel, the covenant people. He spoke for the sovereign Lord to remind, cajole, castigate, reprove, comfort, and give hope to the people of the covenant, constantly reminding them that they were chosen to witness to the nations of the love, mercy, and goodness of God.

Some of the Hebrew prophets, from the 11th to the 8th century BCE, belonged to bands or guilds of ecstatic prophets. Such prophets were spokesmen for God whose uncontrollable actions and words caused them to be feared and, sometimes, held in contempt. In II Kings, chapter 9, verse 11, a prophet—who came to Jehu, the 9th-century-BCE army commander who became king of Israel, in order to anoint him—was called a "madman" (*meshugga'*). Other Hebrew prophets were more independent, such as Nathan and Elijah, though they continued to maintain the quality of being uncontrollable—at least as far as the political authorities were concerned. Both of these early nonwriting prophets spoke out against the oppression of the weak by the strong, a theme that came to be expressed constantly in Judaism throughout the centuries. The activities of such early prophets, including also Micaiah and Elisha in the 9th century BCE, are described in the Former Prophets.

In the 8th century BCE, the writing prophets—*i.e.*, the Latter Prophets—began their activities. Though all the books that bear their names probably have been edited by schools of a prophet or by individuals or groups that were influenced by their ideas, the editors or disciples of the prophets preserved as well as was possible the words, activities, and idiosyncratic themes of the prophetic personalities. Some of the Latter Prophets may have been connected with the priestly class, such as Isaiah, Jeremiah, and Ezekiel; most of the Latter Prophets, however, were independent of priestly connections. All of the Latter Prophets stood out in contrast to the court prophets who, in the tradition of court prophets of most ancient Near Eastern peoples, seldom contradicted what they believed was expected of them by their sovereigns or the people.

Joshua. The Book of Joshua takes its name from the man who succeeded Moses as the leader of the Hebrew tribes—Joshua, the son of Nun, a member of the tribe of Ephraim. In post-biblical times Joshua himself was credited with being the author of the book, though internal evidence gives no such indication. According to the views of the German biblical scholar Martin Noth, which have been accepted by many contemporary biblical critics, the Book of Joshua was the second of a series of five books (Deuteronomy, Joshua, Judges, Samuel, and Kings) written by a Judaean oriented historian after the fall of Jerusalem in 586 BCE. This writer (called the Deuteronomist and designated D) constructed the history

The work of the Deuteronomic historian

of Israel from the death of Moses to the beginning of the Babylonian Exile (586–538 BCE). The Deuteronomist, according to this view, used sources, both oral and written, from various periods to produce the history of Israel in these five books. The Book of Joshua probably contains elements from the J and E documents, as well as local and tribal traditions, all of which were modified by additions and editing until the book assumed its present form. The main theme of the Deuteronomist historian was that under the guidance of and in obedience to Yahweh, Israel would persevere and conquer its many enemies.

This theme is especially and dramatically presented in Joshua. Under the guidance of Yahweh, the people of Israel entered and conquered Canaan in fulfillment of the promise of God to Abraham and his descendants in Genesis, chapter 12. Joshua is interpreted as a second Moses—*e.g.,* he sent out spies, led the people in crossing the Jordan River on dry land as Moses had crossed the Sea of Reeds, and ordered the males to be circumcised with flint knives as Zipporah, Moses' wife, had earlier circumcised the son of Moses (and probably Moses himself). He was obedient to the will of Yahweh, and because of this obedience he was able to lead the Israelite tribes in their battles against the Canaanites. As long as they were faithful to their covenant promise, the land would be theirs as a trust.

The book may be divided into three parts: the story of the conquest of Canaan (chapters 1–12); the division of the land among the tribes of Israel (chapters 13–22); and Joshua's farewell address, the renewal of the Covenant, and Joshua's death (chapters 23–24).

The conquest of Canaan. As told by the Deuteronomist, the conquest of Canaan by Joshua and the Israelite tribes was swift and decisive. No conquest of central Canaan (in the region of Shechem), however, is mentioned in the book; and some scholars interpret this to mean that the central hill country was already occupied either by ancestors of the later Israelite tribes prior to the time of Moses or by portions of Hebrew tribes that had not gone to Egypt. Because these people made peace with the tribes under Joshua, a conquest of the area apparently was not necessary. Archaeological evidence supports portions of Joshua in describing some of the cities (*e.g.,* Iachish, Debir, and Hazor) as destroyed or conquered in the late 13th century BCE, the approximate time of the circumstances documented in Joshua. Some of the cities so reported, however, apparently were devastated at some time prior to or later than the 13th century. Jericho, for example, was razed at the end of the Middle Bronze Age (*c.* 1550 BCE) and most likely had not been rebuilt as a strongly fortified town by the time of Joshua, though the site may well have been inhabited during this period. The city of Ai was destroyed about 600 years before; but it may have been a garrison site for the city of Bethel, which was destroyed later by the "house of Joseph." Though many of the cities of Canaan were conquered by the Israelites under Joshua, historical and archaeological evidence indicates that the process of conquering the land was lengthy and not completed until David conquered the Jebusite stronghold of Jerusalem in the early 10th century BCE. At any rate, the 13th century was an ideal time for a conquest of the area because of the international turmoil involving the great powers of the time: Egypt and Babylonia. A political vacuum existed in the area, permitting small powers to strengthen or to expand their holdings.

Archaeological evidence supporting the biblical narrative

The introductory section of Joshua (chapters 1 and 2), in dealing with the Deuteronomist's view of the ideal man of faith—one who is full of courage and faithful to the law that was given to Moses—relates the story of spies sent to Jericho, where they were sheltered by Rahab, a harlot, whose house was spared by the Israelites when they later destroyed the city. In the Gospel According to Matthew, in the New Testament, Rahab is listed as the grandmother of Jesse, the father of David (the architect of the Israelite empire), which may be the reason why this story was included in Joshua. Also in the New Testament, in the Letter to the Hebrews, Rahab is depicted as an example of a person of faith. After the return of the spies, who reported that the people of Canaan were "fainthearted" in the face of the Israelite threat, Joshua launched the invasion of Canaan; the Israelite tribes crossed the Jordan River and encamped at Gilgal, where the males were circumcised after a pile of stones had been erected to commemorate the crossing of the river. They then attacked Jericho and, after the priests marched around it for seven days, utterly destroyed it in a *ḥerem; i.e.,* a holy war in which everything is devoted to destruction. Prior to the Israelites' further conquests it was discovered that Achan, a member of the tribe of Judah, had broken the *ḥerem* by not devoting everything taken from Jericho to Yahweh. Because he had thus sinned in keeping some of the booty, Achan, his family, and all of his household goods were destroyed and a mound of stones was heaped upon them. The Israelite tribes next conquered Ai, made agreements with the people of the region of Gibeon, and then campaigned against cities to the south, capturing several of them, such as Lachish and Debir, but not Jerusalem or the cities of Philistia on the seacoast. Joshua moved north, first conquering the city of Hazor—a city of political importance—and then defeating a large number (31) of the kings of Canaan, though the conquests of their cities did not necessarily follow.

The significance of the *ḥerem*

Division of the land and renewal of the Covenant. The division of the land among the tribes is recounted in chapters 13–22. Two sources were apparently used by the Deuteronomist in dealing with the division of the land: a boundary list from the pre-monarchical period (*i.e.,* before the late 11th century BCE) and a list of cities occupied by several tribes from the 10th to the 7th century BCE. The tribes who occupied territories were: Reuben, Gad, Manasseh, Caleb, Judah, the Joseph tribes (Ephraim and Manasseh), Benjamin, Simeon, Zebulun, Issachar, Asher, Naphtali, and Dan. Certain cities (*e.g.,* Hebron, Shechem, and Ramoth) were designated Levitical cities. Though the Levites probably did not control the cities politically, as the priestly class they were of cultic significance—and therefore feared and respected—in cities that were the sites of sanctuaries.

As Moses had before him, Joshua gave a farewell address (chapter 23) to his people, admonishing them to be loyal to the Lord of the Covenant; and in the closing chapter (24), the Israelites reaffirmed their loyalty to Yahweh at Shechem: first having heard the story of God's salvatory deeds in the past, they were asked to swear allegiance to Yahweh and to repudiate all other gods, after which they participated in the Covenant renewal ceremony. After the people were dismissed, Joshua died and was buried in the hill country of Ephraim; the embalmed body of Joseph that had been carried with the Hebrews when they left Egypt more than a generation earlier was buried on purchased land; and Eleazar, the priestly successor to Aaron (Moses' brother), was buried at Gibeah.

The Covenant renewal ceremony

Besides the obvious emphases on the conquest of Canaan and the division of the land, the Deuteronomist gave special attention to the ceremony of Covenant reaffirmation. By means of a regularly repeated Covenant renewal the Israelites were able to eschew Canaanite religious beliefs and practices that had been absorbed or added to the religion of the Lord of the Covenant, especially the fertility motifs that were quite attractive to the Hebrew tribes as they settled down to pursue agriculture, after more than a generation of the nomadic way of life.

Judges. The Book of Judges, the third of the series of five books that reflect the theological viewpoint of the Deuteronomic historian, covers the history of the Israelite tribes from the death of Joshua to the rise of the monarchy, a period comprising nearly 200 years (*c.* 1200–*c.* 1020 BCE). Though the internal chronology of Judges points to a period of about 400 years, the editor may have arbitrarily used the formula of 40 years for a generation of rule by a judge; and he may have compiled the list in the form of a series of successive leaders who actually may have led only a particular tribe or a group of tribes during the same generation as another judge. In other words, the reign of two or more judges may well have overlapped.

The Deuteronomic "theology of history." The Deuteronomic "theology of history" shows through very clearly in Judges: unless the people of the Covenant remain faithful and obedient to Yahweh, they will suffer the due consequences of disobedience, whether it be an overtly willful

act or an unthinking negligence in keeping the Covenant promise. The Deuteronomist worked out a formula for his theology of history that was based in a very dramatic way on the historical events of the period: (1) obedience to Yahweh brings peace and well-being; (2) a period of well-being often involves a slackening of resolve to keep the commandments of Yahweh or outright disobedience; (3) disobedience leads to a weakness of the faith that had bound the community together and thus leaves the community open to repression and attacks from external enemies; and (4) external repression forces the community to reassess its position and ask the cause of the calamities, thus leading to repentance and eventual strength to resist all enemies.

Canaanite culture and religion. The Israelite tribes during the period of the guidance and leadership of Moses and Joshua mainly had to contend with nomadic tribes; in their contacts with such groups they absorbed some of the attitudes and motifs of the nomadic way of life, such as independence, a love of freedom to move about, and fear of or disdain for the way of life of settled, agricultural, and urban peoples.

Canaanite cultural accomplishments and religious beliefs and practices

The Canaanites, with whom the Israelites came into contact during the conquest by Joshua and the period of the Judges, were a sophisticated agricultural and urban people. The name Canaan means Land of Purple (a purple dye was extracted from a murex shellfish found near the shores of Palestine). The Canaanites, a people who absorbed and assimilated the features of many cultures of the ancient Near East for at least 500 years before the Israelites entered their area of control, were the people who, as far as is known, invented the form of writing that became the alphabet, which, through the Greeks and Romans, was passed on to many cultures influenced by their successors—namely, the nations and peoples of Western civilization.

The religion of the Canaanites was an agricultural religion, with pronounced fertility motifs. Their main gods were called the Baalim (Lords) and their consorts, the Baalot (Ladies), or Asherah (singular), usually known by the personal plural name Ashtoret. The god of the city of Shechem, which city the Israelites had absorbed peacefully under Joshua, was called Baal-berith (Lord of the Covenant) or El-berith (God of the Covenant). Shechem became the first cultic centre of the religious tribal confederacy (called an amphictyony by the Greeks) of the Israelites during the period of the judges. When Shechem was excavated in the early 1960s, the temple of Baal-berith was partially reconstructed; the sacred pillar (generally a phallic symbol or, often, a representation of the *ashera,* the female fertility symbol) was placed in its original position before the entrance of the temple.

The Baalim and the Baalot, gods and goddesses of the Earth, were believed to be the revitalizers of the forces of nature upon which agriculture depended. The revitalization process involved a sacred marriage (*hieros gamos*), replete with sexual symbolic and actual activities between men, representing the Baalim, and the sacred temple prostitutes (*qedeshot*), representing the Baalot. Cultic ceremonies involving sexual acts between male members of the agricultural communities and sacred prostitutes dedicated to the Baalim were focussed on the Canaanite concept of sympathetic magic. As the Baalim (through the actions of selected men) both symbolically and actually impregnated the sacred prostitutes in order to reproduce in kind, so also, it was believed, the Baalim (as gods of the weather and the Earth) would send the rains (often identified with semen) to the Earth so that it might yield abundant harvests of grains and fruits. Canaanite myths incorporating such fertility myths are represented in the mythological texts of the ancient city of Ugarit (modern Ras Shamra) in northern Syria; though the high god El and his consort are important as the first pair of the pantheon, Baal and his sexually passionate sister-consort are significant in the creation of the world and the renewal of nature.

Attraction of the Israelites toward Canaanite religion

The religion of the Canaanite agriculturalists proved to be a strong attraction to the less sophisticated and nomadic-oriented Israelite tribes. Many Israelites succumbed to the allurements of the fertility-laden rituals and practices of the Canaanite religion, partly because it was new and different from the Yahwistic religion and, possibly, because of a tendency of a rigorous faith and ethic to weaken under the influence of sexual attractions. As the Canaanites and the Israelites began to live in closer contact with each other, the faith of Israel tended to absorb some of the concepts and practices of the Canaanite religion. Some Israelites began to name their children after the Baalim; even one of the judges, Gideon, was also known by the name Jerubbaal ("let Baal contend").

As the syncretistic tendencies became further entrenched in the Israelite faith, the people began to lose the concept of their exclusiveness and their mission to be a witness to the nations, thus becoming weakened in resolve internally and liable to the oppression of other peoples.

The role of the judges. Under these conditions, the successors to Joshua—the judges—arose. The Hebrew term *shofet,* which is translated into English as "judge," is closer in meaning to "ruler," a kind of military leader or deliverer from potential or actual defeat. In a passage from the so-called Ras Shamra tablets (discovered in 1929) the concept of the judge as a ruler is well illustrated:

> Our king is Triumphant Baal,
> Our judge, above whom there is no one!

The magistrates of the Phoenician-Canaanite city of Carthage, which competed with Rome for supremacy of the Mediterranean world in the 3rd century BCE, were called *suffetes,* thus pointing toward the political authority of the judges.

The office of judgeship in the tribal confederacy of the Israelites, which was centred at a covenant shrine, was not hereditary. The judges arose as Yahweh saw fit, in order to lead an erring and repentant people to a restoration of a right relationship with him and to victory over their enemies. The quality that enabled a person selected by Yahweh to be a judge was charisma, a spiritual power that enabled the judge to influence, lead, and control the people caught between the allurements of the sophisticated Canaanite culture and the memory of the nomadic way of life with its rugged freedom and disdain for "civilization." Though many such leaders are mentioned, the Book of Judges focusses attention upon only a few that are singled out as especially significant: Deborah and Barak, Gideon, Abimelech, Jephthah, and Samson. In spite of the Israelites' repeated apostasy, such leaders, under the guidance and spiritual powers granted to them by Yahweh, were able to lead their tribes in successfully defeating or driving back their opponents.

The significance of the office of the judge

The Book of Joshua may be divided into four parts: (1) the conquests of several tribes (chapter 1); (2) a general background for the subsequent events according to the interpretation of the Deuteronomic historian—"And the people of Israel did what was evil in the sight of the Lord and served the Baals"—(chapter 2 through chapter 3, verse 6); (3) the exploits of the judges of Israel (chapter 3, verse 7, through chapter 16); and an appendix (chapters 17 through 21).

Judges, chapter 1, shows that the conquest of Canaan, in contradistinction to the view presented in Joshua, was incomplete, inconclusive, and lengthy. Though conquests of some of the tribes (Judah, Simeon, Caleb, and the "house of Joseph") are noted, the main emphasis is on the cities and areas that the tribes had *not* conquered—*e.g.,* "And Ephraim did not drive out the Canaanites who dwelt in Gezer, but the Canaanites dwelt in Gezer among them" (chapter 1, verse 29).

The second section gives the Deuteronomic interpretation of the consequences of such a policy:

> they forsook the Lord, the God of their fathers, who had brought them out of the land of Egypt; they went after other gods, from among the gods of the peoples who were round about them; and they provoked the Lord to anger. They forsook the Lord, and served the Baals and the Ashtaroth. (chapter 2, verses 12–13)

The purpose of allowing the Canaanites to continue to exist

In chapter 3, an explanation is given as to why the Canaanites had not been annihilated and were allowed to remain with the Israelites: they enabled the Israelites to be tested in the techniques of warfare; the Philistines, for

example, had a monopoly on the smelting of iron in the area—and the iron used in their weapons was far superior to the bronze used by the Israelites for their swords, shields, and armaments—until the secret had been wrested from them by the first king of Israel, Saul, in the latter part of the 11th century BCE. The Canaanites also served to test the faith of the Israelites in the one, true God, Yahweh.

The role of certain lesser judges. The third section relates the exploits of the various judges. Othniel, a member of the tribe of Caleb, delivered the erring Israelites from eight years of oppression by Cushan-rishathaim, king of Mesopotamia. The king, however, was most likely an area ruler, rather than a king of the Mesopotamian Empire. Another judge, Ehud, a left-handed Benjamite, delivered Israel from the oppression of the Moabites. Ehud, a fat man who had hidden a sword under his garments on his right side so that when a search of his person was made it would be overlooked, brought tribute to Eglon, the Moabite king. Upon Ehud's claiming to have a secret message for the king, Eglon dismissed the other people carrying tribute. Ehud then said to the King, "I have a message from God to you," assassinated him, locked the doors to the chamber, and escaped. Rallying the Israelites around him, Ehud led an attack upon the Moabites that was decisive in favour of the Israelites. Shamgar, the third judge, is merely noted as a deliverer who killed 600 Philistines.

The roles of Deborah, Gideon, and Jephthah. The first notably important judge of the tribal confederacy was Deborah, who was primarily a seer, poet, and interpreter of dreams but still a person endowed with the kind of charisma that identified her as a judge sent from Yahweh. The story of the victory of the Israelites under the charismatic leadership of Deborah and the military leadership of Barak, her commander, is related in prose (chapter 4) and repeated in poetry (chapter 5, which is known as the "Song of Deborah"). The Canaanites, under the leadership of Jabin, king of a reestablished Hazor, and his general Sisera, had oppressed an apostate Israel. Deborah sent word to all the tribes to unite against the Canaanites, but only about half the tribes responded. The Canaanites had asserted control over the Valley of Jezreel, which was an important commercial thoroughfare and was commanded by the city of Megiddo. In this valley dominated by the hill of Megiddo (Armageddon)—a site of many later crucial military battles and which later became the symbolic name for the final battle between the forces of good and the forces of evil in apocalyptic literature—the Israelites met the Canaanites near the river Kishon in open battle. A cloudburst occurred, causing the river to flood, thus limiting the manoeuvrability of the Canaanite chariots. The Canaanite general Sisera, seeing defeat for his forces, fled, seeking refuge in the tent of a Kenite woman, Jael. A supporter of the cause of Israel, Jael gave Sisera a drink of milk (fermented?) and he fell asleep "from weariness." Jael pounded a tent peg through his temple, thus ending decisively the threat of the Canaanites of Hazor. The victory song of Deborah in chapter 5 is one of the oldest literary sections of the Old Testament. It is a hymn that incorporates the literary forms of a confession of faith, a praise of Yahweh's theophany (manifestation), an epic, a curse, a blessing, and a hymn of victory.

The significance of Deborah and Gideon

Another important judge, perhaps the most important other than Samuel, was Gideon, whose exploits are related in chapters 6–8. The oppressors of Israel during the time of Gideon were the camel-borne raiders from Midian, roving bands that pillaged the farms and unfortified villages for seven years. A prophet appeared among the Israelites and denounced them for their apostasy, after which, according to the account, an angel of Yahweh visited and then commissioned Gideon, a member of the tribe of Manasseh, to lead the Israelites against the enemies from the Transjordan. After sacrificing to Yahweh, building an altar to the Lord (which he named Yahweh Shalom, or "Yahweh is peace"), and destroying an altar of Baal and an *ashera* (most likely a wooden pole symbolizing the goddess) beside it, he sent out messengers to gather together the tribes in order to meet an armed force of the Midianites and Amalekites that had crossed the Jordan River and were encamped in the Valley of Jezreel. He went to a threshing floor (a common place to seek divinatory advice) and sought a sign from Yahweh—dew on a fleece of wool placed overnight on the threshing floor, with the rest of the area remaining dry. After receiving the positive divinatory sign, Gideon assembled a large force, reduced it to 300 men, and infiltrated the outposts of the Midianite camp with his servant—overhearing a Midianite telling another of his dream about a barley cake rolling into the camp of the Midianites and striking a tent so that it fell down and was flattened (which Gideon interpreted as a sign of victory for the forces under him). He encircled the camp of the Midianites about midnight. On signal, the men broke jars, shouted, waved torches, blew rams' horns, and attacked the encampment. The Midianites, in the confusion, were routed and harassed in their flight. In their pursuit of the fleeing Midianites, Gideon and his forces were refused aid by the cities of Succoth and Penuel, which was a violation of the tribal confederacy agreements. The Midianites, however, were again the objects of a surprise attack and their two kings (Zebah and Zalmunna) were captured and later executed by Gideon because they had killed his brother. The leaders of Succoth were punished and the men of Penuel were killed in retaliation for their refusal to aid the forces of Gideon.

After the victory, the people, recognizing their need for centralized leadership of the confederacy, petitioned to Gideon that he establish a hereditary monarchy, with himself as the first king. Gideon refused, however, on the basis that "the Lord will rule over you."

The request for a hereditary monarchy

After Gideon died, the people returned to worshipping the gods of the Canaanites, especially Baal-berith. Abimelech, one of the 70 sons of the wives and concubines of Gideon, went to Shechem to solicit support for his attempt to establish a monarchy. After receiving financial support from those who controlled the treasury of the shrine of Baal-berith, he hired a band of assassins—who killed all of his brothers except Jotham, the youngest of Gideon's sons. Abimelech was declared king by the Shechemites. The surviving Jotham told a parable about trees that sought a king—after all the larger trees refused the kingship, the bramblebush, which was highly inflammable, accepted the offer. The point of the parable was that as the bramblebush is highly inflammable, so also would the reign of Abimelech be the source of fires of rebellion and revolution. Revolution did occur, and after being wounded at Thebez by a millstone dropped by a woman from a tower, Abimelech asked his armour bearer to kill him. The attempt of Abimelech and the Shechemites to establish a monarchy thus proved to be abortive and premature.

After a brief account of the rule of two judges, Tola of the tribe of Issachar and Jair from Gilead, the Deuteronomist describes the apostasy of the Israelites and the consequent oppression of the tribes by the Philistines from the seacoast and the Ammonites from the Transjordan. The Israelites looked for a leader and found Jephthah, the son of a harlot, who had been rejected by the sons of his father and who had gathered about him a band who made their living by raiding others. Jephthah made several attempts to negotiate with the Ammonites and Moabites; when the Ammonites did not cooperate, Jephthah moved against them. Seized by the Spirit of the Lord—*i.e.*, ecstatically inspired—he began his campaign with a vow to sacrifice the first person he saw upon his return home as a burnt offering to Yahweh. He was victorious over the Ammonites, but the first person he saw on return home was his only child, a daughter. Upon learning of her destined fate, she requested a two-month period to be with her friends to bewail her virginity and approaching death. The story is reminiscent of the fertility myths of the ancient Near East. After she was sacrificed, Jephthah subdued a contingent of the Ephraimites in the Transjordan to bring peace to the area. A password was used to separate the Ephraimites from the men under Jephthah: "shibboleth." Because the Ephraimites could not pronounce the word correctly, in that their dialect was different from the others, they were thus identified and killed.

The judgeship of Jephthah

In chapter 12, three judges are given cursory treatment: Izban of Bethlehem, Elon the Zebulunite, and Abdon the Ephraimite.

The exploits of Samson

The role of Samson. The exploits of the great Israelite strongman judge, Samson (a member of the tribe of Dan), are related in chapters 13–16. Dedicated from birth by his mother to Yahweh, Samson became a member of the Nazirites, an anti-Canaanite reform movement. As a Nazirite, he was required never to cut his hair, drink wine, or eat ritually unclean food. He married a Philistine woman whom he then left when she helped her fellow Philistines avoid payment to Samson in a riddle contest by giving them the answer. Returning later to find her given to another man, he burned the grainfields of the Philistines. They sought revenge by killing Samson's wife and her father. The exploits of Samson against the Philistines from then on are numerous. After he met the temptress Delilah, who wrested from him the secret of his great strength (*i.e.,* his long uncut hair because of his vow), Samson was captured by the Philistines after his hair had been cut short. After imprisonment, blinding, and humiliation, Samson finally avenged his loss of self-respect by pulling down the main pillars of the temple of the Philistine god Dagon, after which the temple was destroyed, along with numerous Philistines. Though Samson was more a folk hero than a judge, he was probably included in the list of judges because his ventures against the Philistines slowed their movements inland against the Israelite towns and villages. The Philistines were a group of "sea peoples" united in a confederacy of five city-states: Gaza, Ashkelon, Ashdod, Gath, and Ekron. To the area they gave their name, which has endured to the 20th century: Palestine.

The final section of the Book of Judges is an appendix divided into two parts: (1) the story of Micah, the repentant Ephraimite, a Levite priest who deserted him to be priest of the tribe of Dan, and the establishment of a shrine at the conquered city of Laish (renamed Dan) with the cult object taken from the house of Micah; and (2) the story of the Benjamites who were defeated in a holy war after they had killed a concubine of a Levite. The book ends with a critique of the period: "In those days there was no king in Israel; every man did what was right in his own eyes" (chapter 21, verse 25).

Samuel. The book of Samuel covers the period from Samuel, the last of the judges, through the reigns of the first two kings of Israel, Saul and David (except for David's death). The division of Samuel and its succeeding book, Kings (Melakhim), into four separate books first appeared in the Septuagint, the Greek translation of the Old Testament from the 3rd to 2nd centuries BCE.

The two main sources of Samuel

Theological and political biases. Containing two primary sources, the book of Samuel is the result of the editorial skill of the Deuteronomic historians of the post-exilic period. The early source, which is pro-monarchical and may have been written by a single author, is found in I Samuel, chapter 9, verse 1, through chapter 10, verse 16, as well as chapter 11 and most of II Samuel. The chapters just noted were probably written by a chronicler during the reign of Solomon; possible authors of these chapters were Abiathar, a priest of the line of Eli (who was Samuel's predecessor at the shrine of Shiloh), or Ahimaaz, a son of Zadok (who originally may have been a priest of the Jebusite city of Jerusalem that David made his capital). The chapters in I Samuel are sometimes called the "Saul" source because it is in them that Saul's charismatic leadership is legitimized in the form of kingship. The chapters of II Samuel, also displaying a pro-monarchical bias—as far as content is concerned—are the "book of David." In the early source, Samuel, a seer, prophetic figure, and priest of the shrine at Shiloh, is viewed mainly as the religious leader who anointed Saul to be king. The later source, which displays a somewhat anti-monarchical bias and shows the marks of disillusionment on the part of the Deuteronomic historians of the post-exilic period, is found in I Samuel, chapter 7, verse 3, to chapter 8, verse 22, chapter 10, verses 17–27, and chapter 12. Sometimes called the Samuel source, the later source interprets the role of Samuel differently; he is viewed as the last and most important judge of the whole nation, whose influence extended to the shrines at Bethel, Gilgal, and Mizpah. The two sources illustrate the two opposing tendencies that lasted for centuries after the conquest of Canaan.

The international and area situation during the 11th–10th centuries BCE

During the period of Samuel, Saul, and David (the 11th–10th century BCE), the Israelites were still threatened by various local enemies. The great nations—Egypt, Assyria, and the Hittite Empire—were either involved in domestic crises or concerned with areas other than Palestine in their expansionist policies. Of the various peoples pressing to break up the Israelite confederacy, the Philistines (the "sea peoples") of the Mediterranean coast proved to be the most dangerous. Expanding eastward with their iron-weapon equipped armies, the Philistines threatened the commercial routes running north and south through Israelite territory. If they captured and controlled such areas as the Valley of Jezreel, they would eventually strangle the economic life of the Israelite confederacy.

To meet this threat, the tribal confederacy had four options open to it. First, the tribes could continue as before, loosely held together by charismatic leaders, who served only as temporary leaders. Second, they could create a hereditary hierocracy (rule by priests), which the priest of the shrine at Shiloh, Eli, apparently attempted to inaugurate. A third possible course of action was to establish a hereditary judgeship, which was the aspiration of the judge Samuel. But in either of these two possibilities, the sons of Eli and Samuel were not of the same stature as their fathers; and the apparent hopes of their fathers could not be realized. The fourth alternative was a hereditary monarchy. The book of Samuel is an account of the eventual success of those who supported the monarchical position, along with the Deuteronomic interpretation that pointed out the weaknesses of the monarchy whenever it departed from the concept of Israel as a covenant people and became merely one kingdom among other similar kingdoms.

The book of Samuel may be divided into four sections: (1) the stories of Samuel, the fall of the family of Eli, and the rise of Saul (I Samuel, chapters 1–15); (2) the accounts of the fall of the family of Saul and the rise of David (I Samuel, chapter 16, to II Samuel, chapter 5); (3) the chronicles of David's monarchy (II Samuel, chapter 6, to chapter 20, verse 22); and (4) an appendix of miscellaneous materials containing a copy of Psalm 18, the "last words of David," which is a psalm of praise, a list of heroes and their exploits, an account of David's census, and other miscellaneous materials.

The early life and "call" of Samuel

The role of Samuel. The first section (chapters 1–15) begins with the story of Samuel's birth, after his mother Hannah (one of the two wives of the Ephraimite Elkanah) had prayed at the shrine at Shiloh, the centre of the tribal confederacy, for a son. She vowed that if she bore a son, he would be dedicated to Yahweh for lifetime service as a Nazirite, as indicated by the words "and no razor shall touch his head."

Three years after she had borne a son, whom she named Samuel—which is interpreted "asked of God," a phrase that fits the meaning of Saul's name but may actually mean "El has heard"—Hannah took the boy to the shrine at Shiloh. Hannah's song of exultation (chapter 12, verses 1–10) probably became the basis of the form and content of the Magnificat, the song that Mary, the mother of Jesus, sang in Luke, chapter 1, verses 46–55, in the New Testament. Eli, the priest at Shiloh (who had heard Hannah's vow), trained the boy to serve Yahweh at the shrine, which Samuel's mother and father visited annually. The sons of Eli, Hophni and Phinehas, are depicted as corrupt, misusing their positions as servants of the shrine to take offerings the people gave to Yahweh for their own gratification, in contrast to Samuel, who "continued to grow in stature and favour with the Lord and with men." Because the sons of Eli failed to heed the admonition of their father, the house of Eli was condemned by a "man of God," who told Eli that his family was to lose its position of trust and power. This condemnation, an interruption of the later source, is the Deuteronomic historian's answer as to why Abiathar, a priest of the family of Eli at the time of David, was excluded from the priesthood at Jerusalem, which became the central shrine of the monarchy.

While a youth (about 12 years old), Samuel experienced a revelation from Yahweh in the shrine at night. First going to Eli three times after hearing his name called, Samuel responded to Yahweh at Eli's suggestion. What

was revealed to him was the fall of the house of Eli, a message that Samuel hesitatingly related to Eli. After this religious experience, Samuel's reputation as a prophet of Yahweh increased.

The fall of Shiloh and the house of Eli

In chapter 4 is an account of the fall of Shiloh and the loss of the ark of the Covenant to the Philistines. Leaving the ark, the symbol of Yahweh's presence, at Shiloh, the Israelites go out to battle against the Philistines near the Mediterranean coast but are defeated. The Israelites return to Shiloh for the ark; but even though they carry it back to the battleground, they are again defeated at great cost—the sons of Eli are killed, and the ark is captured by the enemy. When Eli, old and blind, hears the news of the disaster, he falls over backward in the chair on which he is sitting, breaks his neck, and dies. The wife of his son Phinehas gives birth to a son at this time; and, upon hearing of what had happened to Israel and her family, names the boy Ichabod, meaning "where is the glory?"—because, as she says, "The glory has departed from Israel."

Though the Philistines had captured the ark, they eventually discovered that it did not bring them good fortune. Their god Dagon, an agricultural fertility deity probably meaning "grain," fell to the ground whenever the ark was placed in close proximity to it; and, even more calamitous to them, the Philistines suffered from "tumours," probably the bubonic plague, wherever they carried the ark. After experiencing such disasters for seven months, the Philistines returned the ark to Beth-shemesh in Israelite territory, along with a guilt offering of five golden tumours and five golden mice carried in a cart drawn by two cows. Because many Israelite men in Beth-shemesh also died—"because they looked into the ark of the Lord"—the ark was taken to Kiriath-jearim (the "forest of martyrs" in modern Israel), where it was placed in the house of Abinadab, whose son Eleazar was consecrated to care for it. The ark was not returned to Shiloh, probably because that shrine centre had been destroyed, along with other Israelite towns, by the Philistines.

The request for a monarchy

In chapter 7, verse 3, to chapter 12, verse 25, the Deuteronomic historian depicts the way in which Samuel assumed leadership as judge and Covenant mediator of Israel. The Philistines continued to oppress Israel, though under Samuel's leadership the Israelites were able to reconquer territory lost to their western enemies. When Samuel grew old, his sons were trained to take his place; but they—like the sons of Eli—were corrupt ("they took bribes and perverted justice"), so that the Israelites demanded another form of government—a monarchy. Samuel attempted to dissuade them, pointing out that if they had a highly centralized form of government (*i.e.,* a monarchy), they would have to give up much of their freedom and would be heavily taxed in goods and services. Samuel obeyed both the elders of the people, who demanded a king, and Yahweh, who said, "make them a king."

The rise and fall of Saul. The man selected to become the first monarchical ruler of Israel was Saul, son of Kish, a wealthy Benjamite landowner. Because Kish had lost some donkeys, Saul was sent in search of them. Unsuccessful in his search, he went to the seer-prophet Samuel at Ramah. In the early source, from which this narrative comes, he did not know Samuel's name. The day before Saul went to Ramah, Samuel the seer (*ro'e*), who was depicted by the Deuteronomic historian as a prophet (*navi'*), received notice from Yahweh that Saul was the man chosen to reign over Israel. At the sacrificial meal, Saul, a tall young man, was given the seat of honour, and the next day Samuel anointed him prince (*nagid*) of Israel in a secret ceremony. Before returning home, Saul joined a band of roving ecstatic prophets and prophesied under the influence of the spirit of Yahweh. In chapter 10, verses 17–27, generally accepted as part of the later source, the Deuteronomic historian's views are depicted—Saul was chosen by lot at Mizpah. The early source picks up the story of Saul in chapter 11, which illustrates Saul's military leadership abilities and describes his acclamation as king at Gilgal. Samuel's farewell address, a Deuteronomic reworking of the later source, recapitulates the history of the Israelite tribes from the time of the patriarch Jacob through the period of the judges and forcefully presents the conservative view that the request for a monarchy will bring about adversity to Israel.

The reign of Saul

The early reign of Saul and his confrontations with Samuel until the last judge's death is the subject of chapters 13–15. Saul's early acts as king centred about battles with the Philistines. Because his son Jonathan had defeated one of their garrisons at Geba, the Philistines mustered an army to counterattack near Beth-aven (probably another name for Bethel). Saul issued a request for volunteers, who gathered together for battle but awaited the performance of the sacrifice before the battle by Samuel. Because Samuel did not come for seven days, Saul, acting on his own, presided at the sacrifice. Immediately after the burnt offering had been completed, Samuel appeared (perhaps waiting for such an opportunity to reassert his leading position) and castigated Saul for overstepping the boundaries of his princely prerogatives—even though Saul had been more than patient. Samuel warned him that this type of act (which Saul, in the early source, and later David and Solomon also often performed) would cost Saul his kingdom. In spite of Samuel's apparent animosity, Saul continued to defend the interests of the newly formed kingdom.

The tragedy of Saul was that he was a transitional figure who had to bear the burden of being the man who was of an old order and at the same time of a new way of life among a people composed of disparate elements and leading figures. Both Samuel, the last judge of Israel, and David, the future builder of the small Israelite empire, opposed him. Saul was more a judge—a charismatic leader—than a monarch. Unlike most kings of his time and area, he levied no taxes, depended on a volunteer army, and had no harem. He did not construct a court bureaucracy but relied rather on the trust of the people in his charismatic leadership and thus did not alter the political boundaries or structure of the tribal confederacy.

The dispute between Samuel and Saul

The issue between Saul and Samuel came to a head in the events described in chapter 15 (a section from the later source). Samuel requested Saul to avenge the attacks by the Amalekites on the Israelite tribes during their wanderings in the wilderness after the Exodus from Egypt about 200 years earlier. Saul defeated the Amalekites in a holy war but did not devote everything to destruction as was required by the ban (*herem*). Because Saul had not killed Agag, the Amalekite king, and had saved sheep and cattle for a sacrifice, Samuel informed Saul that he had disobeyed Yahweh and was thus rejected by God, for "to obey is better than to sacrifice." Samuel then asked that Agag be brought to him, and he hacked the Amalekite king to pieces. After that, Saul and Samuel saw each other no more.

The rise and significance of David. The next section contains the account of Saul's fall from power and David's rise to the position of king over all Israel. Samuel, still a charismatic and political power of great consequence, received from Yahweh the message that he was to go to Bethlehem to anoint a new ruler. Because he feared reprisal from Saul, Samuel went to Bethlehem (whose elders had the same fears) under the pretense of presiding at a sacrifice. There he anointed David, son of Jesse, to be future king. David then went to the court of Saul to be the king's armour bearer and court singer.

In a battle with the Philistines David is reported to have killed the 10-foot-tall Philistine champion Goliath of Gath. In II Samuel, chapter 21, verse 19, however, Goliath is killed in a later period by one of David's warriors, Elhanan. According to some biblical scholars, the name of Goliath may have been inserted for an unnamed philistine warrior killed by David apparently while he was armour bearer to Saul and was unrecognized by Saul, thus indicating the reworking of more than one source by the Deuteronomic historian.

The rise of David and the death of Saul

Chapters 18 through 26 depict the rise of David in the court of Saul, his friendship with Jonathan, the beginning of Saul's jealousy of David, the young David's winning of Saul's daughter Michal in marriage for killing a large number of Philistines, Saul's attempt on David's life, David's escape and formation of an outlaw band in the Judaean hills, his acceptance by the priests of the house of Eli at Nob (all of whom were killed by Saul except

Abiathar, who became David's priest), Samuel's death, and other incidents.

Because he feared for his life, David, along with 600 of his men, fled to the Philistine city of Gath, where he became a supposed leader of one of their military contingents against the Israelites. The last four chapters of I Samuel depict the final futile effort of Saul to retain control of his throne and thwart the Philistines: Saul attempted to receive advice from the spirit of the dead Samuel through the necromancer (sometimes called the witch or medium) of Endor, even though he had earlier banned such practices in his realm. Through her mediumship, Samuel foretold the death of Saul and his sons by the Philistines. The armies of the Philistines poured into the Valley of Jezreel. Some of the Philistine leaders distrusted David, who was sent back to his garrison town of Ziklag, which the Amalekites had overrun and in which they had taken many prisoners. Thus, David did not witness the defeat of the Israelites under Saul, who was mortally wounded by the Philistines and whose sons were killed. In an act of heroism so that he, the king of Israel, would not be captured, Saul committed suicide by falling on his own sword. Thus ended the career of the tragic hero who tried to serve Yahweh and Israel but was caught between the old, conservative ways (led by Samuel) and the new, liberal views (championed by David).

The early reign of David

The Second Book of Samuel, as noted earlier, relates the exploits of David and the events of his monarchy. After mourning the death of Saul and executing an Amalekite who claimed to have killed the former king, David began to consolidate his position as the successor to Saul. He was anointed king of Judah at Hebron while Ishbosheth ("man of shame," originally Ishbaal, or "man of Baal"), Saul's son, reigned in the rest of Israel under the guidance of Abner, Saul's general. After seven years, the army of Israel, under Abner, and the army of Judah, under Joab, David's general and nephew, met at Gibeon—each chose 12 champions to fight each other, and all were killed. After the minor battle, a major engagement ensued, with the forces of Judah emerging victorious. A long war of attrition developed between the house of Saul and the house of David. Abner attempted to deliver Israel to David but was killed by Joab to avenge his brother Asahel's death at Abner's hand in the first engagement between the two reigning houses. With Abner dead, Ishbosheth's position became exceedingly insecure, and he was beheaded by two of his own captains, whom David, in turn, executed for murdering the last ruler of the house of Saul.

Because of the course of events, the Israelites asked David to become king over all of Israel, and David made a covenant with the elders of northern Israel. He next engaged in a war with the Jebusite (Canaanite) stronghold of Jerusalem, which he captured. He selected this city as his new capital because it was a neutral site and neither the northerners nor the southerners would be adverse to the selection. From the very beginning of his reign, David showed the political astuteness and acumen that made for him a reputation that has continued for 3,000 years. He built at his new capital a palace, fortified the defenses, and established a harem. The Philistines, concerned about the man whom they had considered a former vassal, decided to move against David, which proved to be their undoing. David effectively contained them in a small area of the Mediterranean coast.

The expansion of the Davidic Empire. The third section of Samuel (II Samuel, chapter 6 through chapter 20, verse 22) contains the account of the reign of David from Jerusalem, ruling over a minor empire that stretched from Egypt in the south to Lebanon in the north and from the Mediterranean Sea in the west to the Arabian Desert in the east. He thus controlled the crossroads of the great empires of the ancient Near East. His second act of political astuteness was to bring theark of the Covenant to Jerusalem; but because of pressures from conservative elements who wanted to retain the tent that housed the ark (which had symbolic value from the days of the Exodus), David was not able to build a temple. Because the ark was now in Jerusalem, however, the city became both the political and the religious cult centre of his kingdom. In chapter 8 is a summary account of David's extension of his kingdom by military means and of the military, administrative, and priestly leaders of Israel.

II Samuel, chapters 9 through 20, verse 22—together with I Kings, chapters 1 and 2, the so-called Succession History, or the Family History of David, which, according to many scholars, forms the oldest section of historiography in Scripture—contains accounts of the domestic problems of David's reign. Though he showed generosity to Mephibosheth, the sole surviving son of the house of Saul, he showed his weakness for the charms of Bathsheba, the wife of Uriah, one of his generals. After ensuring Uriah's death by sending him into the front lines in a battle with the Ammonites, David married Bathsheba, who had become pregnant by the King. When the prophet Nathan came to David and told him of a rich man's unjust actions toward a poor man, David's response was one of anger and a demand for justice, whereupon Nathan said, "You are the man," and that Yahweh would exact retribution by not allowing the child to live. David then repented. He later went to Bathsheba and she conceived and bore another child, Solomon, who was to be the future king of Israel.

Domestic problems of the house of David

Though David was viewed as a master in the art of governing a nation, he was depicted as an unsuccessful father of his family. One son, Amnon (half-brother to Absalom and his sister Tamar), raped Tamar, for which act Absalom later exacted revenge by having Amnon assassinated at a feast. Absalom then fled to Geshur, stayed there three years, was taken back to Jerusalem by Joab, and two years later was reconciled to his father. Absalom's ambition to succeed his father as king caused him to initiate a revolt so that David had to flee from Jerusalem. Absalom was crowned king at Hebron, went to the concubines of David's harem in the palace, and decided to raise a massive army to defeat David. If he had then heeded the advice of Ahithophel, one of David's former counsellors, and attacked David's forces while they were disorganized, he probably would have been successful in retaining the throne. The forces of David under Joab, however, defeated Absalom's army "in the forest of Ephraim." While in flight on a mule, Absalom caught his head in an oak tree, and when Joab heard of his predicament he killed the hanging son of David. When David heard of the death of his rebellious son, he uttered one of the most poignant laments in literature: "O my son Absalom, my son, my son Absalom! Would I had died instead of you, O Absalom, my son, my son!" David then returned to Jerusalem and settled some of the quarrels that had erupted in his absence. A revolt led by the conservative Benjaminite Sheba, under the old rallying cry "every man to his tents, O Israel," was thwarted by Joab, who had to kill David's newly appointed commander Amasa to accomplish this end.

The appendix (chapter 20; verse 23, through chapter 24) has been noted earlier in this section.

Kings. The fourth book of the Former Prophets (I and II Kings in the Septuagint) continues the history of the nation Israel from the death of David, the reign of Solomon, and the divided monarchy through the collapse of both Israel (the northern kingdom) and Judah (the southern kingdom). Whereas Samuel was composed primarily of the early and the later sources with some editing on the part of the Deuteronomic historians, the Deuteronomic editors of Kings, in addition to these two sources, used other sources—such as the book of the acts of Solomon, the Book of the Chronicles of the Kings of Israel, the Book of the Chronicles of the Kings of Judah, temple archives, and traditions centring on certain major kings and prophets. The Deuteronomic historians wrote from the vantage points of the reign of King Josiah of Judah, who died in 609 BCE and was the ruler who accepted the Deuteronomic reform that began in 621 BCE, and of the Babylonian Exile, which traditionally lasted 70 years, though it began in 597 BCE, the temple was destroyed in 586, some exiles returned in 538, and the temple was restored in 516. The Deuteronomic view that national apostasy was the cause of the covenant people's predicament pervades this work.

The sources of the Book of Kings

(The history of the 10th through the early 6th century BCE is covered in the article JUDAISM, and therefore this

article will concentrate only on the reigns of important monarchs and their relationships to the rising power of the prophetic movement in Israel.)

The Book of Kings may be divided into four sections: (1) the last years of David and Solomon's succession to the throne (I Kings, chapter 1, to chapter 2, verse 11); (2) the reign of Solomon (I Kings, chapter 2, verse 12, to chapter 11, verse 43); (3) the beginning of the divided monarchy to the fall of Israel (I Kings, chapter 12, to II Kings, chapter 17); and (4) the last years of Judah (II Kings, chapters 18–25).

The succession of Solomon to the throne. I Kings (chapters 1 and 2) continues the story of David and the struggle for the succession of his throne. The sides were drawn between Adonijah, David's eldest living son, and Solomon, the son of David and Bathsheba. Supporting Adonijah were the "old guard"—the general Joab and the priest Abiathar—and supporting Solomon were the priest Zadok, the prophet Nathan, and the captain of David's bodyguard, Benaiah. With David close to death, Adonijah prepared to seize control of the kingdom; Nathan, however, requested Bathsheba to go to David and persuade David to proclaim Solomon the next monarch. Following the advice of Nathan, David then appointed Solomon the heir to his throne; and Zadok the priest and Nathan the prophet anointed the son of Bathsheba king in Gihon.

The appointment of Solomon as the future king of Israel

After David died, however, Adonijah attempted to regain some semblance of prestige by asking Solomon to give him Abishag, a young Shunammite woman who had been given to David in his old age, as his wife. To this request Solomon answered by ordering Adonijah's execution, which Benaiah carried out. Solomon also ordered the execution of the old general Joab for having killed Abner and Amasa years earlier as a loyal supporter of David, an execution again carried out by Benaiah, who also executed Shimei, a man who had cursed David a long time earlier. Prior to these executions, which David—before he had died—had requested of Solomon, the new king banished the priest Abiathar of the house of Eli to Anathoth, an act that confirmed the position of Zadok as the principal priest of Jerusalem.

The reign of Solomon. David had reigned from about 1000 to 962 BCE, a period in which he consolidated a federation of tribes that had been united under the charismatic leadership of Saul, who had reigned for about two decades before David began to construct his minor empire. Solomon, who inherited a strong monarchy, reigned for 40 years. His reputation as a monarch centred about his great wisdom (chapter 3), his reorganization of the administrative bureaucracy (chapter 4), and his building of the magnificent Temple (chapters 3–8). Though two sons of the prophet Nathan served Solomon, one as a court official and another as a priest, the prophetic movement apparently was little encouraged by the united monarchy's third king. Solomon is perhaps one of the most overrated figures in the Old Testament, in spite of his achievements in wisdom, construction, and commerce; he is recorded as having 1,000 wives and concubines—some of them merely guarantees of commercial treaties, to be sure—and as building a fleet of ships for a nearly landlocked Israel. To accommodate his desire for a seaport, he built the port of Ezion-geber at the head of the Gulf of Aqaba of the Red Sea. A son of the harem, Solomon had had little contact with the people of his realm, and he used many of them in labour battalions in his vast building programs to the economic disadvantage of Israel. By fostering social discontent in such ventures, Solomon prepared the way for the disintegration of the united kingdom and the resurgence of the prophetic movement that reflected the indigenous covenant concept peculiar to Israel.

Whereas David secured Israel's borders and property by military means, Solomon sought to extend Israel's influence through commercial treaties. To secure diplomatic and commercial treaties Solomon contracted marriage with various princesses—who brought with them their native deities. This defection from the Covenant obligations to Yahweh is viewed by the Deuteronomic historian as a continuance of Israel's constant flirting with apostasy, which had occurred under the judges, and the beginning of a long process of internal religious and political disintegration under the monarchical system. Solomon's oppressive taxation and commercial expansion also brought about retaliation and rebellion.

The divided monarchy. After Solomon died (922 BCE), he was succeeded by Rehoboam, who proved to be unfit for the task of reigning. Prior to Solomon's death, Jeroboam the Ephraimite, a young overseer of the forced labour battalions of the "house of Joseph" in the north, had encountered Ahijah, a prophet from the old shrine of the confederacy at Shiloh, and Ahijah had torn a new garment into 12 pieces, prophesying that 10 pieces (tribes) would be given to Jeroboam and only two pieces (tribal political units) would be retained by the house of David. The dismemberment of the united monarchy was to be brought about by Yahweh because Solomon had "not walked in my ways, doing what is right in my sight and keeping my statutes and my ordinances, as David his father did." Though Solomon had worshipped the Sidonian goddess Ashtoreth, the Moabite god Chemosh, and the Ammonite god Milcom, his reign over Israel continued. Jeroboam's initial rebellion proved to be abortive, and he sought political asylum in Egypt under the protection of the pharaoh Sheshonk I (Shishak).

The rise of the divided kingdom: Israel and Judah

Rehoboam, having been crowned king of the united monarchy in Jerusalem, went north to Shechem, a shrine centre of the 10 northern tribes of the old confederacy, to have his position ratified by the northern units of the kingdom. Using this gathering as an opportune time to present their grievances against Solomon's oppressive domestic policies, the northerners, under the leadership of the returned political fugitive Jeroboam, asked the king from Jerusalem to lighten their load. Requesting three days to take their grievances under advisement, Rehoboam sought counsel from his advisers. The older counsellors advised moderation, the younger, retaliation. Assenting to the latter, Rehoboam returned to the people with an answer that was to lead to the disintegration of the united monarchy that had lasted for only about a century under three kings: "My father made your yoke heavy, but I will add to your yoke; my father chastised you with whips, but I will chastise you with scorpions." The response of the northerners was the ancient battle cry, "To your tents, O Israel." Rehoboam, ruling from the cities, sent Adoram, the leader of the forced labour battalions, to Israel (the name to be used henceforth for the northern area); but he was stoned to death. The uncrowned king of the north, unable to quell the rebellion, returned to Jerusalem in rapid flight. Heeding the advice of the prophet Shemaiah, Rehoboam allowed the situation to remain that of a stalemate, thus inaugurating the period of the divided monarchy that lasted in Israel in the north from 922–721 BCE and in Judah in the south until 586 BCE.

Turmoil in the monarchy of the northern kingdom

Though the Davidic monarchy continued in Judah until the fall of Jerusalem in 586 BCE, the monarchial situation in Israel was one of constant turmoil and confusion, except for the periods of a few dynasties. Jeroboam I of Israel (reigned 922–901 BCE) attempted to bring about religious and political reforms. Establishing his capital at Shechem, he set aside two pilgrimage sites (Dan in the north and Bethel in the south) as shrine centres. Though the Deuteronomic historian—with an anti-north prejudice—interpreted Jeroboam's use of golden bulls in the high place sanctuaries as a sin against Yahweh, Jeroboam's actions may have merely been an incorporation of religious symbols similar to the cherubim (winged animals) that guarded the empty throne of Yahweh in the temple of Solomon in Jerusalem. Jeroboam would not have been so politically and religiously naïve as to introduce polytheistic practices among the conservative-minded tribes of northern Israel. Thus, the golden bulls may have been meant to serve as pedestals for the invisible Yahweh just as the ark (throne) may have been the seat of the invisible Yahweh in the Holy of Holies (inner sanctuary) of the Temple in Jerusalem. Gods (such as the storm god Hadad) of other Syrian and Palestinian religions also were represented as standing on the backs of bulls.

Jeroboam remained true to Yahwistic religion, however, in that the God of the Israelites was not represented icono-

graphically. The first king of the northern kingdom also inaugurated other religious reforms or reinstituted ancient practices that were interpreted as decadent by the Deuteronomic historian of the southern kingdom of Judah. He instituted a harvest thanksgiving festival on the 15th day of the eighth month, a change in the religious calendar that would preclude the journey of many northern Israelites to a similar festival in Jerusalem; he reformed the priesthood by installing non-Levites (the traditional shrine functionaries) to serve Yahweh at the shrines, an action that had been carried out in Jerusalem by David but without the opprobrium inferred by the Deuteronomic historian on a similar action by Jeroboam.

The dynasties of the northern kingdom were shortlived. Jeroboam was succeeded by his son Nadab, who reigned for two years before he was overthrown by Baasha, who decimated the house of Jeroboam. Reigning for 24 years, Baasha (who "did what was evil in the sight of the Lord" like all of the northern kings, according to the interpretation of the Deuteronomists) had to concern himself not only with charismatic leaders who were traditionally powerful in the north but also with the rising power of antimonarchical prophets, such as Jehu—who prophesied the end of the house of Baasha (chapter 16). Elah, Baasha's son, ruled only two years before he was assassinated while in a drunken state by Zimri, a chariot commander, who exterminated all of the members of the house of Baasha. Reigning for the brief period of seven days, Zimri was besieged in the citadel at Tirzah by Omri, commander of the army. Zimri burned to death in the king's house. Much of this political turmoil and confusion in the north occurred during the reign of Asa, king of Judah from *c.* 913 to 873 BCE, who inaugurated religious reforms, such as banning male cult prostitutes and the worship of the Canaanite goddess Asherah that had been sponsored by his mother, Maachah, the queen regent.

The dynasty of Omri and the career of Elijah

The significance of Elijah. With the dynasty of Omri (*c.* 876–842), the prophetic movement begins to assume a position of tremendous importance in Israel and Judah. Omri (reigned *c.* 876–869) reestablished Israel's economic and military significance among the Syrian and Palestinian minor kingdoms, so much so that years after his death the Assyrians referred to the northern kingdom as "the land of Omri." He is mentioned in the Moabite Stone of King Mesha (9th century BCE) as a king who "humbled Moab many years." To strengthen an alliance with the Phoenicians, Omri contracted a marriage between Jezebel, princess of Sidon, and his son Ahab. The marriage proved to be fateful for Israel and was a catalyst that brought the prophetic movement into a course of action and a form that became Israel's contribution to Near Eastern prophecy.

The reign of Omri's son Ahab coincided with the activities of the prophet Elijah, as recorded in I Kings, chapter 16, verse 29, to chapter 22, verse 40. Ahab, under the influence of his queen Jezebel, allowed her to foster the worship of the fertility god Baal in Samaria—the capital that Omri had built—and in all Israel, even though he himself remained a worshipper of Yahweh. A temple was built for Baal in Samaria; Jericho was rebuilt (even though the ban against its existence still remained) by Hiel of Bethel, who sacrificed two of his own sons and placed them in the foundation and the gates of the walls of the city. During these apostate activities the great prophet Elijah the Tishbite appeared. A man of erratic behaviour, wearing a garment of hair with a leather belt around his waist, using uncouth language, and preferring the wilderness areas to the towns, Elijah bore many of the outward signs of social rebels. At odds with the court authorities, he began his prophetic career just prior to a retreat in the wilderness during a drought, which he had announced to Ahab, thus pointing out that Yahweh, rather than Baal, is the Lord of nature. In the desert he performed two miracles: he ensured a widow and her son of continuous food for her act of generosity to him and cured her son, apparently dead, who had stopped breathing, by stretching himself on top of the boy three times. Elijah then went to the court of Ahab at Samaria, after having met one of the leading prophets (Obadiah) who had escaped Jezebel's attempt to destroy the leaders of the cult of Yahweh, and stood before Ahab, accusing the king of being the "troubler of Israel" for having followed the cult of Baal. Elijah hurled a challenge to the Baalists, supported by Jezebel, to meet him in a contest on Mt. Carmel.

The Mt. Carmel contest

The contest between Elijah and the 450 prophets of Baal was dramatic. Elijah first taunted the spectators, "How long will you go limping with two different opinions? If the Lord is God, follow him; but if Baal, then follow him." Elijah then laid the ground rules: two bulls were to be sacrificed, one each on an altar, on which firewood was to be laid, but no one was to light the fire—only the God "who answers by fire." The prophets of Baal had the first opportunity, and they prayed to Baal loudly for a full half day, until noon. During this time, Elijah, in coarse language, taunted them. Eliminating the euphemisms in most English versions of the Bible, Elijah mocked the Baalists by saying that Baal might not be responding because he was out urinating ("gone aside"), on a trip, or sleeping. The Baalists then attempted to use sympathetic magic. By cutting themselves they hoped that as their life blood flowed on the ground Baal would send rain, the life blood of the Earth.

When the Baalists had failed, Elijah rebuilt an old altar of Yahweh, poured water on the wood three times (perhaps a remnant of an ancient rainmaking ceremony?), and prayed to Yahweh to answer his servant; "the fire of the Lord fell, and consumed the burnt offering, and the wood, and the stones and the dust, and licked up the water that was in the trench." Though some authorities explain the action by suggesting that Elijah poured naphtha on the wood, this does not explain the ignition of the wood at that particular time and that particular place even if by a bolt of lightning. The Deuteronomic historian emphasized the miracle wrought by Yahweh. The people, upon witnessing the miracle, cried out, "Yahweh, he is God," and proceeded to annihilate the prophets of Baal.

Elijah told Ahab to complete the festivities while he went to the top of Mt. Carmel to perform another rainmaking ceremony. When the rains came in a cloudburst, Ahab was riding in his chariot in the Valley of Jezreel. Elijah, in fear of retaliation from Jezebel, fled to the southern wilderness. At Mt. Horeb (Sinai) after a storm, wind, and an earthquake, Yahweh spoke to Elijah through silence and then revealed that he should anoint Hazael to be king of Syria, Jehu to be king of Israel, and Elisha to be his successor as prophet. I Kings, chapter 20, records a war between Ben-hadad, king of Syria, and Ahab. Though Ahab was victorious, he did not kill Ben-hadad according to the provisions of the *ḥerem* (ban); and a prophet then informed Ahab that he would suffer for his inaction.

The affair of Naboth's vineyard

Upon Ahab's return to Samaria Jezebel attempted to coerce the king into confiscating the vineyards of Naboth of Jezreel, which was a Canaanite centre. Naboth asserted that as an Israelite the land was not his own but was a trust from Yahweh and that he could not sell it. Taken to court on trumped-up charges of blasphemy, Naboth was convicted and stoned to death. Ahab, following Jezebel's advice, then went to Naboth's vineyard and took possession of it. Upon hearing of Ahab's unjust act as king, Elijah proclaimed to him, "In the place where dogs licked up the blood of Naboth shall dogs lick your own blood." The prophet also announced, "The dogs shall eat Jezebel within the bounds of Jezreel."

In I Kings, chapter 22, another prophet, Micaiah, prophesied to Ahab and to King Jehoshaphat of Judah who were preparing for battle against the Syrians that in a vision he saw "all Israel scattered upon the mountains, as sheep that have no shepherd." Micaiah was put in prison to test the validity of his vision. It turned out to be true—Ahab, even though he disguised himself, was mortally wounded by an arrow shot by a Syrian archer. In 850 he was succeeded by his son Ahaziah, who reigned for only two years.

The Second Book of Kings continues the history of the monarchies of Israel and Judah and of the prophetic movement. Ahaziah fell from an upper chamber of his palace in Samaria and sought help from Baalzebub, the god of Ekron. Elijah met the messengers to castigate them for not seeking aid from Yahweh, the God of Israel, and

told a third delegation that had been sent out to return to tell Ahaziah that because of his apostasy he would die. After the death of Ahaziah, Elijah conferred his mantle, the symbol of his prophetic authority, on Elisha, and "Elijah went up by a whirlwind into heaven."

The significance of Elisha. The stories of Elijah and his successor, Elisha, are of a different literary genre from the historical accounts of the political developments of the 9th century. The historical accounts are based on the viewpoints and biases of the monarchy, nobility, and military leaders. The stories of Elijah and Elisha are legendary, popular accounts, probably having arisen among the common people. They demonstrate the predilection of the common people to accent what appears to them as the miraculous and the supernatural, much as has been the case among many Roman Catholics and Eastern Christians in stories of their saints. Elijah was depicted, in several instances, as a second Moses—*e.g.,* he fled to the wilderness to escape the retaliation of a ruler, and he encountered a theophany (manifestation of a deity) of Yahweh on Mt. Horeb. As Moses appointed Joshua as his successor, so also Elijah passed on his prophetic mantle to Elisha. Elisha is depicted in typical folk story embellishments and legendary motifs. The original beginning and ending of the Elijah story apparently was lost, but the Deuteronomic historian incorporated the popular accounts of Elijah and Elisha into the court history that gives scholars significant insights into the religious movements of the 9th century.

The prophetic career of Elisha and the end of the Omri dynasty

During the reigns of King Jehoshaphat of Judah (*c.* 873–849 BCE) and King Jehoram (Joram) of Israel (*c.* 849–842), Elisha began his prophetic career. Elisha was unlike his mentor Elijah in many ways: he did not use uncouth language, he did not shun towns, he wore more fashionable clothing, and he used music to bring about the prophetic spirit—much as Saul had done earlier. A cycle of miracle stories arose around Elisha; he was said to have made bitter water sweet, revived the son of a Shunammite woman from death by breathing into his mouth and lying on top of him, helped a woman to avoid giving up her two sons to a creditor who would make them slaves, informed the Syrian captain Naaman how to be cured from his skin disease, and many other similar actions. In addition to being a miracle worker, Elisha was a political power. He prophesied the defeat of the Moabites as a result of a huge rainfall and advised Joram how to defeat Ben-hadad, king of Syria. By performing this last act Elisha instigated a revolt in Syria; Hazael murdered the sick and dying Ben-hadad.

Elisha sent "one of the sons of the prophets" to anoint Jehu, an army commander, to be the future king of Israel. Rushing in his chariot to Jezreel, Jehu exterminated Jehoram, the last king of the Omri dynasty, his nephew Ahaziah (king of Judah), who was visiting him, and the queen mother Jezebel, who "had painted her eyes, and adorned her head" before she was thrown out of the window and so mangled by the trampling of horses that "they found no more of her than the skull and the feet and the palms of her hands." Jezebel's end had come about in a manner similar to the way in which Elijah had prophesied.

The revolution of Jehu was not only politically inspired. A driving force behind him was the arch conservative Rechabite faction, led by Jehonadab. Despising the Canaanites and their agricultural way of life, the Rechabites—descendants of the ancient Kenites of Midian where Moses had experienced the theophany of the burning bush—lived in tents, refused to drink wine, and attempted to retain as many of the accoutrements of the "good old life" of ancient nomadism as possible. With excessive revolutionary zeal they helped Jehu to annihilate the worshippers of Baal, who were tricked into coming to their temple and there murdered. To further emphasize their revolutionary intent, the followers of Jehu, in addition to the holocaust, made the site of the temple of Baal a latrine.

Because the king of Judah (Ahaziah) had been killed in the revolution—along with the remaining northern members of the house of Omri—the southern kingdom was ruled over by the queen mother, Athaliah, the daughter of Ahab and Jezebel. In her zeal to propagate the faith of her mother, Athaliah seized the opportunity to destroy the line of David that tended to be loyal to Yahweh. Liquidating all the male heirs to the throne of David—except the infant Joash (Jehoash) who received asylum in "the house of the Lord"—Athaliah ruled for six years. With support from the priests led by Jehoiada, the army and "the people of the land" revolted, killing Athaliah and her high priest of Baal, Mattan, and destroying the temple of Baal.

The reign of Jeroboam II and the demise of the northern kingdom

In the north, Jehu was succeeded by his son Jehoahaz (reigned *c.* 815–*c.* 801), who, in turn, was followed by his son Joash, or Jehoash. During the latter king's reign, the prophet Elisha died. Though the Deuteronomic historian says little about Israel's next king, Jeroboam II, he was a major monarch, reestablishing the northern kingdom's ancient boundaries and fostering a period of economic prosperity. During the reign of Jeroboam II (*c.* 786–*c.* 746 BCE), a time of both economic advances and social injustice, Amos, the great prophet of social justice, arose. During Jeroboam's last years another great prophet, Hosea, whose message centred on Covenant love, arose to call an apostate people back to their Covenant responsibilities.

The fall of Israel. After the death of Jeroboam II, however, Israel faced a period of continuous disaster; and no prophetic figure was able to arrest the steady internal decay. From 746–721, when Samaria finally fell to the Assyrians, there were six kings, the last being Hoshea, a conspirator who had assassinated the previous king. The Assyrian king Sargon II deported the leading citizens of Samaria to Persia and imported colonists from other lands to fill their places.

The fall of Judah. The southern kingdom of Judah, under the Davidic monarchy, was able to last about 135 years longer, often only as a weak vassal state. Hezekiah (reigned *c.* 715–*c.* 687), with the advice of the prophet Isaiah, managed to avoid conflict with or outlast a siege of the Assyrians. Hezekiah was succeeded by his son Manasseh, an apostate king who stilled any prophetic outcries, reintroduced Canaanite religious practices, and even offered his son as a human sacrificial victim. Soothsaying, augury, sorcery, and necromancy were also reintroduced. The Deuteronomic historian also notes that many innocent persons were killed during his reign. Manasseh was succeeded by his son Amon, who was assassinated in a palace revolution after a reign of only two years. His son Josiah, who succeeded him, reigned from 640 to 609 BCE, when he was killed in a battle with the pharaoh Necho II of Egypt. During his reign, one of the most significant events in the history of the Israelite people occurred—the Deuteronomic reform of 621 BCE. Occasioned by the discovery of a book of the Law in the Temple during its rebuilding and supported not only by Hilkiah, a high priest, and Huldah, a prophetess, but also by the young prophet Jeremiah, the Deuteronomic Code—or Covenant—as it has been called, became the basis for a far-reaching reform of the social and religious life of Judah. Though the reform was short-lived, because of the pressure of international turmoil, it left an indelible impression on the religious consciousness of the people of the Covenant, Israel, whether they were from the north or the south.

Reforms in Judah and the Babylonian Exile

From 609 to 586 Judah felt the coming oppression of Babylon under King Nebuchadrezzar. After the death of Josiah, four kings ruled in Jerusalem, the last being Zedekiah, who failed to heed the advice of the prophet Jeremiah—who had attempted to persuade the king not to trust the Egyptians in a rebellion against Babylon because there would be only one loser, the House of David. Jehoiachin, the predecessor of the puppet king Zedekiah, had been carried off into exile to Babylon in 598; but about 560 he was released from prison, thus leaving a hope that the Davidic line had not become extinct. Despite this small element of hope, the year 586 BCE marked the beginning of a tragic period for the people of Judah—the Babylonian Exile. During this period of rethinking Covenant faith, the prophet Ezekiel preached, both in Jerusalem and Babylon, offering the people hope for a restoration of the symbols and cultic acts of their covenant religion.

Isaiah. The Book of Isaiah, comprising 66 chapters, is one of the most profound theological and literarily expressive works in the Bible. Compiled over a period of

The division of Isaiah into two or three distinct sections

about two centuries (the latter half of the 8th to the latter half of the 6th century BCE), the Book of Isaiah is generally divided by scholars into two (sometimes three) major sections, which are called First Isaiah (chapters 1–39), Deutero-Isaiah (chapters 40–55 or 40–66), and—if the second section is subdivided—Trito-Isaiah (chapters 56–66).

The prophecies of First Isaiah. First Isaiah contains the words and prophecies of Isaiah, a most important 8th-century BCE prophet of Judah, written either by himself or his contemporary followers in Jerusalem (from *c.* 740 to 700 BCE), along with some later additions, such as chapters 24–27 and 33–39. The first of these two additions was probably written by a later disciple or disciples of Isaiah about 500 BCE; the second addition is divided into two sections—chapters 33–35, written during or after the exile to Babylon in 586 BCE, and chapters 36–39, which drew from the source used by the Deuteronomic historian in II Kings, chapters 18–19. The second major section of Isaiah, which may be designated Second Isaiah even though it has been divided because of chronology into Deutero-Isaiah and Trito-Isaiah, was written by members of the "school" of Isaiah in Babylon: chapters 40–55 were written prior to and after the conquest of Babylon in 539 by the Persian king Cyrus II the Great, and chapters 56–66 were composed after the return from the Babylonian Exile in 538. The canonical Book of Isaiah, after editorial redaction, probably assumed its present form during the 4th century BCE. Because of its messianic (salvatory figure) themes, Isaiah became extremely significant among the early Christians who wrote the New Testament and the sectarians at Qumrān near the Dead Sea, who awaited the imminent messianic age, a time that would inaugurate the period of the Last Judgment and the Kingdom of God.

Themes peculiar to Isaiah and his call to become a prophet

Isaiah, a prophet, priest, and statesman, lived during the last years of the northern kingdom and during the reigns of four kings of Judah: Uzziah (Azariah), Jotham, Ahaz, and Hezekiah. He was also a contemporary of the prophets of social justice: Amos, Hosea, and Micah. Influenced by their prophetic outcries against social injustice, Isaiah added themes peculiar to his prophetic mission. To kings, political and economic leaders, and to the people of the land, he issued a message that harked back nearly five centuries to the period of the judges: the holiness of Yahweh, the coming Messiah of Yahweh, the judgment of Yahweh, and the necessity of placing one's own and the nation's trust in Yahweh rather than in the might of ephemeral movements and nations. From about 742 BCE, when he first experienced his call to become a prophet, to about 687, Isaiah influenced the course of Judah's history by his oracles of destruction, judgment, and hope as well as his messages containing both threats and promises.

Intimately acquainted with worship on Mt. Zion because of his priest-prophet position, with the Temple and its rich imagery and ritualistic practices, and possessed of a deep understanding of the meaning of kingship in Judah theologically and politically, Isaiah was able to interpret and advise both leaders and the common people of the Covenant promises of Yahweh, the Lord of Hosts. Because they were imbued with the following beliefs—God dwelt on Mt. Zion, in the Temple in the city of Jerusalem, and in the person of the King—the messianic phrase "God is with us" (Immanuel) Isaiah used was not a pallid abstraction of a theological concept but a concrete living reality that found its expression in the Temple theology and message of the great prophet.

In chapters 1–6 are recorded the oracles of Isaiah's early ministry. His call, a visionary experience in the temple in Jerusalem, is described in some of the most influential symbolic language in Old Testament literature. In the year of King Uzziah's death (742 BCE), Isaiah had a vision of the Lord enthroned in a celestial temple, surrounded by the seraphim—hybrid human-animal-bird figures who attended the deity in his sanctuary. Probably experiencing this majestic imagery that was enhanced by the actual setting and the ceremonial and ritualistic objects of the Jerusalem Temple, Isaiah was mystically transported from the earthly temple to the heavenly temple, from the microcosm to the macrocosm, from sacred space in profane time to sacred space in sacred time.

Yahweh, in the mystical, ecstatic experience of Isaiah, is too sublime to be described in other than the imagery of the winged seraphim, which hide his glory and call to each other:

> "Holy, holy, holy is the Lord of hosts;
> The whole earth is full of his glory."

With smoke rising from the burning incense, Isaiah was consumed by his feelings of unworthiness ("Woe is me! for I am lost"); but one of the seraphim touched Isaiah's lips with a burning coal from the altar and the prophet heard the words, "Your guilt is taken away, and your sin forgiven." Isaiah then heard the voice of Yahweh ask the heavenly council, "Whom shall I send, and who will go for us?" The prophet, caught up as a participant in the mystical dialogue, responded, "Here am I! Send me." The message to be delivered to the Covenant people from the heavenly council, he is informed, is one that will be unheeded.

The early oracles of Isaiah

The oracles of Isaiah to the people of Jerusalem from about 740 to 732 BCE castigate the nation of Judah for its many sins. The religious, social, and economic sins of Judah roll from the prophet's utterances in staccato-like sequence: (1) "Bring no more vain offerings; incense is an abomination to me. New moon and sabbath and the calling of assemblies—I cannot endure iniquity and solemn assembly," against religious superficiality; (2) "cease to do evil, learn to do good; seek justice, correct oppression; defend the fatherless, plead for the widow," against social injustice; and (3) "Come now, let us reason together, says the Lord: though your sins are like scarlet, they shall be as white as snow," a call for obedience to the Covenant. The prophet also cried out for peace: "and they shall beat their swords into plowshares, and their spears into pruning hooks; nation shall not lift up sword against nation, neither shall they learn war anymore." The sins of Judah, however, are numerous: the rich oppress the poor, the nation squanders its economic resources on military spending, idolatry runs rampant in the land, everyone tries to cheat his fellowman, women flaunt their sexual charms in the streets, and there are many who cannot wait for a strong drink in the morning to get them through the day. One of Isaiah's castigations warns: "Woe to those who are heroes at drinking wine, and valiant men in mixing strong drink, who acquit the guilty for a bribe, and deprive the innocent of his right!"

Isaiah's prophetic attacks on Judah's foreign policy

During the Syro-Ephraimitic war (734–732 BCE), Isaiah began to challenge the policies of King Ahaz of Judah. Syria and Israel had joined forces against Judah. Isaiah's advice to the young King of Judah was to place his trust in Yahweh. Apparently Isaiah believed that Assyria would take care of the northern threat. Ahaz, in timidity, did not want to request a sign from Yahweh. In exasperation Isaiah told the King that Yahweh would give him a sign anyway: "Behold, a young woman shall conceive and bear a son, and shall call his name Immanuel." Thus, by the time this child is able to know how to choose good and refuse evil, the two minor kings of the north who were threatening Judah will be made ineffective by the Assyrians. The name Immanuel, "God is with us," would be meaningful in this situation because God on Mt. Zion and represented in the person of the king would be faithful to his Covenant people. Ahaz, however, placed his trust in an alliance with Assyria under the great conqueror Tiglath-pileser III. In order to give hope to the people, who were beginning to experience the Assyrian encroachments on Judaean lands in 738 BCE, Isaiah uttered an oracle to "the people who walked in darkness": "For to us a child is born, to us a son is given; and the government will be upon his shoulder, and his name will be called Wonderful Counselor, Mighty God, Everlasting Father, Prince of Peace." Isaiah trusted that Yahweh would bring about a kingdom of peace under a Davidic ruler.

From 732 to 731 BCE, the year the northern kingdom fell, Isaiah continued to prophesy in Judah but probably not in any vociferous manner until the Assyrians conquered Samaria. The king of the Assyrians is described as the rod of God's anger, but Assyria also will experience the judgment of God for its atrocities in time of war. During

one of the periods of Assyrian expansion towards Judah, Isaiah uttered his famous Davidic messianic (salvatory figure) oracle in which he prophesies the coming of a "shoot from the stump of Jesse," upon which the Spirit of the Lord will rest and who will establish the "peaceable kingdom" in which "the wolf shall dwell with the lamb." A hymn of praise concludes this first section of First Isaiah.

Chapters 13–23 include a list of oracles against various nations—Babylon, Assyria, Philistia, Moab, Syria, Egypt, and other oppressors of Judah. These probably came from the time when Hezekiah began his reign (*c.* 715). In 705 BCE, Sargon of Assyria died, however, and Hezekiah, a generally astute and reform-minded king, began to be caught up in the power struggle between Babylon, Egypt, and Assyria. Isaiah urged Hezekiah to remain neutral during the revolutionary turmoil. Though Sennacherib of Assyria moved south to crush the rebellion of the Palestinian vassal states, Isaiah—contrary to his previous advocacy of neutrality—urged his king to resist the Assyrians because the Lord, rather than the so-called Egyptian allies, who "are men, and not God," will protect Jerusalem. He then prophesied a coming age of justice and of the Spirit who will bring about a renewed creation.

Second Isaiah (chapters 40–66), which comes from the school of Isaiah's disciples, can be divided into two periods: chapters 40–55, generally called Deutero-Isaiah, were written about 538 BCE after the experience of the Exile; and chapters 56–66, sometimes called Trito-Isaiah (or III Isaiah), were written after the return of the exiles to Jerusalem after 538 BCE.

Deutero-Isaiah and the Servant Songs

The prophecies of Deutero-Isaiah. Second Isaiah contains the very expressive so-called Servant Songs—chapter 42, verses 1–4; chapter 49, verses 1–6; chapter 50, verses 4–9; chapter 52, verse 13; and chapter 53, verse 12. Writing from Babylon, the author begins with a message of comfort and hope and faith in Yahweh. The people are to leave Babylon and return to Jerusalem, which has paid "double for all her sins." As creator and Lord of history, God will redeem Israel, his chosen servant. Through the Servant of the Lord all the nations will be blessed: "I have put my Spirit upon him, he will bring forth justice to the nations." The Suffering Servant, whether the nation Israel or an individual agent of Yahweh, will help to bring about the deliverance of the nation. Though Second Isaiah may have been referring to a hoped-for rise of a prophetic figure, many scholars now hold that the Suffering Servant is Israel in a collective sense. Christians have interpreted the Servant Songs, especially the fourth, as a prophecy referring to Jesus of Nazareth—"He was despised and rejected by men; a man of sorrows and acquainted with grief . . . ," but this interpretation is theologically oriented and thus open to question, according to many scholars.

The oracles of Trito-Isaiah. Chapters 56–66 are a collection of oracles from the restoration period (after 538 BCE). Emphasis is placed upon cultic acts, attacks against idolatry, and a right motivation in the worship of Yahweh. Repentance and social justice are themes that have been retained from the earlier Isaiah traditions, and the ever-present element of hope in the creative goodness of Yahweh that pervaded II Isaiah remains a dominant theme in the last chapters of the Book of Isaiah.

Jeremiah. The prophet Jeremiah began to prophesy about 626 BCE during the reign of the Judaean king Josiah. From the town of Anathoth and probably from the priestly family of Eli, this prophet, who may have been instrumental in the Deuteronomic reform, dictated his oracles to his secretary Baruch. Only a youth in his late teens when he experienced the call by Yahweh to be a "prophet to the nations," Jeremiah was a hesitant reforming prophet, experiencing deep spiritual struggles regarding his adequacy from the very beginning of his call and throughout his prophetic ministry. After the death of Josiah in 609 BCE, however, he became an outspoken prophet against the national policy of Judah, a policy that he knew would lead to the disaster that came to be called the Babylonian Exile. Because of his prophecies, which were unpopular with the military and the revolutionists against the Babylonians, Jeremiah was kidnapped by conspirators after 586 and taken to Egypt, where he disappeared.

The call of Jeremiah and his difficulties with political and revolutionary leaders

The Book of Jeremiah is a collection of oracles, biographical accounts, and narratives that are not arranged in any consistent chronological or thematic order. One 20th-century German biblical scholar, Wilhelm Rudolph, has attempted to arrange the chapters of the book according to certain chronological details. He has divided the work into five sections: (1) prophecies against Judah and Jerusalem, chapters 1–25, during the reigns of kings Josiah (640–609) and Jehoiachim (609–598), and the period after Jehoiachim (597–586); (2) prophecies against foreign nations, chapters 25 and 66–61; (3) prophecies of hope for Israel, chapters 26–35 (probably after the death of Josiah in 609); (4) narratives of Jeremiah's sufferings, chapters 36–45 (from a post-586 period), and (5) an appendix, chapter 52. Jeremiah's own prophetic oracles are found particularly in chapters 1–36 and 46–52. Baruch's writings about Jeremiah are found primarily in chapters 37–45, 26–29, and 33–36.

During the reign of Josiah, after his call, Jeremiah preached to the people of Jerusalem and warned them against the sin of apostasy. Recalling the prophecies of the 8th-century Israelite prophet Hosea, Jeremiah reproached the Judaeans for playing harlot with other gods and urged them to repent. He prophesied that enemies from the north would be the instruments of Yahweh's judgment on the apostate land and Jerusalem would suffer the fate of a rejected prostitute. The idolatry and immorality of the Judaeans would inevitably lead to their destruction. Because of the impending threat from the north, Jeremiah warned the people to flee from the wrath that was to come.

At the beginning of Jehoiachim's reign, Jeremiah preached in the temple that because of Judah's apostasy "death shall be preferred to life by all the remnant that remains of this evil family in all the places where I have driven them, says the Lord of hosts." Because he spoke words that were unpopular, his own townsmen of Anathoth plotted against his life. To symbolize the fate of Judah, Jeremiah adopted some rather bizarre techniques. He buried a waist cloth and wore it when it was spoiled to illustrate the fate of Jerusalem, which had worshipped other gods than Yahweh.

Throughout his career Jeremiah had moments of deep depression, times when he lamented that he had become a prophet. Because of the uncertainty of the times, Jeremiah did not marry.

Jeremiah's depression and allegorical messages

A master of symbolic actions and the use of symbolic devices, Jeremiah used a potter's wheel to show that Yahweh was shaping an evil future for Judah; and he bought a flask, after which he broke it on the ground to illustrate again the fate of Judah. Because of such words and actions, Jeremiah often found himself in trouble. Pashur, a priest, had Jeremiah beaten and placed in stocks. When released, Jeremiah told Pashur he would go into captivity and die. Despite the plots against him, Jeremiah continued to rely on the grace of Yahweh. He was brought to trial for prophesying the destruction of Jerusalem, but his defense attorneys—"certain of the elders"—pointed out that King Hezekiah had not punished the prophet Micah of Moresheth in the 8th century for similar statements.

Continuing to prophesy against the moral and religious corruption of Jerusalem during the reign of Zedekiah (597–586), Jeremiah became even more unpopular for his advocacy to surrender to Babylon.

In spite of his apparent failure to win over the people to his cause, Jeremiah inaugurated a reform that had lasting effects. He helped to bring about a change in religion from the view that primarily accepted corporate responsibility to one that held that religion is more individualistic in terms of responsibility. His words in chapter 31, verse 33, are a summation of his reform: "But this is the covenant which I will make with the house of Israel after those days, says the Lord: I will put my law within them, and I will write it upon their hearts; and I will be their God, and they shall be my people."

Ezekiel. The Book of Ezekiel, written by the prophet-priest Ezekiel, who lived both in Jerusalem prior to the Babylonian Exile (586 BCE) and in Babylon after the Exile, and also by an editor (or editors), who belongs to a "school" of the prophet similar to that of the prophet Isaiah, has captured the attention of readers for centuries

Stylistic and literary problems in Ezekiel

because of its vivid imagery and symbolism. The book has also attracted the attention of biblical scholars who have noticed that, although Ezekiel appears to be a singularly homogeneous composition displaying a unity unusual for such a large prophetic work, it also displays, upon careful analysis, the problem of repetitions, certain inconsistencies and contradictions, and questions raised by terminological differences. Though the book itself indicates that the prophecies of Ezekiel occurred from about 593–571 BCE, some scholars—who are in a minority—have argued that the book was written during widely divergent periods, such as in the 7th century and even as late as the 2nd century BCE. Most scholars, however, accept that the main body of the book came from the 6th century BCE, with the inclusion of some later glosses by redactors who remained loyal to the theological traditions of their master-teacher.

Containing several literary genres, such as oracles, mythological themes, allegory, proverbs, historical narratives, folk tales, threats and promises, and lamentations, the Book of Ezekiel may be divided into three main sections: (1) prophecies against Judah and Jerusalem (chapters 1–24); (2) prophecies against foreign countries (chapters 25–32); and (3) prophecies about Israel's future.

Ezekiel—the man and his message. The man who wrote this book—at least the main body of the work—was undoubtedly one of the leaders of Jerusalem because he was among the first group of exiles to go into captivity—those who were forced to leave their homeland about 597 BCE in a deportation to Babylon on the orders of the conquering king Nebuchadrezzar. Belonging to the priestly class, perhaps of the line of Zadok, Ezekiel was a spiritual leader of his fellow exiles at Tel-abib, which was located near the river Chebar, a canal that was part of the Euphrates River irrigation system. According to his own account, Ezekiel, the priest without a temple, received the call to become a prophet during a vision "In the thirtieth year, in the fourth month, on the fifth day"—perhaps July 31, 593 BCE, if the dating is based on the lunar calendar, though the exact meaning of "thirtieth year" remains obscure. A married man who was often consulted by elders among the exiles, Ezekiel carried out his priestly and prophetic career during two distinct periods: (1) from 593–586 BCE, a date that was doubly depressing for the prophet because it was the period when his wife died and his native city was destroyed; and (2) from 586–571 BCE, the date of his last oracle (chapter 29, verse 17).

The personality of the prophet shows through his oracles, visions, and narrations. Frustrated because the people would not heed his messages from Yahweh, Ezekiel often exhibited erratic behaviour. This need not mean that he was psychologically abnormal. Like many great spiritual leaders, he displayed qualities and actions that did not fall within the range of moderation, and to perform an ex post facto psychological postmortem examination on any great historical figure in the face of a paucity of necessary details may be an interesting game but is hardly scientifically respectable or accurate. To be sure, Ezekiel did engage in erratic behaviour: he ate a scroll on one occasion, lost his power of speech for a period of time, and lay down on the ground "playing war" to emphasize a point, an action that would certainly draw attention to him, which was his purpose. In spite of these peculiarities, Ezekiel was a master preacher who drew large crowds and a good administrator of his religious community of exiles. He held out hope for a temple in a new age in order to inspire a people in captivity. He also initiated a form of imagery and literature that was to have profound effects on both Judaism and Christianity all the way to the 20th century: apocalypticism (the view that God would intervene in history to save the believing remnant and that this intervention would be accompanied by dramatic, cataclysmic events).

Ezekiel's mystical call and symbolic acts

Prophetic themes and actions. The first section of the book (chapters 1–24) contains prophecies against Judah and Jerusalem. Ezekiel's call is recorded in chapter 1 to chapter 3, verse 15. It came in a vision of four heavenly cherubim, who appeared in a wind from the north, a cloud, and flashing fire (lightning?)—traditional symbolic elements of a theophany (manifestation of a god) in ancient Near Eastern religions. These winged hybrid throne bearers—with the faces of a man, a lion, an ox, and an eagle (which became iconographic symbols of the four Gospel writers of the New Testament)—bore the throne chariot of Yahweh. The cherubim, symbolizing intelligence, strength, and—especially—mobility, had beside them four gleaming wheels, or "a wheel within a wheel" (*i.e.,* set at right angles to each other), which further emphasized the omnimobility of the throne chariot. This vision harks back to Isaiah's mystical experience (Isaiah, chapter 6) in which that prophet envisioned the throne of the ark, which symbolized the omnipresence of the invisible Yahweh. High above the cherubim was a firmament, or crystal platform, above which was the throne of Yahweh, who—in a "likeness as if it were of a human form"—spoke to Ezekiel. The Spirit of Yahweh entered him, and he was commissioned to preach to the people of Israel a message of doom to an apostate people. The significance of this vision is that it occurred not to a priest in the holy Temple at Jerusalem but to an exiled prophet-priest in a foreign land. The God of Israel was the God of the nations. The impact of his visionary experience so overwhelmed Ezekiel that he simply sat at Tel-abib for seven days.

Commissioned by Yahweh to be "a watchman for the house of Israel," Ezekiel performed a series of symbolic acts to illustrate the impending fate of the city from which he had been banished: he placed a brick on the ground to symbolize Jerusalem's future siege, lay down on the ground, bound himself to indicate capture, ate food first cooked on fuel composed of human feces and then animal excrement, and then cut his hair and beard. Though these acts were performed in Babylon, news of them was most likely communicated to the people of Jerusalem. Just as Jeremiah had tried to repress the false hopes that the residents of Jerusalem harboured concerning the downfall of Babylon, which had been predicted by the popular nationalistic prophet Hananiah (Jeremiah, chapter 28, verses 5–17), Ezekiel attempted to quash the ill-founded aspirations of the exiles for an immediate return to Jerusalem.

In chapters 6 and 7 Ezekiel prophesies that Jerusalem's "altars shall become desolate," its people will be "scattered through the countries," and "because the land is full of bloody crimes and the city full of violence," Yahweh "will put an end to their proud might and their holy places shall be profane." In chapter 8 he attacked the people of Jerusalem for their idolatry, as manifest by the women sitting before the entrance to the north gate of the Temple of Yahweh weeping in cultic despair for the Mesopotamian fertility deity Tammuz's "annual death."

After prophesying the fall of Jerusalem in chapters 9–11 because "the guilt of the house of Israel and Judah is exceedingly great," Ezekiel performed other symbolic acts such as packing baggage for an emergency exile, digging a hole in his house to illustrate the fact that some will try to escape, and eating and drinking with trembling actions to show the future fear that the Jerusalemites will experience; he also attacked prophets who gave the people false hopes. "Woe to the foolish prophets who follow their own spirit, and have seen nothing. Your prophets have been like foxes among ruins, O Israel." He tried to underline his message of urgency by relating the problem of apostasy to similar situations in Israel's past history.

The death of Ezekiel's wife and the fall of Jerusalem

About the time that Nebuchadrezzar besieged Jerusalem, Ezekiel's wife became ill. Though Ezekiel could mourn her impending death "but not aloud" (*i.e.,* only by himself so that the people would notice his unusual reaction and thus receive the full impact of his prophetic message), he was not to mourn her death publicly. When he did not eat the "bread of mourners" the people asked him for an explanation. He told them, and it was a shattering exposure: Jerusalem would be destroyed "and your sons and daughters whom you left behind shall fall by the sword"; when this happens—in spite of their pining and groaning—they will know the meaning of Ezekiel's actions.

In order to show that Yahweh was the Lord of the whole creation and of all nations, Ezekiel issued prophecies of impending disasters that would be experienced by many neighbouring Near Eastern countries. Nations that exulted

in Judah's defeat—*i.e.,* Ammon, Moab, Edom, Philistia, and Phoenicia—would all suffer the same fate, as well as Egypt, the formerly great empire that had manoeuvred Judah into its disastrous foreign policy of opposing Babylon.

Ezekiel's oracular imagery

Oracles of hope. In the third section, chapters 33–48, Ezekiel proclaimed, in oracles that have become imprinted in theological discourse and folk songs, the hope that lies in the faith that God cares for his people and will restore them to a state of wholeness. As the good shepherd, God will feed his flock and will "seek the lost," "bring back the strayed," "bind up the crippled," and "strengthen the weak." He will also "set up over them one shepherd, my servant David, and he shall feed them." This Davidic ruler will be a *nasi* (prince), the term used for a leader of the tribal confederacy before the inauguration of the monarchy. In chapter 37, Ezekiel had a now-famous vision of the valley of dry bones, which refers not to resurrection from the dead but rather to the restoration of a scattered Covenant people into a single unity. To further emphasize the restoration of the scattered people of Yahweh, Ezekiel uttered the oracle of the two sticks joined together into one, which prophesied the re-unification of Israel and Judah as one nation. Chapters 38 and 39 contain a cryptic apocalyptic oracle about the invasion of an unidentified Gog of Magog. Who this Gog is has long been a matter of speculation; whoever he is, his chief characteristic is that he is the demonic person who leads the forces of evil in the final battle against the people of God. Gog and Magog have thus earned a position in apocalyptic literature over the centuries. Chapters 40–48 are a closing section in which Ezekiel has a vision of a restored Temple in Jerusalem with its form of worship reestablished and a restored Israel, with each of the ancient tribes receiving appropriate allotments. Ezekiel's prophecies while in exile in Babylon were to have a significant influence on the religion of Judaism as it emerged from a time of reassessment of its religious beliefs and cultic acts during the Babylonian Exile (586–538 BCE).

The Twelve. *Hosea.* The Book of Hosea, the first of the canonical Twelve (Minor) Prophets, was written by Hosea (whose name means "salvation," or "deliverance"), a prophet who lived during the last years of the age of Jeroboam II in Israel and the period of decline and ruin that followed the brief period of economic prosperity. The Assyrians were threatening the land of Israel and the people of the Covenant acted as though they were oblivious to the stipulations of their peculiar relation to Yahweh. The Book of Hosea is a collection of oracles composed and arranged by Hosea and his disciples. Like his contemporary Amos, the great prophet of social justice, Hosea was a prophet of doom; but he held out a hope to the people that the Day of Yahweh contained not just retribution but also the possibility of renewal. His message against Israel's "spirit of harlotry" was dramatically and symbolically acted out in his personal life.

The symbolism of Hosea's marriage and the emphasis on Covenant love and knowledge of God

The Book of Hosea may be divided into two sections: (1) Hosea's marriage and its symbolic meaning (chapters 1–3); and (2) judgments against an apostate Israel and hope of forgiveness and restoration (chapters 4–14).

In the first section, Hosea is commanded by Yahweh to marry a prostitute by the name of Gomer as a symbol of Israel's playing the part of a whore searching for gods other than the one true God. He is to have children by her. Three children are born in this marriage. The first, a son, is named Jezreel, to symbolize that the house of Jehu will suffer for the bloody atrocities committed in the Valley of Jezreel by the founder of the dynasty when he annihilated the house of Omri. The second, a daughter, is named Lo Ruḥama (Not pitied), to indicate that Yahweh was no longer to be patient with Israel, the northern kingdom. The third child, a son, is named Lo ʿAmmi (Not my people), signifying that Yahweh was no longer to be the God of a people who had refused to keep the Covenant. In chapter 2, Hosea voiced what probably was a divorce formula—"she is not my wife, and I am not her husband"—to indicate that he had divorced his faithless wife Gomer, who kept "going after other lovers." The deeper symbolism is that Israel had abandoned Yahweh for the cult of Baal, celebrating the "feast days of Baal." Just as Yahweh will renew his Covenant with Israel, however, Hosea buys a woman for a wife—probably Gomer. The woman may have been a sacred prostitute in a Baal shrine, a concubine, or perhaps even a slave. He confines her for a period of time so that she will not engage in any attempt to search for other paramours and thus commit further adulteries.

The second section, chapters 4–14, does not refer to the marriage motif; but the imagery and symbolism of marriage constantly recur. The Israelites, in "a spirit of harlotry," have gone astray and have left their God. Their infidelity emphasized their lack of trustworthiness and real knowledge of love, a love that could not be camouflaged by superficial worship ceremonies. Thus, Hosea emphasized two very significant theological terms: *ḥesed,* or "Covenant love," and "knowledge of God." In attacking the superficiality of much of Israel's worship, Yahweh, through Hosea, proclaimed: "For I desire steadfast (Covenant) love and not sacrifice, the knowledge of God, rather than burnt offerings." Because they have broken Yahweh's Covenant and transgressed his law, however, the Lord's anger "burns against them." For "they sow the wind and they shall reap the whirlwind." Israel will be punished for its rebellion and iniquities, but Hosea's message holds out the hope that the holiness of Yahweh's love—including both judgment and mercy—will effect a triumphant return of Israel to her true husband, Yahweh.

Joel. The Book of Joel, the second of the Twelve (Minor) Prophets, is a short work of only three chapters. The dates of Joel (whose name means "Yahweh is God") are difficult to ascertain. Some scholars believe that the work comes from the Persian period (539–331 BCE); others hold that it was written soon after the fall of Jerusalem in 586 BCE. His references to a locust plague may refer to an actual calamity that occurred; the prophet used the situation to call the people to repentance and lamentation, perhaps in connection with the festival of the New Year, the "Day of Yahweh." " 'Yet even now,' says the Lord, 'return to me with all your heart, with fasting, with weeping, and with mourning; and rend your hearts and not your garments.' " Some scholars, however, believe that the plague of locusts refers to the armies of a foreign power (Babylonia?). In the remaining section of the book (chapter 2, verse 30 to chapter 3, verse 21), Joel, in apocalyptic imagery, predicts the judgment of the nations—especially Philistia and Phoenicia—and the restoration of Judah and Jerusalem.

Amos. The Book of Amos, the third of the Twelve (Minor) Prophets, has been one of the most significant and influential books of the Bible from the time it was written (8th century BCE) down to the 20th century. Comprising only nine chapters of oracles, it was composed during the age of Jeroboam II, king of Israel from 786 to 746 BCE. His reign was marked by great economic prosperity, but the rich were getting richer and the poor poorer. Social injustice ran rampant in the land. The economically weak could find no redress in the courts and no one to champion their cause—until the coming of Amos, a shepherd from Tekoa in Judah, who also said that he was "a dresser of sycamore trees." Amos, thus, was no professional prophet nor a member of a prophetic guild.

The call for social justice by Amos

The book may be divided into three sections: (1) oracles against foreign nations and Israel (chapters 1–2); (2) oracles of indictment against Israel for her sins and injustices (chapters 3–6); and (3) visions and words of judgment (chapters 7–9). Amos was the first of the writing prophets, but his work may be composed of oracles issued both by himself and by disciples who followed his theological views.

His prophetic oracles begin with a resounding phrase: "The Lord roars from Zion." He then goes on to indict various nations—Syria, Philistia, Tyre, Ammon, and Moab—for the crimes and atrocities they have committed in times of peace: "Because they sell the righteous for silver, and the needy for a pair of shoes—they . . . trample the head of the poor into the dust of the earth, and turn aside the way of the afflicted" (chapter 2, verses 6–7).

The second section (chapters 3–6) contains some of the most vehement and cogent invectives against the social injustices perpetrated in Israel. Though the Israelites have

prided themselves on being the elect of God, they have misinterpreted this election as privilege instead of responsibility. In chapter 4, Amos, in language that was sure to raise the ire of the privileged classes, attacked unnecessary indulgence and luxury. To the wealthy women of Samaria he said: "Hear this word, you cows of Bashan, who are in the mountain of Samaria, who oppress the poor, who crush the needy, who say to their husbands, 'Bring, that we may drink!' " (chapter 4, verse 1). After a series of warnings of punishment, Amos proclaimed the coming of the day of Yahweh, which is "darkness, and not light." His attacks against superficial pretenses to worship have become proverbial: "I hate, I despise your feasts, and I take no delight in your solemn assemblies" (chapter 5, verse 21). Another verse from Amos has become a rallying cry for those searching for social justice: "But let justice roll down like waters, and righteousness like an ever-flowing stream" (chapter 5, verse 24).

The third section (chapters 7–9) contains visions of locusts as a sign of punishment, a summer drought as a sign of God's wrath, and a plumb line as a sign to test the faithfulness of Israel. The priest of the shrine at Bethel, Amaziah, resented Amos' incursion on his territory and told him to go back to his home in the south. In reply to Amaziah, Amos prophesied the bitter end of Amaziah's family. Another vision in chapter 8, that of a basket of ripe fruit, pointed to the fact that Israel's end was near. A fifth vision, depicting the collapse of the Temple in Samaria, symbolized the collapse of even the religious life of the northern kingdom. He ended his work with a prophecy that the Davidic monarchy would be restored.

Obadiah. The Book of Obadiah, the fourth book of the Twelve (Minor) Prophets, contains only 21 verses. Nothing is known about the prophet as a person or about his times. It may have been written before the Exile, though many scholars believe that it was composed either some time after 586 BCE or in the mid-5th century, when the Jews returned to the area around Jerusalem. The prophet concentrates on the judgment of God against Edom and other nations, with the final verses referring to the restoration of the Jews in their native land.

Jonah. The Book of Jonah, containing the well-known story of Jonah in the stomach of a fish for three days, is actually a narrative about a reluctant prophet. This fifth book of the Twelve (Minor) Prophets contains no oracles and is thus unique among prophetic books. In II Kings, chapter 14, verses 25–27, there is a reference to a prophet Jonah who lived during the early part of the reign of Jeroboam II (8th century BCE).

Jonah and the universal reign of Yahweh

The story, however, probably comes from a time after the fall of Jerusalem in 586 BCE. Probably living during the Exile, the author used the memory of the hated Assyrians to proclaim the mission of Israel—to teach all nations about the mercy and forgiveness of God. In the short book of four chapters, Jonah, Amittai's son, is commissioned by Yahweh to go to Nineveh, the capital of Assyria, to preach repentance. Attempting to avoid the command of Yahweh, Jonah boarded a ship, which soon was caught up in a storm. The frightened sailors drew lots to discover who was the cause of their unfortunate and calamitous condition. Jonah drew the unlucky lot and was thrown overboard, after which he was swallowed by a fish and stayed in that uncomfortable place for three days and nights. After he cried to the Lord to let him out, the fish vomited Jonah out onto dry land. Jonah, though still reluctant, went to Nineveh to preach repentance. His efforts were successful, which did not please him—because of his hatred for the Assyrians. In the end, however, Jonah realized that God was a universal God, and not the sole property of Israel.

Probably written sometime between 500 and 350 BCE (or perhaps 250 BCE), the message of Jonah protested the exclusiveness of a post-exilic Judaism, with its policy of a pure blood race of Jews that the reformers Ezra and Nehemiah had implemented in the 5th century.

Micah. The Book of Micah, the sixth book of the Twelve (Minor) Prophets, was written by the prophet Micah in the 8th century BCE. Composed of seven chapters, the book is similar in many ways to the Book of Amos. Micah attacked the corruption of those in high places and social injustice, and the book is divided into two sections: (1) judgments against Judah and Jerusalem (chapters 1–3); and (2) promises of restoration for Judah and judgments against other nations (chapters 4–7).

Micah's attacks on corruption and his summation of the prophetic message

In the first section, Micah of Moresheth utters oracles against the corrupt religious and political leaders of Israel and Judah. He also attacks the prophets who attempted to give the people false hopes: "Thus says the Lord concerning the prophets who lead my people astray, who cry 'Peace' when they have something to eat, but declare war against him who puts nothing into their mouths . . . the seers shall be disgraced, and the diviners put to shame" (chapter 3, verses 5–7). In the second section, Israel's future is predicted as being glorious, and it is told that out of Bethlehem will come a ruler of the line of David who will bring peace to the earth. Though he issues an indictment against Judah for its idolatries, Micah proclaims what is necessary to renew the Covenant relationship between God and Israel; "and what does the Lord require of you but to do justice, and to love kindness, and to walk humbly with your God?" (chapter 6, verse 8). In this verse, Micah has given a brief summation of the messages of Amos, Hosea, and Isaiah.

Nahum. The Book of Nahum, seventh of the Twelve (Minor) Prophets, contains three chapters directed against the mighty nation of Assyria. Probably written between 626–612 BCE (the date of the destruction of Nineveh, the Assyrian capital), the book celebrates in oracles, hymns, and laments the fact that Yahweh has saved Judah from potential devastation by the Assyrians.

He begins with the words "The Lord is a jealous God and avenging . . . is slow to anger and of great might, and the Lord will by no means clear the guilty" (chapter 1, verses 2–3). From that beginning he predicts the overthrow of Assyria and the devastating manner in which Nineveh will be destroyed.

Habakkuk. The Book of Habakkuk, the eighth book of the Twelve (Minor) Prophets, was written by a prophet difficult to identify. He may have been a professional prophet of the Temple from the 7th century BCE (probably between 605–597 BCE). Containing three chapters, Habakkuk combines lamentation and oracle. In the first chapter, he cries out for Yahweh to help his people: "O Lord, how long shall I cry for help, and thou wilt not hear?" (chapter 1, verse 2). Though Yahweh will send mighty nations (*e.g.,* the neo-Babylonians will be the executors of his judgment), Habakkuk wonders who will then stop these instruments of God's justice, who use great force. The answer comes in a brief, almost cryptic verse, "but the righteous shall live by his faith." The rest of chapter 2 pronounces a series of woes against those who commit social injustices and engage in debauchery. The last chapter is a hymn anticipating the deliverance to be wrought by Yahweh.

"The righteous shall live by faith"

Zephaniah. The Book of Zephaniah, the ninth book of the Twelve (Minor) Prophets, is written in three chapters. Composed by the prophet Zephaniah in the latter part of the 7th century BCE, the book is an attack against corruption of worship in Judah, probably before the great Deuteronomic reform took place. He attacked the religious syncretism that had become established, especially the worship of Baal and astral deities, and predicted the coming catastrophe of the "Day of the Lord." He denounced both foreign nations and Judah, but issued a promise of the restoration of Israel: "Sing aloud, O daughter of Zion; shout, O Israel! Rejoice and exult with all your heart, O daughter of Jerusalem" (chapter 3, verse 14). The reason for exultation is that Yahweh will deliver his people.

Haggai. The Book of Haggai, the 10th book of the Twelve (Minor) Prophets, is a brief work of only two chapters. Written about 520 BCE by the prophet Haggai, the book contains four oracles. The first oracle calls for Zerubbabel, the governor of Judaea, and Joshua, the high priest, to rebuild the Temple (chapter 1, verses 1–11). A drought and poor harvests, according to Haggai, had been caused because the returnees from the Exile had neglected or failed to rebuild the Temple. The second oracle, addressed to the political and religious leaders and the peo-

ple, sought to encourage them in their rebuilding efforts (chapter 2, verses 1–9). Apparently they were disappointed that the new Temple was not as splendid as the former one, so Haggai reassured them: "My Spirit abides among you, fear not." The third oracle was issued against the people for not acting in a holy manner (chapter 2, verses 10–19), and the fourth proclaimed that Zerubbabel would be established as the Davidic ruler (chapter 2, verses 20–23). His promise, however, remained unfulfilled.

Zechariah. The Book of Zechariah, the 11th book of the Twelve (Minor) Prophets, dates from the same period as that of Haggai—about 520 BCE. Though the book contains 14 chapters, only the first eight are oracles of the prophet; the remaining six probably came from a school of his disciples and contain various elaborations of Zechariah's eschatological themes.

The visions of Zechariah and messianic overtones

Though little is known about Zechariah's life, he probably was one of the exiles who returned to Jerusalem from Babylon. After an initial call to repentance (chapter 1, verses 1–6), Zechariah had a series of eight visions (chapter 1, verse 7 to chapter 6, verse 15). The first is of four horsemen who have patrolled the Earth to make sure that it is at rest. The second vision is of four horns (*i.e.,* nations that have conquered Israel and Judah), which will be destroyed. The third vision is of a man with a measuring line, but Jerusalem will be beyond measurement. The fourth vision shows Joshua the high priest in the heavenly court being prosecuted by Satan (the celestial adversary) and the high priest's eventual acquittal and return to his high position. The fifth vision is of a golden lampstand and an olive tree to emphasize the important positions of Joshua and Zerubbabel, which these two figures symbolize. The sixth and seventh visions—of a flying scroll and a woman of wickedness—symbolize the removal of Judah's previous sins. The eighth vision of four chariots probably refers to the anticipated messianic reign of Zerubbabel, a hope that was thwarted. Chapters 7 and 8 concern fasting and the restoration of Jerusalem.

The remaining chapters—9–14—are additions that contain messianic overtones. Chapter 9, verses 9–10, with its reference to a king riding on the foal of an ass and to a vast kingdom of peace, was used by New Testament Gospel writers in reference to Jesus' entrance into Jerusalem prior to his crucifixion. The book closes on the note of the suffering Good Shepherd, the final battle between Jerusalem and the nations and eventual victory under God, and the universal reign of Yahweh, "king over all the earth."

Malachi. The Book of Malachi, the last of the Twelve (Minor) Prophets, was written by an anonymous writer called Malachi, or "my messenger." Perhaps written from about 500–450 BCE, the book is concerned with spiritual degradation, religious perversions, social injustices, and unfaithfulness to the Covenant. Priests are condemned for failing to instruct the people on their Covenant responsibilities, idolatry is attacked, and men are castigated for deliberately forgetting their marriage vows when their wives become older. In chapter 3, the message is that Yahweh will send a messenger of the Covenant to prepare for, and announce, the day of judgment. If the people turn from their evil ways, God will bless them, and those who "feared the Lord" will be spared. The book ends with a call to remember the Covenant and with a promise to send Elijah, the 9th-century prophet who ascended into heaven in a whirlwind on a chariot, "before the great and terrible day of the Lord comes." (L.F.)

THE KETUVIM

Contents of the Ketuvim

The Ketuvim (the Writings or the Hagiographa), the third division of the Hebrew Bible, comprises a miscellaneous collection of sacred writings that were not classified in either the Torah or the Prophets. The collection is not a unified whole: it includes liturgical poetry (Psalms and Lamentations of Jeremiah), secular love poetry (Song of Solomon), wisdom literature (Proverbs, Book of Job, and Ecclesiastes), historical works (I and II Chronicles, Book of Ezra, and Book of Nehemiah), apocalyptic, or vision, literature (Book of Daniel), a short story (Book of Ruth), and a romantic tale (Book of Esther); it ranges in content from the most entirely profane book in the Bible (Song of Solomon) to perhaps the most deeply theological (Job); it varies in mood from a pessimistic view of life (Job and Ecclesiastes) to an optimistic view (Proverbs). Psalms, Proverbs, and Job constitute the principal poetic literature of the Hebrew Bible and, in many respects, represent the high point of the Hebrew Bible as literature; in fact, Job must be considered one of the great literary products of man's creative spirit.

Although portions of some of the books of the Ketuvim (*e.g.,* Psalms and Proverbs) were composed before the Babylonian Exile (586–538 BCE), the final form was post-exilic, and Daniel was not written until almost the middle of the 2nd century BCE. The books were not included in the prophetic collection because they did not fit the content or the historical-philosophical framework of that collection, because they were originally seen as purely human and not divine writings, or simply because they were written too late for inclusion. Although some of the books individually were accepted as canonical quite early, the collection of the Ketuvim as a whole, as well as some individual books within it, was not accepted as completed and canonical until well into the 2nd century CE. As noted above, there are several indications that the lapse of time between the canonization of the Prophets and of the Ketuvim was considerable; *e.g.,* the practice of entitling the entire Scriptures "the Torah and the Prophets" and the absence of a fixed name.

The needs of the Hellenistic Jews in Alexandria and elsewhere in the Greek-speaking Diaspora led to the translation of the Bible into Greek. The process began with the Torah about the middle of the 3rd century BCE and continued for several centuries. In the Greek canon, as it finally emerged, the Ketuvim was eliminated as a corpus, and the books were redistributed, together with those of the prophetic collection, according to categories of literature, giving rise to a canon with four divisions: Torah, historical writings, poetic and didactic writings, and prophetic writings. Also, the order of the books was changed, and books not included in the Hebrew Bible were added. The early Christians of both the East and West generally cited and accepted as canonical the Scriptures according to the Greek version. When Protestants produced translations based upon the Hebrew original text and excluded or separated (as Apocrypha) the books not found in the Hebrew Bible, they retained the order and the divisions of the Greek Bible. Thus the Ketuvim is not to be found as a distinct collection in the Christian Old Testament.

Order of the books of the Ketuvim

An ancient tradition, preserved in the Babylonian Talmud, prescribed the following order for the Ketuvim: Ruth, Psalms, Job, Proverbs, Ecclesiastes, Song of Solomon, Lamentations, Daniel, Esther, Ezra (which included Nehemiah), and I and II Chronicles. This sequence was chronological according to rabbinic notions of the authorship of the books. Ruth relates to the age of the judges and concludes with a genealogy of David; the Psalms were attributed, for the most part, to David; Job was assigned to the time of the Queen of Sheba, although the rabbis differed among themselves about the date of the hero; Proverbs, Ecclesiastes, and Song of Solomon were all attributed to Solomon; Lamentations, which was ascribed to Jeremiah, refers to the destruction of Jerusalem and the beginning of the Babylonian Exile; the heroes of Daniel were active until early in the reign of Cyrus II, the king of Persia who ended the exile; Esther pertains to the reign of Xerxes I, later than that of Cyrus but earlier than that of Artaxerxes I, the patron of Ezra, reputed also to have written I and II Chronicles.

Despite this tradition, however, it would appear that the sequence of the Ketuvim was not completely fixed, and there is a great variety in ordering found in manuscripts and early printed editions. The three larger books—Psalms, Job, and Proverbs—have always constituted a group, with Psalms first and the other two interchanging. The order of the five Megillot, or Scrolls (Song of Solomon, Ruth, Lamentations, Ecclesiastes, and Esther), has shown the greatest variations. The order that has crystallized has a liturgical origin; the books are read on certain festival days in Jewish places of worship and are printed in the calendar order of those occasions. Chronicles always appears

at either the beginning or the end of the corpus. Its final position is remarkable because the narrative of Ezra and Nehemiah follows that of Chronicles. The final position may have resulted from an attempt to place the books of the Hebrew Bible in a framework (Genesis and Chronicles both begin with the origin and development of the human race, and both conclude with the theme of the return to the land of Israel), but it was more probably the result of the late acceptance of Chronicles into the canon.

Psalms. The Psalms (from Greek *psalmas,* "song") are poems and hymns, dating from various periods in the history of Israel, that were assembled for use at public worship and that have continued to play a central role in the liturgy and prayer life of both Jews and Christians. Known in Hebrew as Tehillim (Songs of Praise), the Psalter (the traditional English term for the Psalms, from the Greek *psalterion,* a stringed instrument used to accompany these songs) consists of 150 poems representing expressions of faith from many generations and diverse kinds of people. These unsystematic poems epitomize the theology of the entire Hebrew Bible.

Characteristics of Hebrew poetry

Hebrew poetry has much in common with the poetry of most of the ancient Near East, particularly the Canaanite poetic literature discovered at Ras Shamra. Its main features are rhythm and parallelism. The rhythm, which is difficult to determine precisely because the proper pronunciation of ancient Hebrew is unknown, is based upon a system of stressed syllables that follows the thought structure of the poetic line. The line, or stich, is the basic verse unit, and each line of verse is normally a complete thought unit. The most common Hebrew line consists of two parts with three stresses to each part (3/3); thus:

Have-mercy-on-me,/O-God, in-your-goodness;
in-your-great-tenderness/wipe-away-my-faults.

(Ps.51:1)

Lines with three or four parts and parts with two, four, or five stresses also occur.

The lines present various kinds of parallelism of members, whereby the idea expressed in one part of a line is balanced by the idea in the other parts. The classical study on Hebrew parallelism was done by Robert Lowth, an 18th-century Anglican bishop, who distinguished three types: synonymous, antithetic, and synthetic. Synonymous parallelism involves the repetition in the second part of what has already been expressed in the first, while simply varying the words.

Yahweh, do not punish me in your rage,
or reprove me in the heat of anger.

(Ps. 38:1)

In antithetic parallelism the second part presents the same idea as the first by way of contrast or negation.

For Yahweh takes care of the way the virtuous go,
but the way of the wicked is doomed.

(Ps. 1:6)

Synthetic parallelism involves the completion or expansion of the idea of the first part in the second part.

As a doe longs for running streams,
so longs my soul for you, my God.

(Ps. 42:1)

Synthetic parallelism is a broad category that allows for many variations, one of which has the picturesque name "staircase" parallelism and consists of a series of parts or lines that build up to a conclusion.

Pay tribute to Yahweh, you sons of God,
tribute to Yahweh of glory and power,
tribute to Yahweh of the glory of his name,
worship Yahweh in his sacred court.

(Ps. 29:1-2)

Although it is evident that Hebrew poetry groups lines into larger units, the extent of this grouping and the principles on which it is based are uncertain. The acrostic poems are a notable exception to this general uncertainty.

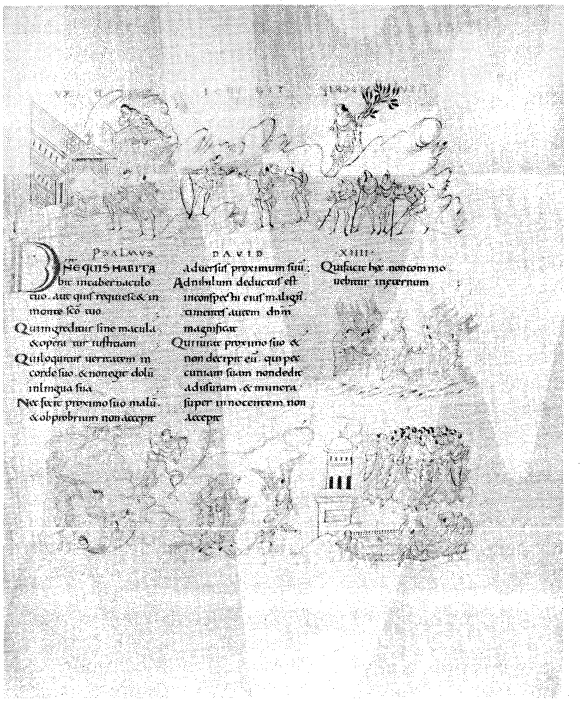
The numeration of the Psalms found in the Hebrew Bible and those versions derived from it differs from that in the Septuagint, the Vulgate, and the versions derived from them. The latter two join Psalms 9 and 10 and 114 and 115 but divide both 116 and 147 into two. The following scheme shows the differences:

Hebrew	Septuagint–Vulgate
1–8	1–8
9–10	9
11–113	10–112
114–115	113
116	114–115
117–146	116–145
147	146–147
148–150	148–150

Although Roman Catholic versions in the past have used the Septuagint–Vulgate way of numbering, recent translations have followed the Hebrew tradition.

The present form of the Psalter is the result of a lengthy literary history. It is divided into five books (Psalms 1–41; 42–72; 73–89; 90–106; and 107–150), probably in imitation of the five books of the Pentateuch. Psalm 1 serves as an introduction to the whole Psalter, while Psalm 150 is a final doxology (an expression of praise to God); the books are divided from each other by short doxologies that form the conclusions of the last psalm of each of the first four books. This division, however, appears to be artificial. There are indications, cutting across the present divisions, that the book was a compilation of existing collections. That there were several collections existing side by side is seen in the way that certain psalms (*e.g.,* Psalms 14 and 53) duplicate each other almost word for word. At some phase of the Psalter's development there must have been an Elohistic collection (Psalms 42–83) distinguished by the use of the divine name Elohim in place of Yahweh,

Compilation of existing collections



Illustrated text of Psalm 15 (Vulgate Psalm 14) from the Utrecht Psalter of Reims, 9th–12th century. In the British Museum (Harley, MS.603).

which is far more common in the rest of the psalms. There appear to be two distinct collections of psalms ascribed to David, one Yahwistic (Psalms 3–41) and the other Elohistic (Psalms 51–72). Further evidence of the book's gradual growth may be seen in the editorial gloss following Psalm 72; it purports to conclude the "prayers of David," although there are more Davidic psalms.

The superscriptions found on most of the psalms are obscure but point to the existence of earlier collections. Psalms are attributed to David, Asaph, and the sons of Korah, among others. It is generally held that Asaph and the sons of Korah indicate collections belonging to guilds of temple singers. Other possible collections include the Songs of Ascents, probably pilgrim songs in origin, the

Hallelujah Psalms, and a group of 55 psalms with a title normally taken to mean "the choirmaster."

It is evident that the process whereby these various collections were formed and then combined was extremely complex. The investigation of the process is made difficult because individual psalms and whole collections underwent constant development and adaptation. Thus, for example, private prayers became liturgical, songs of local sanctuaries were adapted to use in the Temple, and psalms that became anachronistic by reason of the fall of the monarchy or the destruction of the Temple were reworked to fit a contemporary situation. Such problems complicate the determination of the date and original occasion of the psalm.

Date and authorship

For centuries both Jews and Christians ascribed the whole Psalter to David, just as they ascribed the Pentateuch to Moses and much of the wisdom literature to Solomon. This was thought to be supported by the tradition that David was a musician, a poet, and an organizer of the liturgical cult and also by the attribution of 73 psalms to David in the superscriptions found in the Hebrew Bible. These superscriptions, however, need not refer to authorship. Moreover, it is clear that David could not have written all the psalms attributed to him because some of them presuppose the existence of the Temple in Jerusalem, which was not constructed until later. Contrary to the long-established Davidic authorship tradition, at the end of the 19th century most biblical critics spoke of a Persian date (539–333 BCE) and even of the Maccabean era (mid-2nd century BCE) for the majority of the psalms. In the 20th century the Psalter has been considered to be a collection of poems that reflect all periods of Israel's history from before the monarchy to the post-exilic restoration, and it is thought that David played a central role in the formation of the religious poetry of the Jewish people. Scholars, however, are reluctant to assign precise dates.

The most important contribution to modern scholarship on the Psalter has been the work of Hermann Gunkel, a German biblical scholar, who applied form criticism to the psalms. Form criticism is the English name for the study of the literature of the Bible that seeks to separate its literary units and classify them into types or categories (*Gattungen*) according to form and content, to trace their history, and to reconstruct the particular situation in life or setting (*Sitz im Leben*) that gave rise to the various types. This approach does not ignore the personal role of individual composers and their dates, but it recognizes that Hebrew religion, conservative in faith and practice, was more concerned with the typical than with the individual and that it expressed this concern in formal, conventional categories. The study is aided by viewing them in the context of similar literary works in the earlier or contemporary cultures of the ancient Near East.

Major types of psalms

Gunkel identified five major types of psalms, each cultic in origin. The first type is the Hymn, which is a song of praise, consisting of an invitation to praise Yahweh, an enumeration of the reasons for praise (*e.g.*, his work of creation, his steadfast love), and a conclusion in which frequently the invitation is repeated. The life setting of the hymns was generally some occasion of common worship. Two subgroups within the hymn type are the Songs of Zion, which glorify Yahweh's presence in the city of Jerusalem, and the Enthronement Songs, which—though their number, setting, and interpretation have been the subject of much debate—acclaim Yahweh's kingship over the whole world.

The second type is the Communal Lament. Its setting was some situation of national calamity, when a period of prayer, fasting, and penitence would be observed. In such psalms Yahweh is invoked, the crisis is described, Yahweh's help is sought, and confidence that the prayer has been heard is expressed.

The Royal Psalms are grouped on the basis not of literary characteristics but of content. They all have as their life setting some event in the life of the pre-exilic Israelite kings; *e.g.*, accession to the throne, marriage, departure for battle. Gunkel pointed out that in ancient Israel the king was thought to have a special relationship to Yahweh and thus played an important role in Israelite worship. With the fall of the monarchy, these psalms were adapted to different cultic purposes.

In the Individual Lament an individual worshipper cries out to Yahweh in time of need. The structure of these psalms includes: an invocation of Yahweh, the complaint, the request for help, an expression of certainty that Yahweh will hear and answer the prayer, and in many cases a vow to offer a thanksgiving sacrifice. Three aspects have been the subject of extensive study: the identity of the "enemies" who are often the reason for the complaint; the meaning of the term poor, which is frequently used to describe the worshipper; and the sudden transition in mood to certainty that the prayer has been heard. Psalms of this type form the largest group in the Psalter.

The final major type is the Individual Song of Thanksgiving, which presumably had its setting in the thanksgiving sacrifice offered after a saving experience. These psalms begin and conclude with an exclamation of praise to Yahweh. The body of the psalm contains two elements: the story of the one who has been saved and the recognition that Yahweh was the rescuer.

Gunkel also distinguished several minor types of psalms, including Wisdom Poems, Liturgies, Songs of Pilgrimage, and Communal Songs of Thanksgiving.

For Gunkel, although the types of the psalms were originally cultic, the majority of the poems in the existing Psalter were composed privately in imitation of the cultic poems and were intended for a more personal, "spiritualized" worship. Most biblical scholars since Gunkel have accepted his classifications, with perhaps some modifications, but have focussed increased attention on the setting, the *Sitz im Leben,* in which the psalms were sung. Sigmund Mowinckel, a Norwegian scholar, explained the psalms as wholly cultic both in origin and in intention. He attempted to relate more than 40 psalms to a hypothetical autumnal New Year festival at which the enthronement of Yahweh as the universal king was commemorated; the festival was associated with a similar Babylonian celebration. Artur Weiser, a German scholar, sought the cultic milieu of the Hebrew psalms especially in an annual feast of covenant renewal, which was uniquely Israelite.

Psalms is a source book for the beliefs contained in the entire Hebrew Bible. Yet, doctrines are not expounded, for this is a book of the songs of Israel that describe the way Yahweh was experienced and worshipped. Yahweh is creator and saviour; Israel is his elected people to whom he remains faithful. The enemies of this people are the enemies of Yahweh. In these songs are found the entire range of basic human feelings and attitudes before God—praise, fear, trust, thanksgiving, faith, lament, joy. The book of Psalms has thus endured as the basic prayerbook for Jews and Christians alike.

Proverbs. Proverbs is probably the oldest extant document of the Hebrew wisdom movement, of which King Solomon was the founder and patron. Wisdom literature flourished throughout the ancient Near East, with Egyptian examples dating back to before the middle of the 3rd millennium BCE. It revolved around the professional sages, or wise men, and scribes in the service of the court, and consisted primarily in maxims about the practical, intelligent way to conduct one's life and in speculations about the very worth and meaning of human life. The most common form of these wise sayings, which were intended for oral instruction especially in the schools run by the sages for the young men at the court, was the *mashal* (Hebrew: "comparison" or "parable," although frequently translated "proverb"). Typically a pithy, easily memorized aphoristic saying based on experience and universal in application, the *mashal* in its simplest and oldest form was a couplet in which a definition was given in two parallel lines related to each other either antithetically or synthetically. Verse 5 of the 15th chapter of Proverbs is an example of a simple antithetic saying:

Wisdom literature in the ancient Near East

> He who spurns his father's discipline is a fool,
> he who accepts correction is discreet.

Other forms of the *mashal,* such as parables, riddles, allegories, and ultimately full-scale compositions developed later. The word *mashal* was derived from a root that

meant "to rule," and thus a proverb was conceived as an authoritative word.

The two principal types of wisdom—one practical and utilitarian, the other speculative and frequently pessimistic—arose both within and outside Israel. Practical wisdom consisted chiefly of wise sayings that appealed to experience and offered prudential guidelines for a successful and happy life. Such wisdom is found in a collection of sayings bearing the name of Ptahhotep, a vizier to the Egyptian pharaoh about 2450 BCE, in which the sage counsels his son that the path to material success is by way of proper etiquette, strict discipline, and hard work. Although such instructions were largely materialistic and political, they were moral in character and contributed to a well-ordered society.

Speculative wisdom went beyond maxims of conduct and reflected upon the deeper problems of the value of life and of good and evil. Examples are found in ancient Egyptian and Mesopotamian texts—particularly *Ludlul bel nemeqi,* often called the "Babylonian Job"—in which sensitive poets pessimistically addressed such questions as the success of the wicked, the suffering of the innocent, and, in short, the justice of human life.

Hebrew wisdom, which owed much to that of its neighbours, appeared with the establishment of the monarchy and a royal court and found a patron in Solomon. Through the following centuries the wise men were at times the object of rebuke by the prophets, who disliked their pragmatic realism. The exile, however, brought a change in Hebrew wisdom; it became deeply religious. The wise men were convinced that religion alone possessed the key to life's highest values. It was this mood that dominated the final shaping of the Hebrew wisdom literature. Though dependent on older materials and incorporating documents from before the exile, the wisdom books in their present form were produced after the exile. In the Hebrew Bible the book of Proverbs offers the best example of practical wisdom, while Job and Ecclesiastes give expression to speculative wisdom. Some of the psalms and a few other brief passages are also representative of this type of literature. Among the Apocrypha, the Wisdom of Solomon and Ecclesiasticus are wisdom books.

Collections in the book of Proverbs

The book of Proverbs is a collection of units originally independent, some of which can be traced back to the era of Solomon. The present form of the book was the result of a long process of growth that was not completed until post-exilic times. It consists of two principal collections of early origin called "the proverbs of Solomon" and "proverbs of Solomon which the men of Hezekiah king of Judah copied." Appendixes were added to each of the collections. The whole book was preceded by a long introduction and concludes with a poem praising the ideal wife. In addition to sectional titles, changes in literary form and in subject matter help to mark off the limits of the various units, which can be ordered into nine sections.

The introduction (chapters 1–9) constitutes the youngest unit in the book. It consists of a series of poems or discourses in which a father exhorts his son to acquire wisdom and in which wisdom personified intervenes. These chapters have a more speculative quality than the remainder of the book. They do not treat wisdom simply as a human quality and achievement or as a cultural legacy imparted by teachers and parents; they present it as a universal and abiding reality, transcending the human scene. Wisdom is the first of God's works and participated with him in the creation of the world. A constantly debated aspect of this section concerns the identity of "the loose [strange] woman" who is set over against Wisdom.

The "proverbs of Solomon" (10:1–22:16) consist entirely of parallelistic couplets—the *mashal* in its primitive form. There are 375 aphorisms each complete in itself and arranged in no apparent order. The motivation of this section, in contrast to the preceding, is strongly practical: wisdom is a human achievement by means of which man's life can be fulfilled. The wise are contrasted with fools, and the just with the wicked. It is difficult, however, to establish the nature of the difference, if any, between the wicked and the fool or between the just and the wise.

The "sayings of the wise" (22:17–24:22) consist of longer units or sayings introduced by a preface. The most distinctive feature of this section is its close relationship to a piece of Egyptian writing, "The Instruction of Amenemope," which has been dated within the broad limits of 1000–600 BCE. The Hebrew author apparently used this work as a model—the Egyptian work comprises 30 chapters, and the Hebrew text refers to its "thirty sayings"—and as one of the sources in compiling his own anthology. An additional collection of four wise sayings (24:23–34) forms a supplement to the "sayings of the wise."

The second collection of "proverbs of Solomon" (chapters 25–29) consists of 128 sayings that closely resemble the earlier collection, although quatrains as well as couplets are included. The scribes of Hezekiah's court (*c.* 700 BCE) are credited with assembling this collection.

The book concludes with four independent units or collections. The "words of Agur" (30:1–14) differs sharply in spirit and substance from the rest of Proverbs; it has much closer affinities with the book of Job, stressing the inaccessibility of wisdom for man. There is no internal evidence, such as a continuous theme, to show that these 14 verses are a single unit; but in the Septuagint they stand together between the "sayings of the wise" and its supplement. The "numerical sayings" (30:15–33) contain elements of riddle and show a special interest in the wonders of nature and the habits of animals. The "instruction of Lemuel" (31:1–9) is an example of the importance of maternal advice to a ruler in the ancient Near East. Lemuel seems to have been a tribal chieftain of northwest Arabia, in the region of Edom. The final section (31:10–31) is an alphabetical poem in praise of the "perfect wife," who is celebrated for her domestic virtues.

Distinctive character of wisdom literature within Hebrew Scriptures

The wisdom movement constituted a special aspect of the religious and cultural development of ancient Israel. As the primary document of the movement, Proverbs bears a clear impress of this distinctive character, so that in many respects it presents a sharp contrast to the outlook and emphases of Israel's faith as attested in the Hebrew Scriptures generally. This contrast also marks Job and Ecclesiastes, however greatly they may differ from Proverbs in other respects.

Proverbs never refers to Israel's history. In the Hebrew Bible as a whole, this history is constantly recalled not so much for social or political reasons as to declare the faith of Israel that God has acted in its history to redeem his people and make known to them the character of his rule. The great themes of the promise to the patriarchs, the deliverance from slavery, the making of the Covenant at Mt. Sinai, the wilderness wandering, and the inheritance of Canaan were celebrated in Israel's worship to tell the story of God's revelation of himself and of his choice of Israel. None of this is alluded to in Proverbs. The implication seems to be that for Proverbs God's revelation of himself is given in the universal laws and patterns characteristic of nature, especially human nature, rather than in a special series of historical events; that is, the revelation of God is in the order of creation rather than in the order of redemption. Moreover, the meaning of this revelation is not immediately self-evident but must be discovered by men. This discovery is an educational discipline that trusts human reason and employs research, classifying and interpreting the results and bequeathing them as a legacy to future generations. The wise are those who systematically dedicate themselves to this discovery of the "way" of God.

Unlike Job and Ecclesiastes, Proverbs (with the exception of the "words of Agur") is optimistic in that it assumes that wisdom is attainable by those who seek and follow it; that is, man can discover enough about God and his law to ensure the fulfillment of his personal life. This character of God is conceived almost entirely in terms of ethical laws, and the rewards for their observance are defined in terms of human values; *e.g.,* health, long life, respect, possessions, security, and self-control.

Because God is apprehended in static terms, rather than dynamic as elsewhere in the Bible, the viewpoint of Proverbs is anthropocentric. Man's destiny depends upon his responsible action. There is no appeal to divine mercy, intervention, or forgiveness; and the divine judgment is simply the inexorable operation of the orders of life as God

has established them. Implicit in the book is an aristocratic bias. The wise constitute an elite nurtured by inheritance, training, and self-discipline; fools are those who can never catch up, because of either the determinism of birth or the wasted years of neglect. In its social and cultural attitudes, the book is probably the most conservative in the Bible: wealth and status are most important; obedience to the king and all authorities is inculcated; industry and diligence are fostered, for hunger, poverty, and slavery are the fate of the lazy; and age and accepted conventions are accorded great respect.

Job. The Book of Job is not only the finest expression of the Hebrew poetic genius; it must also be accorded a place among the greatest masterpieces of world literature. The work is grouped with Proverbs and Ecclesiastes as a product of the wisdom movement, even though it contains what might be called an anti-wisdom strain in that the hero protests vehemently against the rationalistic ethics of the sages. Yet it is the supreme example among ancient texts of speculative wisdom in which a man attempts to understand and respond to the human situation in which he exists.

The Book of Job consists of two separate portions. The bulk of the work is an extended dialogue between the hero and his friends and eventually Yahweh himself in poetic form. The poem is set within the framework of a short narrative in prose form. The book falls into five sections: a prologue (chapters 1 and 2); the dialogue between Job and his friends (3–31); the speeches of Elihu (32–37); the speeches of Yahweh and Job's reply (38–42:6); and an epilogue (42:7–17).

Prose narrative in Job

The prologue and epilogue are the prose narrative. This is probably an old folktale recounting the story of Job, an Edomite of such outstanding piety that he is mentioned by the prophet Ezekiel in conjunction with Noah and Daniel. The name Job was common in antiquity, being found in texts ranging from the 19th to the 14th century BCE. Whether the folktale is preserved in its original oral form or whether it has been retold by the poet of the dialogue is not known. The fact that an Edomite sheikh is commended by the Hebrew God, however, suggests a date before the 6th century BCE, for Jewish distrust of Edomites became intense during the exile, and the archaic language makes a date in the 8th century probable.

Job is pictured as an ideal patriarch who has been rewarded for his piety with material prosperity and happiness. The Satan (Accuser), a member of the heavenly council of Yahweh, acts with Yahweh's permission as an *agent provocateur* to test whether or not Job's piety is rooted in self-interest. Faced with the appalling loss of his worldly possessions, his children, and finally his own health, Job refuses to curse Yahweh. His capacity for trusting Yahweh's goodness has made him an unsurpassed model of patience. Three of Job's friends, whose names identify them also as Edomites, now arrive to comfort him. At this point the poetic dialogue begins. The conclusion of the tale, as given in the epilogue, describes the restoration of Job, who receives double his original possessions and lives to a ripe old age.

The dialogue between Job and his friends

The picture of Job that is presented in the poetic portion is radically different. Instead of the patient and loyal servant of Yahweh, he is an anguished and indignant sufferer, who violently protests the way Yahweh is treating him and displays a variety of moods ranging from utter despair, in which he cries out accusingly against Yahweh, to bold confidence, in which he calls for a hearing before Yahweh. Most scholars have dated this section to the 4th century BCE, but there is a growing tendency to regard it as two centuries earlier, during the period of the exile. This precise dating is based on the fact that the dialogue shows clear literary dependence on Jeremiah, whereas equally obvious connections with Deutero-Isaiah suggest the dependence of the latter on Job.

The poem opens with a heartrending soliloquy by Job in which the sufferer curses the day of his birth. The shocked friends are roused from their silence, and there follow three cycles of speeches (chapters 4–14, 15–21, and 22–27) in which the friends speak in turn. To each such speech Job makes a reply. The personalities of the friends are skillfully delineated, Eliphaz appearing as a mystic in the prophetic tradition, Bildad as a sage who looks to the authority of tradition, and Zophar as an impatient dogmatist who glibly expounds what he regards as the incomprehensible ways of God.

Eliphaz begins the first cycle by recounting a mystical vision that revealed to him the transcendence of God and the fact that all men are by nature morally frail. He suggests that suffering may be disciplinary, although this is irrelevant to Job's plight. Finally, he urges contrite submission to Yahweh. Job chides his friends for failing him in his hour of need and charges God with being his tormentor.

Bildad suggests that the fault may have lain in Job's children and reiterates Eliphaz' call to humble submission. Job then retorts that the doctrine of Yahweh's omnipotence is no answer but a serious problem, because Yahweh appears to be merely omnipotent caprice. He is convinced that if he could only meet Yahweh in open debate he would be vindicated, but he recognizes the need for an impartial third party who could intervene and protect him from Yahweh's overpowering might.

Zophar re-echoes his predecessors' views on Yahweh but goes the full length of accusing Job himself of sin and once more urges Job to a contrition that for him could only be hypocritical. Job continues to insist that Yahweh is capricious and defiantly challenges him but is bewildered when no reply is forthcoming. His longing for death as a welcome release leads him to ask whether man might not hope for a revival after death, but this daring hope is immediately rejected.

The second cycle opens with Eliphaz accusing Job of blasphemy and almost exultantly describing the fate of the wicked. In his reply Job returns to the idea of a third party to the debate. Now, however, this umpire or judge has become an advocate, a counsel for the defense. After Bildad has again elaborated on the fate of the wicked, Job states that a Vindicator, or Redeemer (Go'el), will establish his innocence. The Vindicator of this crucial but sadly corrupted passage (19:25–27) has long been identified with God himself, so that according to some scholars Job "appeals away from the God of orthodox theology to God as He must be." A few scholars, however, recognize the Vindicator as the third party (the "umpire" or "witness") of earlier chapters. It is also unclear whether this vindication will take place before or after Job's death. Then Zophar, though freely admitting that the wicked may indeed enjoy some prosperity, describes how they fall victim to inevitable nemesis. Job maintains that the wicked do not end thus but live on to an old age.

Eliphaz begins the third cycle by accusing Job at last of specific sins and again counsels Job to humble himself before Yahweh. But Job cannot find this God, who seems to be completely indifferent to him. The conclusion of the dialogue is in serious disorder, with speeches placed in Job's mouth that could only have been uttered by the friends. The final speech of Zophar, which is omitted, seems to be represented by a fragment preserved within the third reply of Job.

Chapter 28 is regarded as a later addition by most scholars, because it is hardly in place at this juncture in the dialogue, especially in the mouth of Job. It is a magnificent hymn in praise of wisdom. Chapters 29–31 contain a monologue by Job; in them occurs an adumbration of the highest moral ideal to be found in the Hebrew Bible.

The speeches of Elihu and the speeches of Yahweh

Although a few scholars have maintained that the speeches of Elihu formed part of the original work, most reject this section as a later insertion. The speeches merely reiterate the dogmas of the friends and unduly delay the appearance of Yahweh. Although the section is in poetic form, its style is different from that of the dialogue. Significantly, there is no mention of Elihu in the dialogue or anywhere else in the book, yet the Elihu speeches are familiar with the dialogue, frequently quoting verbatim from it. Chapter 32 is of interest, because it appears to contain the writer's notes and comments on the dialogue, often citing passages from it. Worthy of notice is the writer's emphasis on the disciplinary value of suffering.

The climax of the poem is reached in the speeches of

Yahweh, who appears in a majestic theophany—a whirlwind—and reveals himself to Job in three speeches interspersed with two short speeches by Job. Biblical scholars have often questioned whether this section—especially the descriptions of Behemoth (the hippopotamus) and Leviathan (the crocodile) in the second Yahweh speech—is a genuine part of the original poem, but there is no doubt that their presence at this point in the book is a dramatic triumph. Throughout these speeches Yahweh does not offer rational answers to Job's questions and accusations; he raises the discussion to a new perspective. With heavy irony Yahweh puts to Job a series of unanswerable questions about the mysteries of the universe; if, the writer is asking, Job is unable to answer the simple questions about the divine activity in the marvels of nature, how can Yahweh explain to him the deeper mystery of his dealings with men. Job's personal problem is ignored, yet he finds his answer in this direct encounter with Yahweh:

> I had heard of thee by the hearing of the ear,
> but now my eye sees thee;
> therefore I despise myself,
> and repent in dust and ashes.

Job stands in a new relationship to Yahweh, one no longer based on hearsay but the result of an act of personal faith expressed in repentance.

A few scholars, beginning in the mid-18th century, have attempted to demonstrate the influence of Greek tragedy upon the form of the book. This has not met with acceptance by most critics; its long monologues are not truly dramatic in nature. Neither is it a philosophical discussion in the style of the Platonic dialogues. It is a deeply religious poem with dramatic possibilities. It skillfully blends many genres: folktale, hymn, individual lament, prophetic oracle, and didactic poem.

The author remains quite unknown except for a few hints provided by the book itself. That he was a Jew is assumed because of his familiarity with much of the Hebrew literature. Nevertheless, the book does not have a Hebrew setting, it is pervaded with foreign elements, and it shows a special knowledge of Egypt, thus leading many to believe that he was well travelled or lived outside the Holy Land. He was a keen observer of the natural world, and his feeling for the agony of the sufferer is a compelling argument that he had known anguish.

The book touches on many subjects, such as disinterested obedience to God under testing, innocent suffering, social oppression, religious experience and pious suffering, a man's relation to God, and the nature of God. Scholars have attempted to discover the basic message of the author. Because of the greater difficulty in understanding the Job of the poetic portion, the traditional interpretation looked to the narrative and saw the message as the need for patient bearing and faith despite tribulation. When certain poetic passages were thought to point to a belief in the resurrection of the body, Job became not only a patient sufferer but also a prophet of the resurrection. This view, however, does not account for the Job of the poetic portion. Thus, in the 19th century, with the advancement of biblical criticism, scholars began to claim that the author was dealing with the problem of unmerited suffering. The book presents a deep view of suffering, and Job's experience teaches that man must rest in faith and resign himself to the incomprehensible ways of God.

It would seem, however, that the question raised by Job is both deeper and broader than the question of how to account for the infliction of physical adversity on the innocent. Job's physical suffering is the outward symbol of his intense inward agony, the agony of a man who feels himself lost in a meaningless universe and abandoned even by God. What torments Job—and the author—is the question of the justice of God and the justice and honour of man before God. His passionate pleading of his own righteousness and his calling upon God for a hearing lead him to an encounter with God. This encounter does not answer the question of why the innocent suffer, but it is the only answer to the plea of a man seeking to find his God and to justify himself to him. The complacent believer who has been shattered by suffering, doubt, and despair is confirmed in faith and repents.

The Megillot (the Scrolls). The five books known as the Megillot or Scrolls are grouped together as a unit in modern Hebrew Bibles according to the order of the annual religious festivals on which they are read in the synagogues of the Ashkenazim (central and eastern European Jews and their descendants). They did not originally form a unit and were found scattered in the Bible in their supposed historical position. In the so-called Leningrad Codex of the year 1008 CE, on which the third and subsequent editions of *Biblica Hebraica* edited by Rudolf Kittel are based, the five are grouped together but in a historical order. Nevertheless, their appearance usually follows the order of the liturgical calendar:

Song of Solomon	Pesah (pass over)	March–April
Ruth	Shavuot (Feast of Weeks	May–June
Lamentations	Tisha be-Av (Fast of Av 9)	July–August
Ecclesiastes	Sukkot (Feast of Tabernacles)	September–October
Esther	Purim (Feast of Lots)	February–March

The five books have little in common apart from their roles in the liturgy. Although the Song of Solomon and Lamentations are poetic in form and Ruth and Esther are stories of heroines, the contrast in the moods and purposes of both pairs sharply distinguishes the books. Ecclesiastes is a product of the Hebrew wisdom movement and exhibits the most pessimistic tone of any book in the Hebrew Bible.

Song of Solomon. The Song of Solomon (also called Song of Songs and Canticle of Canticles) consists of a series of love poems in which lovers describe the physical beauty and excellence of their beloved and their sexual enjoyment of each other. The Hebrew title of the book mentions Solomon as its author, but this seems improbable, primarily because of the late vocabulary of the work. Although the poems may date from an earlier period, the present form of the book is late, perhaps as late as the 3rd century BCE, and its author remains unknown.

Interpretations of the Song of Solomon

The Song of Solomon has been interpreted in different ways, four of which are noteworthy. The allegorical interpretation takes the book as an allegory of God's love for Israel or of Christ's love for the church. Such a view seems gratuitous and incompatible with the sensuous character of the poems. The dramatic interpretation is based on the dialogue form of much of the book and attempts to find a plot involving either a maiden in Solomon's court and the King or the maiden, the King, and a shepherd lover. The absence of drama in Semitic literatures and the episodic character of the book make this theory highly improbable. The cultic-mythological interpretation connects the book with the fertility cults of the ancient Near Eastern world. The condemnation in the Hebrew Bible of such rituals makes it difficult to accept this view, unless it is assumed that the original meaning of the poems was forgotten. The literal interpretation considers it to be a collection of secular love poems, without any religious implications, that may have been sung at wedding festivities. According to this commonly accepted view, the poems were received into the biblical canon despite their secular nature and their lack of mention of God because they were attributed to Solomon and because they were understood as wedding songs and marriage was ordained by God.

The reasons for the Song of Solomon being read at Passover, which celebrates the Exodus from Egypt, are not entirely clear. Possibly, they include the fact that spring is referred to in the book and that according to the allegorical interpretation the book could refer to God's love for Israel, which is so well evidenced by the events of the Exodus and especially the Covenant at Mt. Sinai.

Ruth. The Book of Ruth is a beautiful short story about a number of good people, particularly the Moabite great-grandmother of David. Though events are set in the time of the judges, linguistic and other features suggest that the present form dates from post-exilic times. But it gives the impression of being based on an ancient tradition, perhaps on written source. It was certainly grounded on a solid core of fact, for no one would have invented a Moabite ancestress for Israel's greatest king.

The book describes how, during a time of famine, Elimelech, a Bethlehemite, travelled to Moab with his wife, Naomi, and his two sons, Mahlon and Chilion. After his

death, the sons married Moabite women, and then they too died, leaving no children. There was thus no one to keep the family line alive and no one to provide for Naomi. Ruth, the widow of Mahlon, dedicated herself to the care of Naomi and insisted on returning with her to her native land and adopting her God. They arrived in Bethlehem during the harvest, and Ruth went out to work for the two women in the field of Boaz, a wealthy landowner. Naomi urged Ruth to seek marriage with Boaz because he was a kinsman of her late husband, and the firstborn son of such a marriage would count as a son of the deceased. (This resembles the levirate marriage that obliged a man to marry the widow of his deceased brother if the brother died without male issue.) Ruth crept under Boaz' cloak while he slept, and he accepted the implied proposal of marriage. After a nearer kinsman forfeited his claim to Ruth, Boaz married her and a son was born. Thus, loyal Ruth was provided with an excellent husband, the dead Mahlon with a son to keep his name alive, and Naomi with a grandson to support her in her old age.

Many purposes have been assigned to the book: to entertain, to delineate the ancestry of David, to uphold levirate marriage as a means of perpetuating a family name, to commend loyalty in family relationships, to protest the narrowness of Ezra and Nehemiah, the leaders of the post-exilic restoration in relation to marriages with non-Jews, to inculcate kindness toward converts to Judaism, to teach that a person who becomes a worshipper of Yahweh will be blessed by him, and to illustrate the providence of God in human affairs. The book may have served all these "purposes," but the author's objective cannot be determined with certainty.

Lamentations of Jeremiah. The Lamentations of Jeremiah consists of five poems (chapters) in the form of laments for Judah and Jerusalem when they were invaded and devastated by the Babylonians in 586 BCE, for the sufferings of the population, and for the poet himself during and after the catastrophe. These grief-stricken laments are intermingled with abject confessions of sin and prayers for divine compassion. The first four poems are alphabetic acrostics; the fifth is not, although like the others it has 22 stanzas, which is the number of letters in the Hebrew alphabet. The formal structure served as a mnemonic device and perhaps was meant to convey the note of wholeness, of Israel's total grief, penitence, and hope. The moving quality of these elegies has suited them for liturgical use. Besides their place in the Jewish liturgy commemorating the anniversary of the destruction of Jerusalem, the laments are employed by the Christian Church to pour out its grief over the Passion and death of Jesus Christ.

Dating the composition of Lamentations

Most critics place the composition of the book before the return of the Jews from exile in 537/536 BCE. Certain passages appear to be word pictures by an eyewitness and would, therefore, have been written shortly after the destruction of Jerusalem. Until the 18th century, the work was universally ascribed to the prophet Jeremiah, and this was supported by a prologue found in the Septuagint and in some manuscripts of the Vulgate. Since that time, however, many scholars have rejected the attribution to Jeremiah chiefly because the ideas and sentiments expressed in Lamentations are unlike those in Jeremiah. Moreover, it is unlikely that the spontaneity and naturalness so characteristic of Jeremiah's utterances could be accommodated to a poetic form as complicated and artificial as that in Lamentations. It is probable that the laments were the product of more than one poet.

Ecclesiastes. The book of Ecclesiastes is a work of the Hebrew wisdom movement, associated by its title and by tradition with King Solomon. It is evident, however, that the book is of much later composition; the author may have identified himself with the famous king and wise man of the past to give greater authority to his work. The language of the book, including the relatively large number of Aramaic forms, and its content point to a date in the early Greek period (later 4th or early 3rd century BCE). That the book was written prior to the 2nd century BCE, however, is shown by its influence on Ecclesiasticus, which was written early in that century, and its appearance among the manuscripts discovered at Khirbat Qumrān, on the northwestern shore of the Dead Sea, where a Jewish community existed in the mid-2nd century.

The name Ecclesiastes is a transliteration of the Greek word used in the Septuagint to translate the Hebrew Qohelet, a word connected with the noun *qahal* ("assembly"). Qohelet seems to mean the one who gathers or teaches an assembly; the author used the word as a pseudonym. He appears to be a wisdom teacher writing late in life expressing skeptical personal reflections in a collection of popular maxims of the day and longer compositions of his own. The book has been described as a sage's notebook of random observations about life. Some interpreters have questioned the unity of authorship, but, given the notebook character of the work, there seems to be little need for questioning its basic integrity.

Although the phrase "vanity of vanities! all is vanity" stressed at both the beginning and the end of the book sums up its theme, it does not convey the variety of tests that the skeptical Qohelet applies to life. He examines everything—material things, wisdom, toil, riches—and finds them unable to give meaning to life. He repeatedly returns to life's uncertainties, to the hidden and incomprehensible ways of God, and to the stark and final fact of death. The only conclusion to this human condition is to accept gratefully the small day-to-day pleasures that God gives to man.

Qohelet stands in sharp contrast to the conventional wisdom schools. He recognizes the relative value of wisdom as against foolishness, but he rejects the oversimplified and optimistic view of wisdom as security for life. He offers a religious skepticism that rejects all facile answers to life's mysteries and God's ways.

Book of Esther. The Book of Esther is a romantic and patriotic tale, perhaps with some historical basis but with so little religious purpose that God, in fact, is not mentioned in it. The book may have been included in the Hebrew canon only for the sake of sanctioning the celebrations of the festival Purim, the Feast of Lots. There is considerable evidence that the stories related in Esther actually originated among Gentiles (Persian and Babylonian) rather than among the Jews. There is also reason to believe that the version given in the Septuagint goes back to older sources than the version given in the Hebrew Bible.

Origin of Esther

Laying the scene at Susa, a residential city of the Persian kings, the book narrates that Haman, the vizier and favourite of King Ahasuerus (Xerxes I; reigned 486–465 BCE), determined by lot that the 13th of Adar was the day on which the Jews living in the Persian Empire were to be slain. Esther, a beautiful Jewess whom the King had chosen as queen after repudiating Queen Vashti, and her cousin and foster father Mordecai were able to frustrate Haman's plans. Haman then schemed to have Mordecai hanged; instead, he was sent to the gallows erected for Mordecai, and Jews throughout the empire were given permission to defend themselves on the day set for their extermination. The governors of the provinces learned in time that Mordecai, who had saved the King from being assassinated by two discontented courtiers, had succeeded to Haman's position as vizier; thus, they supported the Jews in the fight against their enemies.

In the provinces, the Jews celebrated their victory on the following day, but at Susa, where, at Esther's request, the King permitted them to continue to fight on the 14th of Adar, they rested and celebrated their success a day later. Therefore, Esther and Mordecai issued a decree obligating the Jews henceforth to commemorate these events on both the 14th and 15th of Adar.

Theme and language characterize Esther as one of the latest books of the Hebrew Bible, probably dating from the 2nd century BCE. Nothing is known of its author. According to the postbiblical sources, its inclusion in the canon, as well as the observance of the feast of 14th and 15th Adar, still met with strong opposition on the part of the Jewish authorities in Jerusalem as late as the 3rd century CE; yet, despite its lack of specific religious content, the story has become in popular Jewish understanding a magnificent message that the providence of God will preserve his people from annihilation.

Daniel. The Book of Daniel presents a collection of

popular stories about Daniel, a loyal Jew, and the record of visions granted to him, with the Babylonian Exile of the 6th century BCE as their background. The book, however, was written in a later time of national crisis—when the Jews were suffering severe persecution under Antiochus IV Epiphanes (reigned 175–164 BCE), the second Seleucid ruler of Palestine.

The exiled Jews had been permitted to return to their homeland by Cyrus II the Great, master of the Medes and Persians, who captured Babylon in 539 BCE from its last king, Nabonidus, and his son Belshazzar. The ancient Near East was then ruled by the Persians until Alexander the Great brought it under his control in 331. After Alexander's death in 323, his empire was divided among his generals, with Palestine coming under the dominion of the Ptolemies until 198, when the Seleucids won control. Under the Persian and Ptolemaic rulers the Jews seem to have enjoyed some political autonomy and complete religious liberty. But under Antiochus IV Jewish fortunes changed dramatically. In his effort to Hellenize the Jews of Palestine, Antiochus attempted to force them to abandon their religion and practice the common pagan worship of his realm. Increasingly sterner restrictions were imposed upon the Jews, the city of Jerusalem was pillaged, and, finally, in December 167 the Temple was desecrated. The outcome of this persecution was the open rebellion among the Jews, as described in the books of Maccabees. This period of Hellenistic Judaism is treated more fully in JUDAISM.

The conflict between the religion of the Jews and the paganism of their foreign rulers is also the basic theme of the Book of Daniel. In Daniel, however, it is regarded as foreseen and permitted by God to show the superiority of Hebrew wisdom over pagan wisdom and to demonstrate that the God of Israel will triumph over all earthly kings and will rescue his faithful ones from their persecutors.

Apocalyptic literature

To develop this theme the author makes use of a literary and theological form known as apocalypse (from the Greek *apokalypsis,* "revelation" or "unveiling"), which was widely diffused in Judaism and then in Christianity from 200 BCE to 200 CE. Apocalyptic literature professes to be a revelation of future events, particularly the time and manner of the coming of the final age when the powers of evil will be routed in bloody combat and God's kingdom will be established. This revelation usually occurs as a vision expressed in complicated, often bizarre symbolism. The literature is generally pseudonymous, proposed under the name of some authoritative figure of the distant past, such as Daniel, Moses, Enoch, or Ezra. This allows the author to present events that are past history to him as prophecies of future happenings.

The Book of Daniel, the first of the apocalyptic writings, did not represent an entirely new type of literature. Apocalypse had its beginnings in passages in the works of the prophets. In fact, it has been said that the apocalyptic was really an attempt to rationalize and systematize the predictive side of prophecy. There were significant differences, however. The prophet, for the most part, declared his message by word of mouth, which might subsequently be recorded in writing. The apocalyptist, on the other hand, remained completely hidden behind his message, which he wrote down for the faithful to read. The prophets normally spoke in their own name a message for their own day. The apocalyptists normally wrote in the name of some notable man of the past a message for the time of the age to come.

Like the prophets before them, the apocalyptists saw in the working out of history, which they divided into well-defined periods, a purpose and a goal. The evil in the world might lead men to despair, but the predetermined purpose of God could not be frustrated. A future age of righteousness would replace the present age of ungodliness, and God's purpose would at last be fulfilled. This literature, then, is a mixture of pessimism—times would become worse and worse, and God would destroy this present evil world—and of optimism—out of turmoil and confusion God would bring in his kingdom, the goal of history.

For many centuries the apocalyptic character of the Book of Daniel was overlooked, and it was generally considered to be true history, containing genuine prophecy. In fact, the book was included among the prophetic books in the Greek canon. It is now recognized, however, that the writer's knowledge of the exilic times was sketchy and inaccurate. His date for the fall of Jerusalem, for example, is wrong; Belshazzar is represented as the son of Nebuchadrezzar and the last king of Babylon, whereas he was actually the son of Nabonidus and, though a powerful figure, was never king; Darius the Mede, a fictitious character perhaps confused with Darius I of Persia, is made the successor of Belshazzar instead of Cyrus. By contrast, the book is a not inconsiderable historical source for the Greek period. It refers to the desecration of the Temple in 167 and possibly to the beginning of the Maccabean revolt. Only when the narrative reaches the latter part of the reign of Antiochus do notable inaccuracies appear—an indication of a transition from history to prediction. The book is thus dated between 167 and 164 BCE.

Date, hero, and languages of Daniel

Other considerations that point to this 2nd-century date are the omission of the book from the prophetic portion of the Hebrew canon, the absence of Daniel's name in the list of Israel's great men in Ecclesiasticus, the book's linguistic characteristics, and its religious thought, especially the belief in the resurrection of the dead with consequent rewards and punishments.

The name Daniel would appear to refer to a legendary hero who was used in different ways at different times and who became particularly popular in the storytelling of the Persian and Greek Diaspora as a personification of the practical and theological problems faced by the Jews in that environment. Whether there is any connection between the Daniel of this book and the one mentioned as a wise man without equal in the Book of Ezekiel and as a righteous man in the tale of Aqhat, a Ugaritic text dated from about the middle of the 14th century, is uncertain.

The book is written in two languages: the beginning (1:1–2:4a) and the final chapters (8–12) in Hebrew and the rest in Aramaic. This offers no proof of multiple authorship, however, because the linguistic divisions do not correspond to the division by literary form: chapters 1–6 are stories of Daniel and his friends in exile, and chapters 7–12 are Daniel's apocalyptic visions. Furthermore, there is a singleness of religious outlook, spirit, and purpose throughout. Nevertheless, the problem of the languages has never been satisfactorily answered.

The stories and visions of Daniel

The stories of the first six chapters, which probably existed in oral tradition before the author set them down, begin with the account of how Daniel and his three companions (Hananiah, Mishael, and Azariah, who were given the names Shadrach, Meshach, and Abednego by the Babylonians) came to be living at the Babylonian court and how they remained faithful to the laws of their religion. This is followed by five dramatic episodes calculated to demonstrate the wisdom and might of Israel's God and the unconquerable steadfastness of his loyal people. Thus, through God's gift of wisdom, Daniel excels the professional sages of the pagan court by revealing and interpreting Nebuchadrezzar's dream of a great image, made of four metals, which was shattered by a stone cut without human hand, and then the King's further dream of a tree reduced to a stump, which presaged the punishment of his arrogance by madness, and, finally, the writing on the wall, which spelled Belshazzar's doom at his sacrilegious feast. By trust in God, Daniel's companions, who refused to worship Nebuchadrezzar's golden idol, are miraculously delivered from a fiery furnace, and Daniel himself, thrown into a den of lions for holding fast to his tradition of prayer, is divinely protected.

The last six chapters of the book are apocalyptic. In chapter 7 Daniel is granted a vision of four beasts from the abyss, which are brought under divine judgment, and of "one like a son of man," who is brought before God to be invested with his universal and everlasting sovereignty. The mythological beasts are interpreted as four empires (the Babylonian Empire, the kingdom of the Medes, the Persian Empire, and the empire of Alexander) and the manlike figure as Israel. The vision of a battle between the ram (Medes and Persians) and the goat (the Greek

Empire) in chapter 8 introduces the iniquities of Antiochus IV Epiphanes and is an assurance to the stricken Jews that the end of their tribulation is near. In chapter 9 the author reinterprets the prophecy of Jeremiah that Jerusalem's desolation would end after 70 years. By making these 70 years mean 70 "weeks of years" (*i.e.,* 490 years), the author is again able to focus attention on the period of Antiochus' persecution in the 2nd century and on the imminence of his determined doom. A precise understanding of the author's scheme is not possible, however, because 490 years calculated from the beginning of the exile extends far beyond the time of Antiochus. The remaining chapters provide the fourth commentary on the crisis provoked by the Seleucid tyrant. The greater part of this vision is a sketch of the events that affected the Jews from the Persian period to the time of Antiochus and prepared for his reign of terror. After chapter 11, verse 39, the account of Antiochus' life ceases to correspond with historical fact; an inaccurate prediction of his end is the prelude to the announcement of the end of Israel's tribulation and the inauguration of God's kingdom.

The purpose of the whole book, stories and visions alike, is to encourage Israel to endure under the threat of annihilation and to strengthen its faith that "the Most High rules the kingdom of men" and will in the end give victory to his people and establish his kingdom.

The Chronicler's History

Ezra, Nehemiah, and Chronicles. The final books of the Hebrew Bible are the books of Chronicles and Ezra–Nehemiah, which once formed a unitary history of Israel from Adam to the 4th century BCE, written by an anonymous Chronicler. That these books constituted a single work—referred to as the Chronicler's history, in distinction to the Deuteronomic history and the elements of history from the priestly code of the Torah—appears evident because the same language, style, and fundamental ideas are found throughout and because the concluding verses of II Chronicles are repeated at the beginning of Ezra. The purpose of this history seems to have been to trace the origin of the Temple and to show the antiquity and authenticity of its cult and of the formal, legalistic type of religion that dominated later Judaism.

The history that these books record has already been treated in the historical section of this article and is found in greater detail in JUDAISM. The concern in this section will be chiefly with the literary and theological aspects of the books, but their contents can be summarized. In I and II Chronicles the author repeats much of the material from earlier historical books, concentrating upon the history of the kingdom of Judah. The First Book of the Chronicles begins with an extensive genealogy of Israel from Adam to the restoration but is primarily a biography of David that adds further facts to the story as given in Samuel. The Second Book of the Chronicles begins with Solomon and goes through the division of the kingdom to the reign of Zedekiah; once again the Chronicler had access to materials that supplemented the account in I and II Kings. In the Book of Ezra he describes the return of the Jews from the Babylonian Exile and the reconstruction of the Temple. He includes lists of the families who returned and the texts of the decrees under which they returned. In the Book of Nehemiah the reconstruction of the city walls of Jerusalem becomes the basis for a meditation upon the relation between God and his people. This book, too, contains lists of those who participated in the reconstruction, but much of it concentrates upon the description of Nehemiah and his persistence in performing his assignment.

Divisions, Author, Date, and Sources

The fourfold division of the books derives from the Greek and Latin versions; the more basic twofold division into Chronicles and Ezra–Nehemiah is more complex. This original division apparently resulted from the inclusion of the material known as Ezra–Nehemiah in the Hebrew canon before that known as Chronicles because it contained fresh information not found in any other canonical book. When Chronicles was later admitted to the canon, it was placed in order after Ezra–Nehemiah; although the book has retained this position in the Hebrew Bible, the Greek version restored it to its proper sequence. That Chronicles was thus "left aside" may account for the choice of *Paraleipomena* ("Things Omitted") as the Greek title of the book, but the usual and perhaps correct explanation is that Chronicles contains stories, speeches, and observations that were omitted from the parallel accounts in earlier books.

Jewish tradition has identified Ezra as the author of these books, and some modern scholars concur. According to many critics, however, the Chronicler was a Levite cantor in Jerusalem. This position is supported by the author's concern with the Levites and cultic musicians. The date of the work is more difficult to pinpoint. In its final form it has to be later than Ezra, who came to Judah about 400 BCE. An indication of the latest date at which the entire work could have been completed is its silence about the Hellenizing of Judaism that took place after Alexander the Great. This, together with language considerations that point to the late Persian period, has led the majority of commentators to postulate a 4th-century date. Some scholars, however, claim that a time before 300 BCE would be too short to account for the genealogy at the beginning of I Chronicles, which is carried down to the eighth generation after Zerubbabel, one of the leaders of the band that returned from Babylon. Thus, they push the final date to about 200 BCE or even slightly later. It is possible that the 4th-century work of the Chronicler went through a series of minor additions and adaptations until sometime early in the 2nd century, when it reached its final form.

The Chronicler had numerous historical sources—both biblical and extrabiblical—at his disposal. He was closely dependent on the books of Samuel and Kings for all of Chronicles except the first nine chapters. Sometimes he even repeated the actual words of his model, though slight textual variations suggest to some that the Hebrew copy he had before him differed a little from that of the canon and corresponded to that which lay behind the Septuagint. But he was also able to consult the final version of the Torah and the whole of the Deuteronomic history. His use of the personal memoirs of Nehemiah is undisputed; the nature of his Ezra source is less clear, but some have regarded a portion of narrative written in the first person as an autobiographical source. He included many lists, genealogies, census reports, and other official documents that may have been preserved as Temple records. The text refers by name to certain documents representing royal histories and prophetic writings about which, as they have not survived, only speculation is possible.

Theological purpose

The Chronicler made use of all these sources, but he was not shackled by them. Although his work has won increasing respect as a historical document, especially as an indispensable source for the restoration period, his purpose was chiefly theological, not historical. He was convinced of the definitiveness of the divine covenant with David. The holy community that was brought into existence by this covenant, that has been maintained by God through the vicissitudes of history, and that has its worship centred on the Temple in Jerusalem is the true kingdom of God. It is the true Israel and is the Chronicler's only concern. Thus, he mentions the northern kingdom and the kings of Israel only to the extent that they figure in the events of Judah. Loyalty to the Davidic line of succession, to Jerusalem, and to the Temple worship were the central elements in the life of God's people according to this writer. All success and failure were the result of such loyalty or disloyalty. Thus, if a king's reign was long and successful, the Chronicler saw it as the reward of God for a life led in obedience to his will; conversely, a king suffered misfortune only if he had sinned. Significantly, the Chronicler devotes much attention to David's part in the development of the liturgy, especially the organization and functions of the Levites, and omits important but uncomplimentary stories about the King that are found in the Deuteronomic history.

In short, the Chronicler traced the reformed liturgy of his day back to David and laid a solid foundation for the acceptance and conservation of the religious community that he envisioned—a devout community that worshipped joyfully in the Temple with sacrifice and praise and obeyed the Law of Moses. He knew well that the realization of that community in his day was not perfect and that the future had something better in store, but he seems to

have been content to accept the existing Davidic leaders in order not to abandon the dynastic hope because of their shortcomings. These books thus provided an apologia for orthodox Judaism (perhaps in the face of opposition from the Samaritans, the inhabitants of the former northern kingdom), and they offer to the modern reader some insight into the post-exilic community in Jerusalem, withdrawn into itself and trying to justify, explain, and preserve its existence and its spirituality. (R.F.)

Intertestamental literature

NATURE AND SIGNIFICANCE

Definitions. A vast amount of Jewish literature written in the intertestamental period (mainly 2nd and 1st centuries BCE) and from the 1st and 2nd centuries CE was preserved, for the most part, through various Christian churches. A part of this literature is today commonly called the Apocrypha (Hidden; hence, secret books; singular apocryphon). At one time in the early church this was one of the terms for books not regarded by the church as canonical (scripturally acceptable), but in modern usage the Apocrypha is the term for those Jewish books that are called in the Roman Catholic Church deuterocanonical works—*i.e.*, those that are canonical for Catholics but are not a part of the Jewish Bible. (These works are also regarded as canonical in the Eastern Orthodox churches.) When the Protestant churches returned to the Jewish canon (Hebrew Old Testament) during the Reformation period (16th century), the Catholic deuterocanonical works became for the Protestants "apocryphal"—*i.e.*, non-canonical.

Meaning of Apocrypha and Pseudepigrapha among Roman Catholics, Protestants, and Jews

In 19th-century biblical scholarship a new term was coined for those ancient Jewish works that were not accepted as canonical by either the Catholic or Protestant churches; such books are now commonly called Pseudepigrapha (Falsely Inscribed; singular pseudepigraphon), *i.e.*, books wrongly ascribed to a biblical author. The term Pseudepigrapha, however, is not an especially well suited one, not only because the pseudepigraphic character is not restricted to the Pseudepigrapha alone—and, indeed, not even all Pseudepigrapha are ascribed to any author, since there are among them anonymous treatises—but also because the group of writings so designated by this name necessarily varies in the different modern collections. Theoretically, the name Pseudepigrapha can designate all ancient Jewish writings that are not canonical in the Catholic Church. The writings of the philosopher Philo of Alexandria (1st century BCE–1st century CE) and the historian Josephus (1st century CE) and fragments of other postbiblical Hellenistic Jewish historians and poets, however, usually are excluded. Rabbinic literature (2nd century BCE–2nd century CE) also is generally excluded; such literature existed for centuries only in oral form. The edition of the Pseudepigrapha edited by the British biblical scholar R.H. Charles in 1913, however, contains a translation of *Pirqe Avot* ("Sayings of the Fathers"), an ethical tractate from the Mishna (a collection of oral laws), and even the non-Jewish *Story of Aḥikar* (a folklore hero), though other genuine Jewish writings from antiquity are omitted. Some of the Jewish Pseudepigrapha were discovered only in the last two centuries, and the Dead Sea Scrolls (the first of them discovered in the 1940s), most of which belong to this category, are not yet all published. Thus, in the broader meaning of the terms, the Apocrypha and Pseudepigrapha are a bloc of Jewish literature written in antiquity from the later Persian period (*c.* 4th century BCE) and not canonized by the Jews.

Texts and versions. A small portion of this literature is preserved in the original languages: Hebrew, Aramaic, and Greek. Most of the Hebrew or Aramaic works, however, exist today only in various translations: Greek, Latin, Syriac, Ethiopian, Coptic, Old Slavonic, Armenian, and Romanian. All the works of the Apocrypha are preserved in Greek, because they have for the Greek Church a canonical value. Those books not considered canonical by the early church have often fallen into oblivion, and their Greek text was often lost; many of the ancient Jewish Pseudepigrapha are today preserved only in fragments or quotations in various languages, and sometimes only their titles are known from old lists of books that were rejected by the church.

Of this literature only the Apocrypha (contained in Latin and Greek Bibles) were read in the liturgical services of the church. The Pseudepigrapha, in their various versions, were in most cases nearly forgotten; and manuscripts of most of them were rediscovered only in modern times, a process that continues. The discovery of the Dead Sea Scrolls at Qumrān in the Judaean desert not only furnished new texts and fragments of unknown and already known Pseudepigrapha but also contributed solutions to problems concerning the origin of other Jewish religious writings (including some Old Testament books), the connection between them, and even their composition and redaction from older sources. The new original texts also strengthened interest in the Jewish literature of the intertestamental period because of its importance for the study of both ancient Judaism and early Christianity. As a result of such discoveries, better critical editions of the Apocrypha and Pseudepigrapha have been published, as well as new studies of their content.

Rediscoveries and new discoveries of texts

The Apocrypha, whose texts originated mostly before the rise of Christianity, were regarded as canonical in the early church but contain no Christian interpolations. Many of the Pseudepigrapha, however, were interpolated by Christian writers. The nature and the extent of these Christian interpolations is often difficult to define since a Christian interpolator not only changes the text according to his Christian views or introduces specific Christian terminology, but he also may introduce in a Jewish text ideas, motifs, or terminology that are common to both Judaism and Christianity. For these reasons it is sometimes difficult to decide if a passage in a pseudepigraphon, or even sometimes the whole work, is Jewish or Christian.

Persian and Hellenistic influences. Some of the Apocrypha (*e.g.*, Judith, Tobit) may have been written already in the Persian period (6th–4th century BCE), but with these possible exceptions, all the Apocrypha and Pseudepigrapha were written in the Hellenistic period (*c.* 300 BC–*c.* AD 300). Yet the influence of Persian culture and religion sometimes can be detected even in comparatively late Jewish works, especially in Jewish apocalyptic literature (see below *Apocalypticism*). The Persian influence was facilitated by the fact that both the Jewish and Persian religions are iconoclastic (against the veneration or worship of images) and opposed to paganism and display an interest in eschatology (doctrines of last times).

Although such an affinity did not exist between Judaism and Hellenistic culture, literary activity among Hellenistic Jews was generally Greek in character: the Greek-writing Jewish authors thought mainly in Greek concepts, used genuine Greek terminology, and wrote many of their works in Greek literary forms.

Though Hellenistic Jewish authors sometimes imitated biblical forms, they learned such forms from their Greek Bible (the Septuagint). Many Greek products written by Jews served as religious propaganda and probably influenced many pagans to become proselytes, or at least to abandon their heathen faith and to become "God fearing." Thus, the Jewish literature written in Greek could be later used by Christianity for similar purposes.

Greek influence

Greek influence on Jewish writings written in Hebrew or Aramaic in Palestine in the intertestamental period was by no means as significant as upon Jewish works written in Greek among the Hellenistic Diaspora (Jews living outside Palestine). In Palestine, religion and culture formed a unity, and the Hellenization of the upper classes in Jerusalem before the Maccabean wars (168–142 BCE) was restricted to some families who had accepted Greek civilization for practical purposes. Jews in Palestine developed a flourishing autonomous culture based upon religious ideals. Living without interruption in their powerful religious tradition and with their own non-Greek education, the Palestinian Jews were able to produce literary works without significant evidences of Greek influence. The language of this literature was both Aramaic and Hebrew. Under the national revival in the Maccabean period, Hebrew became prevalent as the language of Jewish literature in Palestine; but since Aramaic was a spoken language in Palestine

during the whole period, some of the extant literary works of Palestinian Jews in the Maccabean and Roman period probably were originally written also in Aramaic.

Apocalypticism. In intertestamental Jewish literature a special trend developed: namely, apocalypticism. *Apokalypsis* is a Greek term meaning "revelation of divine mysteries," both about the nature of God and about the last days (eschatology). Apocalyptic writings were composed both in Judaism and Christianity; one of them (the Book of Daniel) was accepted in the Jewish canon and another (the book of Revelation) in the New Testament. Other apocalypses form a part of the Pseudepigrapha, and influences of apocalypticism or similar approaches are found in some of the Apocrypha. The sectarian Dead Sea Scrolls are the works of an apocalyptic movement, though not all are written in the style of apocalypses. *The Sibylline Oracles* are, in their Jewish passages, a part of Jewish Hellenistic literature; inasmuch as they contain eschatological prophecies of future doom and salvation, they are apocalyptic, but in their polemics against idolatry and their apology for Jewish faith, they are a product of Jewish Hellenistic propagandistic literature. Because one of the central themes of apocalypticism is that of future salvation, messianic hopes involving the advent of a deliverer are usually the object of intertestamental Jewish apocalypticism.

APOCRYPHAL WRITINGS

Apocryphal works indicating Persian influence. *Esdras.* The "Greek Ezra," sometimes named I (or II or III) Esdras, enjoyed considerable popularity in the early church but lost its prestige in the Middle Ages in the Latin Chruch. At the reforming Council of Trent (1545–63), the Roman Catholic Church no longer recognized it as canonical and relegated it in the Latin Bible to the end, as an appendix to the New Testament. One of the reasons for its noncanonicity in the West is that the "Greek Ezra" contains parallel material to the biblical books of Chronicles, Ezra, and Nehemiah but differs in textual recension (points of critical revision) and occasionally in the order of the stories. The content of the book is a history of the Jews from the celebration of the Passover in the time of King Josiah (7th century BCE) to the reading of the Law in the time of Ezra (5th century BCE). Though written in an idiomatic Greek, "Greek Ezra" is probably a Greek translation from an unknown Hebrew and Aramaic redaction of the materials contained in the biblical books of Chronicles, Ezra, and Nehemiah. An important part of this book (3:1–5:6), the story of the three youths at the court of Darius, has no parallel in the canonical books. This story concerns a debate between three guardsmen before Darius, king of Persia, about the question of what they consider to be the strongest of all things; the first youth asserts that it is wine, the second says that it is the king, and the third, who is identified with the biblical Zerubbabel (a prince of Davidic lineage who became governor of Judah under Darius), expresses his opinion that "women are strongest, but truth is victor over all things." He is acclaimed as the victor, and, as a reward, he requests that Darius rebuild Jerusalem and its Temple. The story evidently was written in two stages: originally, the competition was about wine, the king, and women, but later, truth was added. Truth is one of the central concepts of Persian religion and the competition itself is before a Persian king; thus it seems likely that the story is Persian in origin and that it became Jewish by the identification of the third youth with Zerubbabel.

The three youths at the court of Darius

Judith. The book of Judith is similar to the biblical Book of Esther in that it also describes how a woman saved her people from impending massacre by her cunning and daring. The name of the heroine occurs already in Gen. 26:34 as a Gentile wife of Esau, but in the book of Judith it evidently has symbolic value. Judith is an exemplary Jewess. Her deed is probably invented under the influence of the account of the 12th-century-BCE Kenite woman Jael (Judg. 5:24–27), who killed the Canaanite general Sisera by driving a tent peg through his head.

The story is clearly fiction, and the anachronisms in it are intentional: they show that the story itself is a mere fiction. The book speaks about the victory of Nebuchadnezzar, "who reigned over the Assyrians at Nineveh" (the name is of the 7th–6th century BCE king of Babylon, Nebuchadrezzar) in the time of an unknown Arphaxad, king of the Medes. Since the western nations of Nebuchadnezzar's empire had refused to come to his aid, the King ordered his commander in chief, Holofernes (a Persian name), to force submission upon the rebellious nations. In subduing these nations Holofernes destroyed their sanctuaries and proclaimed that Nebuchadnezzar alone should henceforth be worshipped as a god. Thus, the Jews, who had recently returned from the Babylonian Captivity (6th century BCE) and rebuilt the Temple, were compelled to prepare for war. Holofernes laid siege to Bethulia (otherwise unknown), described as an important strategic point on the way to Jerusalem. Because of a long siege, the inhabitants wanted to surrender their city, but Judith persuaded the people to delay the surrender for five days. Judith was a virtuous, pious, and beautiful widow. She removed her mourning garments, left the city, entered Holofernes' camp, and was brought before him. On the fourth day, Holofernes decided to seduce Judith and invited her to come into his tent; he then drank more wine than ever before. After he fell into a drunken stupor, Judith cut off his head with his sword and returned with the head to Bethulia. The Jews put Holofernes' head outside the city wall, and the following morning, upon learning of the death of their commander in chief, the Assyrian soldiers dispersed and were pursued by the Jews of Bethulia, who took abundant spoil. The Jews were not threatened again during Judith's lifetime—she lived to be 105—or for long thereafter.

The killing of Holofernes

Many suggestions have been made about the book of Judith's date of composition. Though current scholarly opinion is that the book was written in the warlike patriotic atmosphere of the early Maccabean period (*c.* 150 BCE) by a Palestinian Jew, there are no Maccabean elements in the book. It shows no direct or indirect Greek influences, the deification of kings existed already in the ancient Near East, and the political situation described in the book has nothing in common with the Maccabean period. All the apparently intentional historical mistakes, however, can be understood if it is suggested that the book of Judith was written under Persian rule. Holofernes is, as noted above, a typical Persian name; and the whole political and social situation described in the book fits the Persian world, as do the Jewish life and institutions reflected in the book. Thus, there are no serious indications that the book of Judith is a Maccabean product, and there are many allusions to the time of the Persian rule over Palestine. Only a Greek translation of the book is extant, but, from its style, it is clear that the book was originally written in Hebrew. In his preface to the book of Judith, the Latin biblical scholar Jerome (*c.* 347–419/420 CE) states that he used for his translation a "Chaldaean" (*i.e.,* Aramaic) text and that he also used an older Latin translation from Greek. His translation differs in many points from the original text.

Tobit. The other Jewish short story possibly dating from Persian times is the book of Tobit, named after the father of its hero. From the fragments of the book discovered at Qumrān, scholars now know that the original form of the name was Tobi. Tobit was from the Hebrew tribe of Naphtali and lived as an exile in Nineveh; his son was Tobias. Obeying the tenets of Jewish piety, Tobit buried the corpses of his fellow Israelites who had been executed. One day, when he buried a dead man, the warm dung of sparrows fell in his eyes and blinded him. His family subsequently suffered from poverty, but then Tobit remembered that he had once left a deposit of silver at Rages (today Teheran) in Media. He sent his son Tobias along with a companion, who was in reality the angel Raphael under the guise of an Israelite, to retrieve the deposit. During the journey, while Tobias was washing in the Tigris, a fish threatened to devour his foot. Upon instructions from Raphael, Tobias caught the fish and removed its gall, heart, and liver, since it was believed that the smoke from the heart and liver had the power to exorcise demons and that ointment made from the gall would cure blindness. On the way he stopped at Ecbatana (in Persia), where Raguel, a member of Tobias' family, lived. His daughter Sarah had been

married seven times, but the men had been slain by the demon Asmodeus on the wedding night, before they had lain with her. On the counsel of Raphael, Tobias asked to marry Raguel's daughter, and on the wedding night Tobias put Asmodeus to flight through the stench of the burning liver and heart of the fish. Raphael went to Rages and returned with the deposit. When he returned with his young wife and Raphael to Nineveh, Tobias restored his father's sight by applying the gall of the fish to his eyes. Raphael then disclosed that he was one of God's seven angels and ascended into heaven.

"The Grateful Dead"

The story of the book of Tobit is a historicized and Judaized version of the well-known folktale of "The Grateful Dead" (or "The Grateful Ghost"), in which a young man buries the corpse of a stranger despite injunctions against such an act; later the youth wins a bride through the intercession of the dead man's spirit. Asmodeus (in Persian, Aeshma Daeva, the demon of wrath) occurs as a powerful demon in rabbinic literature as well as in folktales. In the Jewish form of the story, "The Grateful Dead" is replaced by the angel Raphael. According to the *Ethiopic Enoch* (20:3; 22:3), Raphael is appointed over the spirits of the souls of the dead (for *Enoch,* see below). Because the cause of this situation is not mentioned in the book of Tobit, the story itself in its Jewish form probably existed before it became the subject of the book of Tobit. The present work is a literary product; the interesting plot gave to the author many occasions to insert religious and moral teachings in the manner of wisdom literature, which is concerned with practical, everyday issues. The book contains prayers, psalms, and aphorisms, most of them put in the mouth of Tobit. It is the oldest Jewish witness of the golden rule (4:15): "And what you hate, do not do to anyone." Eschatological hopes are also described: at the end of time, all Jewish exiles will return, Jerusalem will be rebuilt of precious stones and gold, and all nations will worship the true God. In these eschatological images, however, the figure of the Messiah does not occur.

The religious, social, and literary atmosphere of the book does not contain elements from the Greek period. Thus, the book probably was written already in the Persian period or in the early days of Greek rule (3rd century BCE). The book exists today in three principal recensions, and it is often difficult to determine, in a particular passage, what was the original text. The book was written in Hebrew or Aramaic; the Greek recensions differ, perhaps because they are based on different Semitic versions. These questions may be answered when the Hebrew and Aramaic fragments of the book, which were found among the Dead Sea Scrolls, are published.

The Story of Ahikar. According to the book of Tobit, Ahikar, the cupbearer of the Assyrian king Esarhaddon, was Tobit's nephew; he is a secondary personage in the plot, and his own story is mentioned. Ahikar is the hero of a Near Eastern non-Jewish work, *The Story of Ahikar.* The book exists in medieval translations, the best of them in Syriac. The story was known in the Persian period in the Jewish military colony in Elephantine Island in Egypt, a fact demonstrated by the discovery of fragmentary Aramaic papyri of the work dating from 450–410 BCE. Thus, the author of the book of Tobit probably knew *The Story of Ahikar,* in which, as in the book of Tobit, the plot is a pretext for the introduction of speeches and wise sayings. Some of Tobit's sayings have close parallels in the words of the wise Ahikar.

Baruch. The apocryphon of Baruch, which is extant in Greek and was included in the Septuagint, is attributed to Baruch, secretary to the Old Testament prophet Jeremiah (7th–6th century BCE). It was Baruch who read Jeremiah's letter to the exiles in Babylon. After hearing his words, the Jews repented and confessed their sins. The first part of the book of Baruch (1:1–3, 8), containing a confession of sins by the Jews following the destruction of Jerusalem and the exiles' prayer for forgiveness and salvation, may date from the Persian or at least from the pre-Maccabean period. This early section was originally written in Hebrew and seems to be very ancient. The other two parts (3:9–4:4 and 4:5–5:9) were written in Greek or freely translated from Hebrew or Aramaic. The first is a praise of wisdom: only Israel received wisdom from God, which is the Law of Moses. The last part of the book of Baruch contains Jerusalem's lament over her desolation and her consolation.

Apocryphal works lacking strong indications of influence. *The Letter of Jeremiah.* The Letter of Jeremiah, like the book of Baruch, was conserved—together with the Greek translation of the Book of Jeremiah—in the Septuagint. The oldest witness of the letter is a fragment of a Greek papyrus, written about 160 BCE and found among the Dead Sea Scrolls at Qumrān. Whether the letter was originally written in Greek or is a translation from Hebrew or Aramaic is difficult to decide. The letter attacks the folly of idolatry as did Jeremiah's letter "to those who were to be taken to Babylon as captives." Though, according to some experts, the idolatry described in the book fits Babylonian cults, the only clear indication of its date is that of the Qumrān fragment.

Prayer of Manasseh. In some manuscripts of the Septuagint and in two later Christian writings, a pseudepigraphic Prayer of Manasseh is contained. This prayer was composed with reference to II Chron. 33:11–18, according to which the wicked Judaean king Manasseh repented and prayed. In the present form the prayer is Greek in origin, but it may have existed in a Hebrew version, of which the Greek is a free adaptation. The prayer was probably composed (or translated) in the 1st century BCE.

Additions to Daniel and Esther. Two of the Old Testament Hagiographa (Ketuvim; see above *The Hebrew canon*)—Daniel and Esther—contain, in their Greek translations, numerous additions.

The Prayer of Azariah and the Song of the Three Young Men. The first addition to Daniel (in Greek and Latin translations Dan. 3:24–68) contains the Prayer of Azariah and the Song of the Three Young Men. These are the prayers of Hananiah, Mishael, and Azariah, the three young men who praised God after they had been placed in the midst of the fiery furnace during a persecution of Jews in Babylon, as told in the Book of Daniel. The first prayer is said by Azariah alone; the second, a thanksgiving prayer, is said by all three after having been saved by God. The two poems are not found in the original Daniel and were never a part of it. They were translated from Hebrew originals or adapted from them. A passage from the second, a liturgical hymn of praise, is a poetic expansion of the doxology that was sung in the Temple when the holy name of God was pronounced. Like the other additions to Daniel, the two prayers were probably composed before 100 BCE.

Susanna. The second addition to Daniel, the story of Susanna, and the third one, Bel and the Dragon, are preserved in two Greek versions. In both stories the hero is the wise Daniel. Susanna was the pious and beautiful wife of Joakim, a wealthy Jew in Babylon. Two aged judges became inflamed with love for her. They tried to force her to yield to their lust, and, when she refused, they accused her of committing adultery with a young man, who escaped. She was condemned to death, but when Daniel cross-examined the two elders separately, the first stated that Susanna had been surprised under a mastic tree, the other under a holm tree. Susanna was thus saved and the two false witnesses executed.

The wisdom of Daniel

The short story, perhaps invented even before the extant Book of Daniel was composed, could very well be added to Daniel (whose name means God is my Judge). The story was written in its present form in Greek, since it contains two Greek puns, but a written Semitic prototype may have existed.

Bel and the Dragon. The third Greek addition to the Book of Daniel is the story of Bel and the Dragon. The Babylonians worshipped the idol of the god Bel and daily provided him with much food, but Daniel proved to the King that the food was in reality eaten by the priests. The priests were punished by death and Bel's temple destroyed. The Babylonians also worshipped a dragon, but Daniel declined to worship him. To destroy the beast, Daniel boiled pitch, fat, and hair together: the dragon ate it and burst asunder. After Daniel's sacrilege of slaying the dragon, the King was forced to cast Daniel into the lions' den, but nothing happened to him. Indeed, he was given

a dinner by the prophet Habakkuk, who was brought there by the hair of his head by an angel. On the seventh day the King found Daniel sitting in the den; so he led Daniel out and cast his enemies into the den, where they were devoured.

The two stories are an attack against idolatry. As the addition ends with the story about Daniel in the lions' den, which is also narrated in the canonical Book of Daniel with another motivation, it is probable that this short treatise originated in a tradition that was parallel to the canonical Book of Daniel and that the two stories were translated from a Hebrew or Aramaic original.

Greek and Roman anti-Semitism

Greek additions to Esther. The Hebrew Book of Esther had a religious and social value to the Jews during the time of Greek and Roman anti-Semitism, though the Hebrew short story did not directly mention God's intervention in history—and even God himself is not named. To bring the canonical book up-to-date in connection with contemporary anti-Semitism and to stress the religious meaning of the story, additions were made in its Greek translation. These Greek additions are (1) the dream of Mordecai (Esther's uncle), a symbolic vision written in the spirit of apocalyptic literature; (2) the edict of King Artaxerxes (considered by some to be Artaxerxes II, but more probably Xerxes) against the Jews, containing arguments taken from classical anti-Semitism; (3) the prayers of Mordecai and of Esther, containing apologies for what is said in the Book of Esther—Mordecai saying that he refused to bow before Haman (the grand vizier) because he is flesh and blood and Esther saying that she strongly detests her forced marriage with the heathen king; (4) a description of Esther's audience with the King, during which the King's mood was favourably changed when he saw that Esther had fallen down in a faint; (5) the decree of Artaxerxes on behalf of the Jews, in which Haman is called a Macedonian who plotted against the King to transfer the kingdom of Persia to the Macedonians; and (6) the interpretation of Mordecai's dream and a colophon (inscription at the end of a manuscript with publication facts), where the date, namely, "the fourth year of the reign of Ptolemy and Cleopatra" (*i.e.*, 114 BCE), is given. This indicates that the additions in the Greek Esther were written in Egypt under the rule of the Ptolemies.

I and II Maccabees. *I Maccabees.* The first two of the four books of Maccabees are deuterocanonical (accepted by the Roman Catholic Church). The First Book of the Maccabees is preserved in the Greek translation from the Hebrew original, the original Hebrew name of it having been known to the Christian theologian Origen of Alexandria. At the beginning, the author of the book mentions Alexander the Great, then moves on to the Seleucid king of Syria, Antiochus Epiphanes (died 164/163 BCE), and his persecution of the Jews in Palestine, the desecration of the Jerusalem Temple, and the Maccabean revolt. After the death of the priest Mattathias, who had refused to obey Antiochus, his son Judas Maccabeus succeeded him and led victorious wars against the Syrian Greeks. Exactly three years after its profanation by Antiochus, Judas captured the Temple, cleansed and rededicated it, and in honour of the rededication initiated an annual festival (Ḥanukka) lasting eight days. After Judas later fell in battle against the Syrian Greeks, his brother Jonathan succeeded him and continued the struggle. Only in the time of Simon, Jonathan's brother and successor, did the Maccabean state become independent. A short mention of the rule of Simon's son John Hyrcanus I (135/134–104 BCE) closes the book. The author, a pious and nationalistic Jew and an ardent adherent of the family of Maccabees, evidently lived in the time of John Hyrcanus. The book imitates the biblical style of the historical books of the Old Testament and contains diplomatic and other important—though not necessarily authentic—official documents.

Origin of Ḥanukka

II Maccabees. The Second Book of the Maccabees, or its source, was probably written in the same period as I Maccabees. The book is preceded by two letters to the Jews of Egypt: the first from the year 124 BCE and the second one written earlier (164 BCE) commemorating the rededication of the Temple. In the preface of the book, the author indicates that he has condensed into one book the lost five-volume history compiled by Jason of Cyrene. II Maccabees describes the persecution under Antiochus Epiphanes and the Maccabean wars until the victory of Judas Maccabeus over Nicanor, the commander of the Syrian elephant corps, in 161 BCE. The book, written in Greek, is an important document of Hellenistic historiography. Descriptions of the martyrdom of the priest Eleazar and of the seven brothers under Antiochus, in which Greek dramatic style is linked with Jewish religious spirit, became important for Christian martyrology. The book also furnished proof texts for various Jewish and subsequently Christian doctrines (*e.g.*, doctrines of angels and the resurrection of the flesh).

Wisdom literature. *Ecclesiasticus (or Sirach).* There are two deuterocanonical works of the genre known as wisdom literature, one Hebrew and one Greek. The Hebrew work is called Ecclesiasticus, in the Latin Bible and in Greek manuscripts Sophia Iēsou hyiou Sirach (the Wisdom of Jesus the Son of Sirach); the original Hebrew title was probably Ḥokhmat Yeshua' Ben-Sira, the Wisdom of Ben-Sira. Written in Hebrew about 180–175 BCE, it was translated into Greek by the author's grandson in Egypt. A Syriac translation also was made. Portions (about three-fifths) of the Hebrew text were found in medieval copies in a synagogue of Cairo and a part of the book in a fragment of a scroll from Massada in Palestine (written *c.* 75 BCE). Small Hebrew fragments also were found among the Dead Sea Scrolls; one of them, the Psalms scroll, contains a large part of a poem about wisdom that is a part of the appendix (chapter 51) and that was not written by the author. The Proverbs of Ben-Sira are often quoted in rabbinic literature.

The book is written in the poetical style of the wisdom books of the Old Testament (*e.g.*, Proverbs, Job) and deals with the themes of practical and theoretical morality. The religious and moral position of the author is conservative—he does not believe in the afterlife, but he reflects the contemporary religious positions. He identifies wisdom, the origin of which is divine, with "the Law which Moses commanded," an idea that became important for later Judaism. He also reflects contemporary debates about freedom of will and determinism, and, though realistic in his basic opinions, he sometimes expresses eschatological hopes of salvation for his people. His piety is ethical, though lacking in asceticism; and he invites his readers to enjoy life, which is short (in this point some Greek influence is palpable, but it is not very deep). At the end of the book the author praises, in chronological order, "the fathers of old," from the beginning of history to his contemporary, the high priest Simon, whose appearance in the Temple is poetically described. After some verses comes the colophon with the author's name—the last chapter being an appendix not composed by the author.

The themes of intertestamental wisdom literature

The Wisdom of Solomon. The other deuterocanonical wisdom book, the Wisdom of Solomon, was written in Greek, though it purports to have been written by King Solomon himself. The hypothesis that the first half of the book was translated from Hebrew seems to be without foundation and probably came into existence because, in this section, the author imitated in Greek the Old Testament poetical style. The Wisdom of Solomon was probably written in Alexandria (Egypt) in the 1st century BCE.

The book has three parts. The first (chapters 1–5) concerns the contrast between pious and righteous Jews and the wicked, sinful, and mundane Jews who persecute the righteous; the lot of the righteous is preferable to the sorrows and final condemnation of the sinners. In the second part (chapters 6–9) Solomon speaks about the essence of wisdom and how he attained it. In the third part (chapters 10–19) the author proves the value of wisdom by telling—not in an exact chronological order—how, in the history of Israel from the beginning until the conquest of Palestine, God exalted Israel and punished the heathens, the Egyptians, and the Canaanites. He also describes the folly of heathenism and its origins in human aberrations.

The author fuses Judaism and Hellenism both in style and in thought. Though he imitates biblical style, he is also influenced by Greek rhetoric. He also freely uses Greek philosophical and other terms and is influenced by

Jewish apocalyptic literature. Some close parallels to the Dead Sea sect (at Qumrān), both in eschatology and in anthropology (doctrines about man), can be found in the Wisdom of Solomon.

THE PSEUDEPIGRAPHAL WRITINGS

Works indicating a Greek influence. *The Letter of Aristeas.* An important document of Jewish Hellenistic literature is *The Letter of Aristeas,* a pseudepigraphon ascribed to Aristeas, an official of Ptolemy II Philadelphus, a Greek monarch of Egypt in the 3rd century BCE. The letter is addressed to his brother and gives an account of the translation of the Pentateuch (first five books of the Old Testament) into Greek, by order of Ptolemy. According to the legend, reflected in the letter, the translation was made by 72 elders, brought from Jerusalem, in 72 days. The letter, in reality written by an Alexandrian Jew about 100 BCE, attempts to show the superiority of Judaism both as religion and as philosophy. It also contains interesting descriptions of Palestine, of Jerusalem with its Temple, and of the royal gifts to the Temple.

Translation of the Pentateuch into Greek

IV Maccabees. Another Jewish Hellenistic work combining history and philosophy is *The Fourth Book of Maccabees.* The theme of the book, reflecting the views of the Greek Stoics, is "whether the Inspired Reason is supreme ruler over the passions." This thesis is demonstrated by the martyrdom of the elderly scribe Eleazar and the unnamed seven brothers and their mother, taken from II Macc. 6:18–7:41. The idea of the expiatory force of martyrdom is stressed more in *IV Maccabees* than in its source. The author probably lived in the 1st century BCE and may have been from Antioch (in Syria), where the tombs of the Maccabean martyrs were venerated by the Jews.

III Maccabees. The Greek book called *The Third Book of Maccabees* itself has nothing to do with the Maccabean period. Its content is a legend, a miraculous story of deliverance, which is also independently told—in another historical context—by Josephus (*Against Apion* II, 5). In III Maccabees the story takes place during the reign of Ptolemy IV Philopator (reigned 221–203 BCE). The central episode of the book is the oppression of Egyptian Jews, culminating with an anti-Jewish decree by the King. The Jews who were registered for execution were brought into the hippodrome outside of Alexandria; the King had ordered 500 elephants to be drugged with incense and wine for the purpose of crushing the Jews, but by God's intercession "the beasts turned round against the armed hosts [of the king] and began to tread them under foot and destroy them." The Jews fixed annual celebrations of this deliverance. The book was probably written at the end of the 1st century BCE by an Alexandrian Jew in a period of high anti-Jewish tension.

The Lives of the Prophets. The little book called the *Lives of the Prophets* is a collection of Jewish legends about Old Testament prophets. It is preserved in Greek and in versions and recensions in various languages, all based on the Greek. The purpose of the work was to furnish to the readers of the Bible further information about the prophets. The collection evidently passed through Christian hands since it includes an assumed prophecy of Jeremiah about the birth of Christ. Thus, the date of composition of the supposed original Jewish work and the question as to whether it was originally written in Hebrew or Greek are difficult to resolve. Scholars are inclined toward a 1st-century-CE date in Palestine—with the exception of the life of "Jeremiah," which is Egyptian in origin.

The Ascension of Isaiah. According to the *Lives of the Prophets,* Jeremiah was stoned to death and Isaiah was sawn asunder. These two legends are reflected in two originally Jewish works. The *Ascension of Isaiah,* in which the martyrdom of Isaiah is narrated, is as a whole extant only in Ethiopic, translated from a Greek original, which itself is also known from fragments. The book contains important Christian passages from the 1st century CE, but the story about Isaiah's martyrdom is most likely based upon a Jewish written source. According to this legend, Isaiah was killed by the wicked king Manasseh, who served Beliar-Sammael, the chief of the evil spirits, instead of God. Isaiah, with his followers, had fled to the wilderness, but upon being captured he was sawn asunder with a wooden saw, and his followers fled to the region of Tyre and Sidon. The activity of Beliar is known also from the writings of the sect that preserved the Dead Sea Scrolls and similar writings, and the story itself resembles in some way the history of the Dead Sea sect; but no fragment of the Jewish part of the book was found among the Dead Sea Scrolls. The original *Martyrdom of Isaiah* was written probably in Hebrew or Aramaic before the 1st century CE.

The legendary death of Isaiah

Paralipomena of Jeremiah. In the last chapter of the Greek text of the *Paralipomena* (additional stories) *of Jeremiah,* there is a hint of the Christian part of the *Ascension of Isaiah:* the people stoned Jeremiah to death because he, like Isaiah before him, prophesied the coming of Christ. In a parallel legend (preserved in Arabic), both the violent death of Jeremiah and the Christian motif are lacking. The book begins shortly before and ends shortly after the Babylonian Exile and contains mostly otherwise unknown legends. The legend about the long sleep of Abimelech (the biblical Ebed-melech—an Ethiopian eunuch who rescued Jeremiah from a cistern), who slept and so did not see the destruction of Jerusalem by the Babylonians—is based upon a legendary understanding of Psalm 126:1; a similar legend about another person is preserved in the Talmud (the authoritative rabbinical compendium of Jewish law, lore, and commentary). The book is basically Jewish, and the last chapter was Christianized. The Jewish work was probably written at the end of the 1st century CE or at the beginning of the 2nd, originally in either Hebrew, Aramaic, or Greek.

The Testament of Job. Though there are scholars who think that the *Testament of Job* was once written in Hebrew or Aramaic, it is more probable that the existing Greek text of the book is the original or even a rewritten later version of a Greek work; a fragment of an older form is probably preserved in the Greek translation of Job (2:9). Job is identified, according to some Jewish traditions, with the biblical Jobab (king of Edom), and his (second) wife is Dinah, Jacob's daughter. Job knew by revelation that, for destroying an idol, he would undergo suffering but that a happy end would be the final outcome. Thus, in contrast to the biblical Book of Job, this work does not deal with the question of God's righteousness but places great emphasis on resurrection and eternal life. These special motifs in the book indicate that the book probably was written by a member of an unknown Jewish group that upheld a high mystical spirituality. The extreme "pietistic" tendency of the book is noted in the exaggeration of Job's love for suffering and of his charity to the poor. At the end of the book Job's soul was taken to heaven in a heavenly chariot. The book was probably written before 70 CE.

Life of Adam and Eve. The many Christian legends in many languages about the lives of Adam and Eve probably have their origin in a Jewish writing (or writings) about the biblical first man and woman. The most important of these works are the Latin *Vita Adae et Evae* (*Life of Adam and Eve*) and a Greek work closely parallel to it, named erroneously by its first editor the *Apocalypse of Moses.* The narrative runs from the Fall to the deaths of Adam and Eve. The religious message in the story involves the repentance of Adam and Eve after their expulsion from paradise—and the description of their deaths does not show any traces of the idea of original sin, which was important in later Christian theology. Nonetheless, there are definitely Christian passages in the various versions, and the treatment of Adam in the literature of the Ebionites (an early Jewish Christian sect) shows an affinity for the story. Thus, the Jewish source probably was composed in the 1st century CE in Jewish circles that influenced the Ebionites. The original language of this supposed source is unknown.

Apocalyptic and eschatological works. *III Baruch.* Apocalyptic literature was much concerned about sources of information about the heavenly world and about the places of the damned and saved souls. In later Jewish and early Christian apocalypses, in which the hero undertakes a heavenly trip and sees the secrets that are hidden from others, these sources of information are highly significant. *III Baruch,* a book written in Greek—in which Baruch,

Interests of apocalyptic literature

the disciple of the prophet Jeremiah, visits the universe and sees its secrets and the places of the souls and of the angels—is such an apocalypse. In the Greek text the number of heavens visited by Baruch is five, but it is possible that originally he was said to have seen seven heavens. There are Christian passages in the book, but it seems to have been a Jewish work from the 1st century CE later rewritten by a Christian.

II Enoch. Similar in content is *II Enoch,* or *The Book of the Secrets of Enoch,* which is preserved only in an Old Slavonic translation. The oldest text does not contain any Christian additions nor any passage from which it could be concluded that the book was written in Greek. Thus, the book could have been written originally in Hebrew or Aramaic, probably in the 1st century CE. The hero who visits the heavens is the biblical Enoch (son of Jared). The author of the book knew at least some of the treatises contained in *I Enoch.* The book also contains the story of the miraculous birth of the biblical priest-king Melchizedek.

The Psalms of Solomon. Other Jewish apocalypses or books containing eschatological elements did not deal with the mysteries of celestial worlds but rather with the political aspect of apocalyptic thought and with the last days and the messianic age. This latter theme is one of the important motifs of the *Psalms of Solomon,* a book written originally in Hebrew; only the Greek translation of the *Psalms* is preserved. The title is evidently a later addition—the author himself apparently had no intention to give the impression that his 18 psalms were composed by the biblical king Solomon. The *Psalms of Solomon* were written in Jerusalem about the middle of the 1st century BCE, and, though persons are not named, they reflect the dramatic events of the Jewish history of that period, especially the Roman general Pompey's conquest of Jerusalem in 63 BCE and his violent death in Egypt. In Psalm 17, the author denounces the Hasmonean dynasty as illegal and describes the coming of the Davidic Messiah (a kingly saviour from the line of David). His religious opinion resembles the teachings of the Pharisees (a sect that espoused a reinterpretation of Jewish laws and customs), especially in his faith in the resurrection of the body and in the question of free will, though he most likely was not a Pharisee but rather a member of the community of Ḥasidim, a Jewish pietistic group that had joined the Maccabean revolt from its beginning.

The Assumption of Moses. The *Assumption of Moses* originally contained apocalyptic material—no longer extant—in the form of a legend. According to Origen, the dispute between the archangel Michael and the devil for the body of Moses was narrated in the *Assumption of Moses.* This legend, which has parallels in the rabbinic literature, probably formed the end of the *Assumption of Moses,* the first part of which was discovered in a Latin manuscript. The Latin version was translated from Greek, but the original language was Semitic, probably Hebrew.

The coming of the Kingdom of Heaven

The main content of the preserved part is Moses' prophecy about the future, from his time until the Kingdom of Heaven will be revealed. According to the custom of apocalyptic literature, names of persons and groups are not mentioned, but from the last events hinted at in the book it can be assumed that it was written at the beginning of the 1st century CE, while Jesus was alive. In its older version, the book apparently was written at the beginning of the Maccabean revolt, some years before the Book of Daniel; after a description of the pre-Maccabean Hellenistic priests (chapter 5) and before the description of the persecutions by Antiochus Epiphanes (chapter 8), chapters 6–7 contain Jewish history from the time of the later Hasmonean rulers to the time of the sons of Herod—as well as polemics against leading religious circles, which are accused of religious hypocrisy, as are the Pharisees in the Christian Gospels. The author of these chapters (6–7), a contemporary of Jesus, evidently erroneously identified the wicked pre-Maccabean priests with the wicked late Maccabean priestly rulers and also interpreted Antiochus Epiphanes as a kind of eschatological Antichrist. No messianic figure is mentioned in the eschatological description of the Kingdom of God: God himself and his angel will bring the salvation.

The Sibylline Oracles. *The Sibylline Oracles* is a collection of oracles in Greek verse containing pagan, Jewish, and Christian material from various periods. It comprised 15 books (books IX, X, and XV are lost), of which 4,240 verses are extant. Sibyl is the name (or title) of a legendary ancient pagan prophetess. In the Hellenistic period, eastern nations fabricated Sibylline oracles as propagandistic literature against Greek and, later, against Roman occupation. The political anti-Roman and anti-pagan tone is typical of the Jewish and Christian parts of Sibylline oracles; they also contain religious propaganda for the respective religion. Because Jewish parts used pagan material and Christian authors interpolated Jewish parts or used Jewish material, it is sometimes difficult to decide what verses are pagan, Jewish, or Christian. *The Sibylline Oracles* perhaps became a part of Jewish (and Christian) apocalyptic literature because of their emphasis on eschatology. The oldest Jewish "Sibyl" is contained in the third book: it dates from about 140 BCE and describes the coming of the Messiah. Book IV was written by a Jew about 80 CE: the eruption of Vesuvius (79) is viewed as a divine punishment for the massacre of Jews in the Roman war (70). Book V was written by a Jew about 125.

II Esdras (or IV Esdras). Two important apocalyptic pseudepigrapha (II Esdras and the *Apocalypse of Baruch*), in which the political and eschatological aspects are central to the aim of the books, were written in Palestine at the end of the 1st century CE as a consequence of the catastrophic destruction of the Second Temple in Jerusalem (70). Both were written as if they reflected the doom that befell the people of Israel after the destruction of the First Temple (586 BCE) by the Babylonians. II Esdras (or IV Esdras) was written in Hebrew, but only various translations from a lost Greek version are preserved. The Latin version (in which chapters 1–2 and 15–16 have been added by a Christian hand) at one time was printed at the end of the Latin Bible. The book consists of six visions attributed to the biblical Ezra (who is, at the beginning of the book, erroneously identified with Salathiel, the father of Zerubbabel, a leader of the returning exiles from Babylon). The tragedy of his nation evokes in the heart of the author questions about God's righteousness, the human condition, the meaning of history, and the election of Israel; "Ezra" does not find consolation and full answer in the words of the angel who was sent to him, which also contain revelations about the last days. In the fourth vision "Ezra" sees a mourning woman; she disappears and a city (the New Jerusalem) stands in her place. In the fifth vision a monstrous eagle appears, the symbol of the Roman Empire, and a lion, the symbol of the Messiah. The final victory of the Messiah is described in the last vision of the man (Son of man) coming from the sea. In chapter 14 "Ezra" is described as dictating 94 books: 24 are the books of the Hebrew Bible, and the other 70 are esoteric.

Questions about the meaning of history

The Apocalypse of Baruch. The *Apocalypse of Baruch* was written about the same time as II (IV) Esdras, and the less profound *Apocalypse* probably depends much upon II Esdras. The *Apocalypse of Baruch* survives only in a Syriac version translated from Greek; originally the book was composed in Hebrew or Aramaic and is ascribed to Baruch, the disciple of Jeremiah and a contemporary of the destruction of the First Temple. If II Esdras asks questions about important problems of human history and the tragic situation of Israel after the destruction of the Second Temple, the *Apocalypse of Baruch* apparently was written to give a positive, traditional answer to these doubts.

Pseudepigrapha connected with the Dead Sea Scrolls. There are three Pseudepigrapha that are closely connected with the writings of the Dead Sea sect: the *Book of Jubilees,* the Ethiopic *Book of Enoch,* and the *Testaments of the Twelve Patriarchs.* It is not accidental that fragments of the two first books and of two sources of the third were found among the Dead Sea Scrolls.

The Book of Jubilees. From the fragments of the *Book of Jubilees* among the Dead Sea Scrolls, scholars note that the book was originally written in biblical Hebrew. The whole book is preserved in an Ethiopic version translated from Greek.

The book is written in the form of a revealed history of

Israel from the creation until the dwelling of Moses on Mt. Sinai, where the content of the book was revealed to Moses by "the angel of the presence." The *Book of Jubilees* in fact is a legendary rewriting of the book of Genesis and a part of Exodus. One of the main purposes of the author is to promote, in the form of divine revelation, a special sectarian interpretation of Jewish law. All the legal prescriptions noted in the book were practiced by the Dead Sea sect; in connection with the solar calendar of 52 weeks, one of the Dead Sea Scrolls even mentions the *Book of Jubilees* as the source. The (unpublished) *Temple Scroll,* a book of sectarian prescriptions that paraphrases—also as divine revelation—a part of the Mosaic Law and was composed by the Dead Sea sect before 100 BCE (*i.e.,* in the same period as the *Book of Jubilees*), closely resembles some parts of the *Book of Jubilees.* Thus, the *Book of Jubilees* could be accepted by the Dead Sea sect and apparently was written in the same circles, immediately before the sect itself came into existence. The apocalyptical hopes expressed in the book are also identical to those of the Dead Sea sect.

The Book of Enoch. Another book that was written during the period of the apocalyptic movement in which the Dead Sea sect came into existence is the *Book of Enoch,* or *I Enoch.* It was completely preserved in an Ethiopic translation from Greek, and large parts from the beginning and end of the Greek version have been published from two papyri. Aramaic fragments of many parts of the book were found among the Dead Sea Scrolls, as were Hebrew fragments of the *Book of Noah,* either one of the sources of *Enoch* or a parallel elaboration of the same material. Passages of the *Book of Noah* were included in *Enoch* by its redactor (editor). Scholars generally agree that the somewhat haphazard redaction of the book was made in its Greek stage, when a redactor put together various treatises of the Enochic literature that were written at various times and reflected various trends of the movement.

The *Book of Noah*

Besides the passages from the *Book of Noah,* five treatises are included in the *Book of Enoch.* The hero of all of them is the biblical Enoch. The first treatise (chapters 1–36) speaks about the fall of the angels, who rebelled before the Flood, and describes Enoch's celestial journeys, in which divine secrets were revealed to him. It was probably written in the late 2nd century BCE.

The second part of the *Book of Enoch* is the "Parables" (or Similitudes) of Enoch (37–71). These three eschatological sermons of Enoch refer to visions; their original language was probably Hebrew rather than Aramaic. This treatise is an important witness for the belief in the coming of the Son of man, who is expressly identified with the Messiah; in chapters 70–71, which are probably a later addition, the Son of man is identified with Enoch himself. The treatise probably dates from the 1st century BCE.

As Aramaic fragments from the Dead Sea Scrolls show, the astronomical book entitled "The Book of the Heavenly Luminaries" (chapters 72–82) is in the present form abbreviated in the *Book of Enoch.* All these astronomical mysteries were shown to Enoch by the angel Uriel. The treatise propagates the same solar calendar that is also known from the *Book of Jubilees* and from the Dead Sea sect. This treatise was probably written before the year 100 BCE.

The fourth treatise (chapters 83–90) contains two visions of Enoch: the first (chapters 83–84), about the Flood, is in reality only a sort of introduction to the second one ("the vision of seventy shepherds"), which describes the history of the world from Adam to the messianic age; the personages of the visions are allegorically described as various kinds of animals. The symbolic description of history continues to the time of Judas Maccabeus; then follows the last assault of Gentiles and the messianic period. Thus, the treatise was written in the early Hasmonean period, some time after the biblical Book of Daniel.

The fifth treatise (chapters 91–107) contains Enoch's speech of moral admonition to his family. The moral stress and the social impact is similar to parts of Jesus' teaching; even the form of beatitudes (blessings) and woes is present. The treatise shows some affinities to the Dead Sea Scrolls, but the author was not a member of the Dead Sea sect; he opposes the central teaching of the sect, the doctrine of predestination (98: 4–5). The treatise apparently was written at the end of the 1st century BCE. Chapter 105, lacking in the Greek version, is a late interpolation, probably of Christian origin.

The author of the treatise himself apparently incorporated into it a small apocalypse, the "Apocalypse of Weeks" (93:1–10; 91:12–17); in it the whole of human history is divided into ten weeks; seven of them belong to the past and the last three to the future.

Testaments of the Twelve Patriarchs. The third pseudepigraphon that shows important affinities with the Dead Sea sect is the *Testaments of the Twelve Patriarchs,* the last speeches of the 12 sons of the Hebrew patriarch Jacob. In its extant form, containing Christian passages, the book was written in Greek. Fragments of two original Semitic sources of the book were found among the Dead Sea Scrolls: the Aramaic "Testament of Levi" (fragments of it were also discovered in Aramaic in the medieval Geniza, or synagogue storeroom, in Cairo) and a Hebrew fragment of the "Testaments of Naphtali." A Hebrew "Testament of Judah," which was used both by the *Book of Jubilees* and the *Testaments of the Twelve Patriarchs* in their description of the wars of the sons of Jacob, also probably existed.

Whether Hebrew and Aramaic prototypes for all the 12 testaments of the patriarchs existed is difficult to ascertain. The present book was originally written in Greek. In it each of the sons of Jacob before his death gives moral advice to his descendants, based upon his own experience. All the testaments, with the exception of Gad, also contain apocalyptic predictions.

Between the *Testaments of the Twelve Patriarchs* and the Dead Sea sect there is a historical and ideological connection. The sources of the book were found among the scrolls, the source of the "Testament of Levi" is quoted in a sectarian writing (the Damascus Document), a dualistic outlook is common to the book and the sect, and the devil is named Belial in both. There are, however, important differences: in regard to the nature of the dualism between good and evil, there is in the *Testaments* the concept of the good and bad inclination, known from rabbinical literature, which does not exist in the scrolls; though the sect believed in an afterlife of souls, the *Testaments* reflect the belief in the resurrection of the body; there are no traces of the doctrine of predestination in the testaments, a doctrine that is so important for the sect. Only the "Testament of Asher" preaches, as did the Dead Sea sect, hatred against sinners; the other testaments stress, as does rabbinic literature and especially Jesus, the precept of love for God and neighbour. Thus, it is probable that the testaments of the patriarchs were composed in circles in which doctrines of the Dead Sea sect were mitigated and combined with some rabbinic doctrines. A similar humanistic position, founded both on doctrines of the Dead Sea sect and of the Pharisees, is typical of Jesus' message, and there are important parallels between his message and the *Testaments of the Twelve Patriarchs.*

Connection between the *Testaments of the Twelve Patriarchs* and the Dead Sea sect

QUMRĀN LITERATURE (DEAD SEA SCROLLS)

New literary documents from the intertestamental period were found in the caves of Qumrān in the vicinity of the Dead Sea in the 1940s, but only a portion of them has yet been published. All the Dead Sea Scrolls were written before the destruction of the Second Temple; with the exception of small Greek fragments, they are all in Hebrew and Aramaic. The scrolls formed the library of an ancient Jewish sect, which probably came into existence at the end of the 2nd century BCE and was founded by a religious genius, called in the scrolls the Teacher of Righteousness. Scholars have tried to identify the sect with all possible groups of ancient Judaism, including the Zealots and early Christians, but it is now most often identified with the Essenes; all that the sectarian scrolls contain fits previous information about the Essenes, and the Dead Sea Scrolls help scholars to interpret the descriptions about the Essenes in ancient sources.

Apocryphal and pseudepigraphal writings. The importance of the discovery is very great; the scrolls of books

of the Old Testament caused a new evaluation of the history of the text of the Hebrew Bible; fragments of the Apocrypha (Sirach and Tobit) and of already known and unknown Pseudepigrapha enlarge knowledge about Jewish literature of the intertestamental period, and the properly sectarian scrolls are important witnesses about an ancient sect that influenced, in some points, the origins of Christianity.

Among the previously unknown Pseudepigrapha were large parts of an Aramaic scroll, the *Genesis Apocryphon,* which retells stories from Genesis in the manner of a number of apocryphal books. The chapters that are preserved are concerned with Lamech, his grandfather Enoch, Noah, and Abraham, and the narrators in the scroll are the respective biblical heroes. There is a close affinity between this scroll and the *Book of Jubilees* and *Book of Enoch,* fragments of these books having been also found among the Dead Sea Scrolls. Another pseudepigraphon that resembles the Dead Sea sect in spirit is the *Testaments of the Twelve Patriarchs;* fragments of two of its sources, namely, the Aramaic "Testament of Levi" and a Hebrew "Testament of Naphtali," are extant in the Qumrān library. All these books were composed in an apocalyptic movement in Judaism, in the midst of which the Dead Sea sect originated. It is sometimes difficult to ascertain if a work was written within the sect itself or if it represents the broader movement. The largest scroll, the *Temple Scroll,* is as yet unpublished. It describes—by the mouth of God himself and in Hebrew—not the Temple of the last days but the Temple as it should have been built. There are strong ties between the *Temple Scroll* and the *Book of Jubilees* and the prescriptions in it fit the conceptions of the sect; the work was composed by the sectarians themselves.

The Temple Scroll

Pesharim. An important source of knowledge about the history of the Dead Sea sect is the *pesharim* ("commentaries"; singular *pesher*). The sectarian authors commented on the books of Old Testament prophets and the book of Psalms and in the commentaries explained the biblical text as speaking about the history of the sect and of events that happened in the time of its existence. According to the manner of apocalyptic literature in the *pesharim,* persons and groups are not named with their proper names but are described by symbolic titles—*e.g.,* the Teacher of Righteousness for the founder of the sect. The most important sectarian commentaries are the *pesharim* on Habakkuk and on Nahum.

The War of the Sons of Light Against the Sons of Darkness. One of the most interesting Dead Sea Scrolls is *The War of the Sons of Light Against the Sons of Darkness,* a description of the eschatological war between the Sons of Light—*i.e.,* the sect—and the rest of mankind, first with the other Jews and then with the Gentiles. At the end the Sons of Light will conquer the whole world, and in this war they will be helped by heavenly hosts; the Sons of Darkness, aided by the devil Belial and his demonic army, and, finally, all wicked ones will be destroyed. The work contains prayers and speeches that will be uttered in the eschatological war as well as military and other ordinances. Thus, the book also could be called the *Manual of Discipline* for the last war.

Books of ordinances. Other books of ordinances of the sect have been preserved, containing prescriptions and other material. Three such compositions are written on one scroll: the *Manual of Discipline,* the *Rule of the Congregation,* and the manual of *Benedictions.* The *Manual of Discipline* is the rule (or statement of regulations) of the Essene community; the most important part of this work is a treatise about the special theology of the sect. The *Rule of the Congregation* contains prescriptions for the eschatological future when the sect is expected to be the elite of the nation. The manual of *Benedictions,* preserved only in a fragmentary state, contains benedictions that are to be said in the eschatological future.

Another sectarian book of ordinances is the Damascus Document (the Zadokite Fragments). The work was already known from two medieval copies before the discovery of the Dead Sea Scrolls, but fragments of it also were found in Qumrān, and the connection between this work and the Dead Sea sect is evident. The Damascus Document was written in a community in Damascus, which was not as rigidly organized as the Essenes. The work contains the rules of this community and reminiscences of the sect's history. Some scholars think that "Damascus" is only a symbolical name for Qumrān.

Hodayot. One of the most important Essene works is the *Hodayot* ("Praises")—a modern Hebrew name for the *Thanksgiving Psalms.* This scroll contains sectarian hymns of praise to God. In its view of the fleshly nature of man, who can be justified only by God's undeserved grace, it resembles St. Paul's approach to the same problem. Some scholars think that the work, or a part of it, was written by the Teacher of Righteousness.

The Thanksgiving Psalms

Among other fragments of scrolls liturgical texts of prayers were found, as well as fragments of horoscopes written in a cryptic script.

(D.Fl.)

New Testament canon, texts, and versions

THE NEW TESTAMENT CANON

Conditions aiding the formation of the canon. The New Testament consists of 27 books, which are the residue, or precipitate, out of many 1st–2nd-century AD writings that Christian groups considered sacred. In these various writings the early church transmitted its traditions: its experience, understanding, and interpretation of Jesus as the Christ and the self-understanding of the church. In a seemingly circuitous interplay between the historical and theological processes, the church selected these 27 writings as normative for its life and teachings—*i.e.,* as its canon (from the Greek *kanōn,* literally, a reed or cane used as a measuring rod and, figuratively, a rule or standard). Other accounts, letters, and revelations—*e.g.,* the *Didachē* (Teaching of the Twelve Apostles), *Gospel of Peter, First Letter of Clement, Letter of Barnabas, Apocalypse* (Revelation) *of Peter, Shepherd of Hermas*—exist, but through a complex process the canon was fixed for both the Eastern and Western churches in the 4th century. The canon contained four Gospels (Matthew, Mark, Luke, and John), Acts, 21 letters, and one book of a strictly revelatory character, Revelation. These were not necessarily the oldest writings, not all equally revelatory, and not all directed to the church at large.

The Old Testament in its Greek translation, the Septuagint (LXX), was the Bible of the earliest Christians. The New Covenant, or Testament, was viewed as the fulfillment of the Old Testament promises of salvation that were continued for the new Israel, the church, through the Holy Spirit, which had come through Christ, upon the whole people of God. Thus, the Spirit, which in the Old Testament had been viewed as resting only on special charismatic figures, in the New Testament became "democratized"—*i.e.,* was given to the whole people of the New Covenant. In postbiblical Judaism of the first Christian centuries, it was believed that the Spirit had ceased after the writing of the Book of Malachi (the last book of the Old Testament canon) and that no longer could anyone say "Thus saith the Lord," as had the prophets, nor could any further holy writ be produced.

Significance of the Old Testament

The descent of the Spirit on the community of the Messiah (*i.e.,* the Christ) was thus perceived by Christians as a sign of the beginning of the age to come, and the church understood itself as having access to that inspiration through the Spirit. Having this understanding of itself, the church created the New Testament canon not only as a continuation and fulfillment of the Old Testament but also as qualitatively different, because a new age had been ushered in. These 27 books, therefore, were not merely appended to the traditional Jewish threefold division of the Old Testament—the Law (Torah), the Prophets (Nevi'im), and the Writings (Ketuvim)—but rather became the New Testament, the second part of the Christian Bible, of which the Old Testament is the first.

Because of a belief that something almost magical occurs—with an element of secrecy—when a transmitted oral tradition is put into writing, there was, in both the Old and New Testaments, an expression of reluctance about committing sacred material to writing. When such

sacred writings are studied to find the revealed word of God, a settled delimiting of the writings—*i.e.,* a canon—must be selected. In the last decade of the 1st century, the Synod of Jamnia (Jabneh), in Palestine, fixed the canon of the Bible for Judaism, which, following a long period of flux and fluidity and controversy about certain of its books, Christians came to call the Old Testament. A possible factor in the timing of this Jewish canon was a situation of crisis: the fall of Jerusalem and reaction to the fact that the Septuagint was used by Christians and to their advantage, as in the translation of the Hebrew word *'alma* ("young woman") in chapter 7, verse 14, of Isaiah—"Behold, a young woman shall conceive and bear a son, and shall call his name Immanuel"—into the Greek term *parthenos* ("virgin").

As far as the New Testament is concerned, there could be no Bible without a church that created it; yet conversely, having been nurtured by the content of the writings themselves, the church selected the canon. The concept of inspiration was not decisive in the matter of demarcation because the church understood itself as having access to inspiration through the guidance of the Spirit. Indeed, until *c.* AD 150, Christians could produce writings either anonymously or pseudonymously—*i.e.,* using the name of some acknowledged important biblical or apostolic figure. The practice was not believed to be either a trick or fraud. Apart from letters in which the person of the writer was clearly attested—as in those of Paul, which have distinctive historical, theological, and stylistic traits peculiar to Paul—the other writings placed their emphases on the message or revelation conveyed, and the author was considered to be only an instrument or witness to the Holy Spirit or the Lord. When the message was committed to writing, the instrument was considered irrelevant, because the true author was believed to be the Spirit. By the mid-2nd century, however, with the delay of the final coming (the Parousia) of the Messiah as the victorious eschatological (end-time) judge and with a resulting increased awareness of history, increasingly a distinction was made between the apostolic time and the present. There also was a gradual cessation of "authentically pseudonymous" writings in which the author could identify with Christ and the Apostles and thereby gain ecclesiastical recognition.

Acceptance of anonymous or pseudonymous writings

The process of canonization. The process of canonization was relatively long and remarkably flexible and detached; various books in use were recognized as inspired, but the Church Fathers noted, without embarrassment or criticism, how some held certain books to be canonical and others did not. Emerging Christianity assumed that through the Spirit the selection of canonical books was "certain" enough for the needs of the church. Inspiration, it is to be stressed, was neither a divisive nor a decisive criterion. Only when the canon had become self-evident was it argued that inspiration and canonicity coincided, and this coincidence became the presupposition of Protestant orthodoxy (*e.g.,* the authority of the Bible through the inspiration of the Holy Spirit).

The need for consolidation and delimitation. Viewed both phenomenologically and practically, the canon had to be consolidated and delimited. Seen historically, however, there were a number of reasons that forced the issue of limiting the canon. Oral tradition had begun to deteriorate in post-apostolic times, partly because many or most of the eyewitnesses to the earliest events of Jesus' life and death and the beginning of the church had died. Also, the oral tradition may simply have suffered in transmission. Papias (died *c.* 130), a bishop of Hieropolis, in Asia Minor, was said by Irenaeus (died *c.* 200), a bishop of Lugdunum (now Lyon, France) to have been an eyewitness of the Apostle John. Papias had said, "For I did not suppose that the things from the books would aid me so much as the things from the living and continuing voice." Eusebius (*c.* 260–*c.* 340), a church historian, reported these comments in his *Ecclesiastical History* and pointed out inconsistencies in Papias' recollections, doubted his understanding, and called him "a man of exceedingly small intelligence." Large sections of oral tradition, however, which were probably translated in part from Aramaic before being written down in Greek—such as the Passion (suffering of Christ) narrative, many sayings of Jesus, and early liturgical material—benefitted by the very conservativism implicit in such traditions. But because the church perceived its risen Lord as a living Lord, even his words could be adjusted or adapted to fit specific church needs. Toward the end of the 1st century, there was also a conscious production of gospels. Some gospels purported to be words of the risen Lord that did not reflect apostolic traditions and even claimed superiority over them. Such claims were deemed heretical and helped to push the early church toward canonization.

Faced with heresy and claims to late revelations, the early church was constrained to retain the historical dimension of its faith, the *ephapax,* or the "once for all," revelation of God in Jesus Christ.

Impulse toward canonization from heretical movements. Gnosticism (a religious system with influence both on Judaism and Christianity) tended to foster speculation, cutting loose from historical revelation. In defense the orthodox churches stressed the apostolic tradition by focussing on Gospels and letters from apostolic lives and distinguished them from Gnostic writings, such as the *Gospel of Truth* (mentioned by Irenaeus) and now found in Coptic translation in a collection of Gnostic writings from Egypt; it is a Coptic manuscript of a Valentinian Gnostic speculation from the mid-2nd century—*i.e.,* a work based on the teachings of Valentinus, a Gnostic teacher from Alexandria. In the same collection is the *Gospel of Thomas* in Coptic, actually a collection of sayings purporting to be the words of the risen Christ, the living Lord. This "gospel" also occurred in Greek (*c.* 140), and warnings against it as heretical were made by the Church Fathers in the 2nd to the 4th centuries.

Gnosticism

In a general prophetic apocalyptic mood, another heresy, Montanism, arose. This was an ecstatic enthusiastic movement claiming special revelation and stressing "the age of the spirit." Montanus (died *c.* 175) and two prophetesses claimed that their oracular statements contained new and contemporary authoritative revelations. This break with the apostolic time caused vigorous response. An anti-Montanist reported that "the false prophet is one who speaks in ecstasy after which follow freedom . . . and madness of soul."

The single most decisive factor in the process of canonization was the influence of Marcion (flourished *c.* 140), who had Gnostic tendencies and who set up a "canon" that totally repudiated the Old Testament and anything Jewish. He viewed the Creator God of the Old Testament as a cruel God of retribution and the Jewish Law. His canon consisted of *The Gospel,* a "cleaned up" Luke (the least Jewish), and the *Apostolikon* (ten Pauline letters with Old Testament references and analogies edited out, without Hebrews, I and II Timothy, and Titus). This restrictive canon acted as a catalyst to the formation of a canon more in line with the thought of the church catholic (universal).

Late-2nd-century canons. By the end of the 2nd century, Irenaeus used the four canonical Gospels, 13 letters of Paul, I Peter, I and II John, Revelation, *Shepherd of Hermas* (a work later excluded from the canon), and Acts. Justin Martyr (died *c.* 165), a Christian apologist, wrote of the reading of the Gospels, "the memoirs of the Apostles," in the services, in which they were the basis for sermons. In his writings he quoted freely from the Gospels, Hebrews, the Pauline Letters, I Peter, and Acts. Justin's Syrian pupil, Tatian (*c.* 160), although he quotes from John separately, is best known for his *Diatessaron* (literally, "through four" [gospels], but also a musicological term meaning "choral" "harmony"), which was a life of Christ compiled from all four Gospels but based on the outline and structure of John. This indicates both that Tatian was aware of four gospel traditions and that their canonicity was not fixed in final form at his time in Syria. Although Tatian was later declared a heretic, the *Diatessaron* was used until the 5th century and influenced the Western Church even after four separated gospels were established.

The influence of Tatian's *Diatessaron*

The first clear witness to a catalog of authoritative New Testament writings is found in the so-called Muratorian Canon, a crude and uncultured Latin 8th-century manuscript translated from a Greek list written in Rome *c.* 170–

180, named for its modern discoverer and publisher Lodovica Antonio Muratori (1672–1750). Though the first lines are lost, Luke is referred to as "the third book of the Gospel," and the canon thus contains [Matthew, Mark] Luke, John, Acts, 13 Pauline letters, Jude, two letters of John, and Revelation. Concerning the *Apocalypse of Peter,* it notes that it may be read, although some persons object; it rejects the *Shepherd of Hermas* as having been written only recently in Rome and lacking connection with the apostolic age. The Wisdom of Solomon (a Jewish intertestamental writing), is included in the accepted works as written in Solomon's honour.

Criteria of canonicity

Some principles for determining the criteria of canonicity begin to be apparent: apostolicity, true doctrine (*regula fidei*), and widespread geographical usage. Such principles are indicated by Muratori's argument that the Pauline Letters are canonical and universal—the Word of God for the whole church—although they are addressed to specific churches, on the analogy of the letters to the seven churches in Revelation; in a prophetic statement to the whole church, seven specific churches are addressed, then the specific letters of Paul can be read for all. Thus, the catholic status of the Pauline letters to seven churches is vindicated on the basis of the revelation of Jesus Christ to John, the seer and writer of Revelation. Wide usage in the church is indicated in calling Acts the Acts of *all* the Apostles and in the intention of the "general address"—*e.g.,* "To those who are called," in Jude—of the Catholic (or general) Letters—*i.e.,* I and II Peter, I, II, and III John, James, and Jude. The criterion of accordance with received teaching is plain in the rejection of heretical writings. The Muratorian Canon itself may have been, in part, a response to Marcion's heretical and reductive canon.

The criteria of true doctrine, usage, and apostolicity all taken together must be satisfied, then, in order that a book be judged canonical. Thus, even though the *Shepherd of Hermas,* the *First Letter of Clement,* and the *Didachē* may have been widely used and contain true doctrines, they were not canonical because they were not apostolic nor connected to the apostolic age, or they were local writings without support in many areas.

During the time of the definitive formation of the canon in the 2nd century, apparent differences existed in the Western churches (centred in or in close contact with Rome) and those of the East (as in Alexandria and Asia Minor). It is not surprising that the Roman Muratorian Canon omitted Hebrews and accepted and held Revelation in high esteem, for Hebrews allows for no repentance for the baptized Christian who commits apostasy (rejection of faith), a problem in the Western Church when it was subjected to persecution. In the East, on the other hand, there was a dogmatic resistance to the teaching of a 1,000-year reign of the Messiah before the end time—*i.e.,* chiliasm, or millenarianism—in Revelation. There was also a difference in the acceptance of Acts and the Catholic Letters. With the continued expansion of the church, particularly in the 2nd century, consolidation was necessary.

Canonical standards of the 3rd and 4th centuries. Clement of Alexandria, a theologian who flourished in the late 2nd century, seemed to be practically unconcerned about canonicity. To him, inspiration is what mattered, and he made use of the *Gospel of the Hebrews,* the *Gospel of the Egyptians,* the *Letter of Barnabas,* the *Didachē,* and other extracanonical works. Origen (died *c.* 254), Clement's pupil and one of the greatest thinkers of the early church, distinguished at least three classes of writings, basing his judgment on majority usage in places that he had visited: (1) *homologoumena* or *anantirrhēta,* "undisputed in the churches of God throughout the whole world" (the four Gospels, 13 Pauline Letters, I Peter, I John, Acts, and Revelation); (2) *amphiballomena,* "disputed" (II Peter, II and III John, Hebrews, James, and Jude); and (3) *notha,* "spurious" (*Gospel of the Egyptians, Thomas,* and others). He used the term "scripture" (*graphē*) for the *Didachē,* the *Letter of Barnabas,* and the *Shepherd of Hermas,* but did not consider them canonical. Eusebius shows the situation in the early 4th century. Universally accepted are: the four Gospels, Acts, 14 Pauline Letters (including Hebrews), I John, and I Peter. The disputed writings are of two kinds:

Standards of Origen and Eusebius

(1) those known and accepted by many (James, Jude, II Peter, II and III John, and (2) those called "spurious" but not "foul and impious" (*Acts of Paul, Shepherd of Hermas, Apocalypse of Peter, Letter of Barnabas, Didachē* and possibly the *Gospel of the Hebrews*); finally there are the heretically spurious (*e.g., Gospel of Peter, Acts of John*). Revelation is listed both as fully accepted ("if permissible") and as spurious but not impious. It is important that Eusebius feels free to make authoritative use of the disputed writings. Thus canon and authoritative revelation are not yet the same thing.

Determination of the canon in the 4th century. Athanasius, a 4th-century bishop of Alexandria and a significant theologian, delimited the canon and settled the strife between East and West. On a principle of inclusiveness, both Revelation and Hebrews (as part of the Pauline corpus) were accepted. The 27 books of the New Testament—and they only—were declared canonical. In the Greek churches there was still controversy about Revelation, but in the Latin Church, under the influence of Jerome, Athanasius' decision was accepted. It is notable, however, that, in a mid-4th-century manuscript called Codex Sinaiticus, the *Letter of Barnabas* and the *Shepherd of Hermas* are included at the end but with no indication of secondary status, and that, in the 5th-century Codex Alexandrinus, there is no demarcation between Revelation and I and II Clement.

The Syriac canon

In the Syriac Church, Tatian's *Diatessaron* was used until the 5th century, and in the 3rd century the 14 Pauline Letters were added. Because Tatian had been declared a heretic, there was a clear episcopal order to have the four separated Gospels when, according to tradition, Rabbula, bishop of Edessa, introduced the Syriac version known as the Peshitta—also adding Acts, James, I Peter, and I John—making a 22-book canon. Only much later, perhaps in the 7th century, did the Syriac canon come into agreement with the Greek 27 books.

Developments in the 16th century. With the advent of printing and differences between Roman Catholics and Protestants, the canon and its relationship to tradition finally became fixed. During the Counter-Reformation Council of Trent (1545–63), the canon of the entire Bible was set in 1546 as the Vulgate, based on Jerome's Latin version. For Luther, the criterion of what was canonical was both apostolicity, or what is of an apostolic nature, and "was Christum treibet"—what drives toward, or leads to, Christ. This latter criterion he did not find in, for example, Hebrews, James, Jude, and Revelation; even so, he bowed to tradition, and placed these books last in the New Testament.

TEXTS AND VERSIONS

Textual criticism. *The physical aspects of New Testament texts.* To establish the reliability of the text of ancient manuscripts in order to reach the text that the author originally wrote (or, rather, dictated) involves the physical aspects of the texts: collection, collation of differences or variant readings in manuscripts, and comparison in matters of dating, geographical origins, and the amount of editing or revision noted, using as many copies as are available. Textual criticism starts thus with the manuscripts themselves. Families of manuscripts may be recognized by noting similarities and differences, degrees of dependence, or stages of their transmission leading back to the earliest text, or autograph. The techniques used in textual studies of ancient manuscripts are the same whether they deal with secular, philosophical, or religious texts. New Testament textual criticism, however, operates under unique conditions because of an abundance of manuscripts and the rather short gap between the time of original writing and the extant manuscripts, shorter than that of the Old Testament.

Significance of the large number of New Testament manuscripts

Compared with other ancient manuscripts, the text of the New Testament is dependable and consistent, but on an absolute scale there are far more variant readings as compared with those of, for example, classical Greek authors. This is the result, on the one hand, of a great number of surviving manuscripts and extant manuscript fragments and, on the other, of the fact that the time gap be-

tween an oral phase of transmission and the written stage was far shorter than that of many other ancient Greek manuscripts. The missionary message—the kerygma (proclamation)—with reports of the Passion, death, Resurrection, and Ascension of Jesus Christ and collections of his deeds and sayings was, at first, oral tradition. Later it was written down in Gospel form. The letters of Paul, Apostle to the Gentiles who founded or corresponded with churches, were also collected and distributed as he had dictated them. All autographs of New Testament books have disappeared. In sharp contrast to the fact that the oldest extant full manuscript of a work by the Greek philosopher Plato (died 347 BC) is a copy written in 895—a gap of more than 1,000 years bridged by only a few papyrus texts—there was a time gap of less than 200 or 300 years between the original accounts of the New Testament events and extant manuscripts. In fact, a small (about 2.5 inches by 3.5 inches [6.4 by 8.9 centimetres]) papyrus fragment with verses from the 18th chapter of the Gospel According to John can be dated *c.* 120–130; this earliest known fragment of the New Testament was written 40 years or less after the presumed date of the production of that Gospel (*c.* 90).

Excluding papyri found preserved in the dry sands, as in Egypt (where the Gospel According to John was evidently popular judging from the large number of fragments found there), the approximate number of New Testament manuscripts dating from the 3rd to 18th centuries are: 2,000 of the four Gospels; 400 of Acts, Pauline, and Catholic letters together; 300 of Pauline letters alone; 250 of Revelation; and 2,000 lectionaries—*i.e.,* collections of gospel (and sometimes Acts and letter) selections, or pericopes, meant to be used in public worship. Quotations from the Church Fathers—some of which are so extensive as to include almost the whole New Testament—account for more than 150,000 textual variants. Of the quotations in the Fathers, however, it is difficult to make judgments because the quotations may have been intended to be exact from some particular text traditions, but others may have been from memory, conflations, harmonizations, or allusions. Of the many New Testament manuscripts to date, however, only about 50 contain the entire 27 books of the New Testament. The majority have the four Gospels, and Revelation is the least well attested. Prior to the printing press (15th century), all copies of Bibles show textual variations.

Types of writing materials and methods. In Hellenistic times (*c.* 300 BC–*c.* AD 300), official records were often inscribed on stone or metal tablets. Literary works and detailed letters were written on parchment or papyrus, though short or temporary records were written or scratched on potsherds (ostraca) or wax tablets. Scrolls were made by gluing together papyrus sheets (made from the pith of the papyrus reed) or by sewing together parchment leaves (made from treated and scraped animal skins); they were written in columns and read by shifting the roll backward and forward from some wooden support on one or both ends. Such scrolls were used for literary or religious works and seldom exceeded 30 feet (nine metres) in length because of their weight and awkwardness in handling.

Scrolls and codices

In contrast, the church used not scrolls but the codex (book) form for its literature. A codex was formed by sewing pages of papyrus or parchment of equal size one upon another and vertically down the middle, forming a quire; both sides of the pages thus formed could be written upon. In antiquity, the codex was the less honourable form of writing material, used for notes and casual records. The use of the book form testifies to the low cultural and educational status of early Christianity—and, as the church rose to prominence, it brought "the book" with it. Not until the time of the Roman emperor Constantine in the 4th century, when Christianity became a state religion, were there parchment codices containing the whole New Testament.

Some very early New Testament manuscripts and fragments thereof are papyrus, but parchment, when available, became the best writing material until the advent of printing. The majority of New Testament manuscripts from the 4th to 15th centuries are parchment codices. When parchment codices occasionally were deemed no longer of use, the writing was scraped off and a new text written upon it. Such a rewritten (*rescriptus*) manuscript is called a palimpsest (from the Greek *palin,* "again," and *psaō,* "I scrape"). Often the original text of a palimpsest can be discerned by photographic process.

In New Testament times there were two main types of Greek writing: majuscules (or uncials) and minuscules. Majuscules are all capital (uppercase) letters, and the word uncial (literally, 1/12 of a whole, about an inch) points to the size of their letters. Minuscules are lowercase manuscripts. Both uncials and minuscules might have ligatures making them into semi-connected cursives. In Greco-Roman times minuscules were used for the usual daily writing. In parchments from the 4th to the 9th centuries, both majuscules and minuscules were used for New Testament manuscripts, but by the 11th century all the manuscripts were minuscules.

In these early New Testament manuscripts, there were no spaces between either letters or words, rarely an indication that a word was "hyphenated," no chapter or verse divisions, no punctuation, and no accents or breathing marks on the Greek words. There was only a continuous flow of letters. In addition, there were numerous (and sometimes variable) abbreviations marked only by a line above (*e.g.,* $\overline{\text{IC}}$ for IHCOUC, or Jesus, and $\overline{\text{KC}}$ for *kyrios,* or Lord. Not until the 8th–9th century was there any indication of accents or breathing marks (both of which may make a difference in the meaning of some words); punctuation occurred sporadically at this period; but not until the Middle Ages were the texts supplied with such helps as chapters (*c.* 1200) and verses (*c.* 1550).

Problems of the continuous flow of letters

Occasionally, the parchment was stained (*e.g.,* purple), and the ink was silver (*e.g.,* Codex Argenteus, a 5th–6th-century Gothic translation). Initial letters were sometimes illuminated, often with red ink (from which comes the present English word rubric, based on the Latin for "red," namely *ruber*).

Types of manuscript errors. Since scribes either copied manuscripts or wrote from dictation, manuscript variants could be of several types: copying, hearing, accidental, or intentional. Errors in copying were common, particularly with uncial letters that looked alike. In early manuscripts OC (for *hos,* "[he] who"), for example, might easily be mistaken for the traditional abbreviation of God: $\overline{\Theta\text{C}}$ (for ΘEOC, *theos*). Dittography (the picking up of a word or group of words and repeating it) and haplography (the omission of syllables, words, or lines) are errors most apt to occur where there are similar words or syllables involved. In chapter 17, verse 15, of John, in one manuscript the following error occurs: "I do not pray that thou shouldest take *them from the* [world, but that thou shouldst keep *them from the*] evil one" becomes "I do not pray that thou shouldst take them from the evil one." This is obviously a reading that omitted the words between two identical ends of lines—*i.e.,* an error due to *homoioteleuton* (similar ending of lines).

Especially in uncial manuscripts with continuous writing, there is a problem of word division. An English example may serve to illustrate: GODISNOWHERE may be read "God is now here" or "God is nowhere." Internal evidence from the context can usually solve such problems. Corrections of a manuscript either above the line of writing or in the margin (and also marginal comments) may be read and copied into the text and become part of it as a gloss.

Errors of hearing are particularly common when words have the same pronunciation as others but differ in spelling (as in English: "their, there"; "meet, meat"). This kind of error increased in frequency in the early Christian Era because some vowels and diphthongs lost their distinctive sound and came to be pronounced alike. For example, the Greek vowels *ē, i,* and *u* and the diphthongs *ei, oi,* and *ui* all sounded like the ēē (as in "feet"). Remarkable mistranslations can occur as, for example, in I Corinthians, chapter 15, verse 54: "Death is swallowed up in victory"—becomes by itacism (pronunciation of the Greek letter ē) "Death is swallowed up in conflict" (*neikos*). Another problem of itacism is the distinction between declensions of the 1st and 2nd persons in the plural ("we" and "you") in Greek,

which can sound the same (*hemeis,* "we"; *humeis,* "you"), because the initial vowels are not clearly differentiated. Such errors can cause interpretative difficulties.

A different category of error occurs in dictation or copying, when sequences of words, syllables, or letters in a word are mixed up, synonyms substituted in familiar passages, words read across a two- (or more) column manuscript instead of down, or assimilated to a parallel. Intentional changes might involve corrections of spelling or grammar, harmonizations, or even doctrinal emendations, and might be passed on from manuscript to manuscript. Paleographers—*i.e.,* scientists of ancient writing—can note changes of hands in manuscript copying or the addition of new hands such as those of correctors of a later date.

Significance and reliability of paleography

Paleography, a science of dating manuscripts by typological analysis of their scripts, is the most precise and objective means known for determining the age of a manuscript. Script groups belong typologically to their generation; and changes can be noted with great accuracy over relatively short periods of time. Dating of manuscript material by a radioactive-carbon test requires that a small part of the material be destroyed in the process; it is less accurate than dating from paleography.

Attempts to approximate an original manuscript and critical scholarship. Textual criticism of the Greek New Testament attempts to come as near as possible to the original manuscripts (which did not survive), based on reconstructions from extant manuscripts of various ages and locales. Assessment of the individual manuscripts and their relationships to each other can produce a fairly reliable text from various readings that may have been the result of copying and recopying of manuscripts. It is not always age that matters. Older manuscripts may be corrupt, and a reading in a later manuscript may in reality be ancient. No single witness or group of witnesses is reliable in all its readings.

Erasmus' edition

When Erasmus, the Dutch Humanist, prepared the Greek text for the first printed edition (1516) of the New Testament, he depended on a few manuscripts of the type that had dominated the church's manuscripts for centuries and that had had its origin in Constantinople. His edition was produced hastily, he even translated some parts for which he did not have a Greek text from Jerome's Latin text (Vulgate). In about 1522 Cardinal Francisco Jiménez, a Spanish scholarly churchman, published his Complutensian Polyglot at Alcalá (Latin: Complutum), Spain, a Bible in which parallel columns of the Old Testament are printed in Hebrew, the Vulgate, and the Septuagint (LXX), together with the Aramaic Targum (translation or paraphrase) of Onkelos to the Pentateuch with a translation into Latin. The Greek New Testament was volume 5 of this work, and the text tradition behind it cannot be determined with any accuracy. During the next decades new editions of Erasmus' text profited from more and better manuscript evidence and the printer Robert Estienne of Paris produced in 1550 the first text with a critical apparatus (variant readings in various manuscripts). This edition became influential as a chief witness for the *Textus Receptus* (the received standard text) that came to dominate New Testament studies for more than 300 years. This *Textus Receptus* is the basis for all the translations in the churches of the Reformation, including the King James Version.

Large extensive New Testament critical editions prepared by the German scholars C. von Tischendorf (1869–72) and H. von Soden (1902–13) had Sigla (signs) for the various textual witnesses; they are complex to use and different from each other. The current system, a revision by an American scholar, C.R. Gregory (adopted in 1908), though not uncomplicated has made uniform practice possible. A more pragmatic method of designation and rough classification was that of the Swiss scholar J.J. Wettstein's edition (1751–52). His textual apparatus was relatively uncomplicated. He introduced the use of capital Roman, Greek, or Hebrew letters for uncials and Arabic numbers for minuscules. Later, a Gothic P with exponents came into use for papyri and, in the few cases needed, Gothic or Old English O and T with exponents for ostraca and talismans (engraved amulets). Lectionaries are usually designated by an italicized lowercase *l* with exponents in Arabic numbers.

Known ostraca—*i.e.,* broken pieces of pottery (or potsherds) inscribed with ink—contain short portions of six New Testament books and number about 25. About nine talismans date from the 4th to 12th centuries; they are good-luck charms with a few verses on parchment, wood, or papyrus. Four of these contain the Lord's Prayer. These short portions of writing, however, are hardly of significance for a study of the New Testament textual tradition.

Texts and manuscripts. In referring to manuscript text types by their place of origin, one posits the idea that the major centers of Christendom established more or less standard texts: Alexandria; Caesarea and Antioch (Eastern); Italy and Gallia plus Africa (Western); Constantinople, the home for the Byzantine text type or the *Textus Receptus.* While such a geographical scheme has become less accurate or helpful, it still serves as a rough classification of text types.

Characteristics of major codices

Uncials. The main uncials known in the 17th and 18th centuries were: A, D, D^p, E^a, and C.

A, Codex Alexandrinus, is an early-5th-century manuscript containing most of the New Testament but with lacunae (gaps) in Matthew, John, and II Corinthians, plus the inclusion of the extracanonical I and II Clement. In the Gospels, the text is of the Byzantine type, but, in the rest of the New Testament, it is Alexandrian. In 1627 the A uncial was presented to King Charles I of England by the Patriarch of Constantinople; it has been in the British Museum, in London, since 1751.

D, Codex Bezae Cantabrigiensis, is a 5th-century Greco-Roman bilingual text (with Greek and Latin pages facing each other). D contains most of the four Gospels and Acts and a small part of III John and is thus designated D^{ea} (e, for *evangelia,* or "gospels"; and a for *acta,* or Acts). In Luke, and especially in Acts, D^{ea} has a text that is very different from other witnesses. Codex Bezae has many distinctive longer and shorter readings and seems almost to be a separate edition. Its Acts, for example, is one-tenth longer than usual. D represents the Western text tradition. D^{ea} was acquired by Theodore Beza, a Reformed theologian and classical scholar, in 1562 from a monastery in Lyon (in France). He presented it to the University of Cambridge, England, in 1581 (hence, Beza Cantabrigiensis).

D^p, Codex Claromontanus, of the same Western text type although not remarkably dissimilar from other known texts, contains the Pauline Letters including Hebrews. D^p (p, for Pauline epistles) is sometimes referred to as D_2. Beza acquired this 6th-century manuscript at about the same time as D^{ea}, but D^p was from the Monastery of Clermont at Beauvais (hence, Claramontanus). It is now in the Bibliothèque Nationale, in Paris.

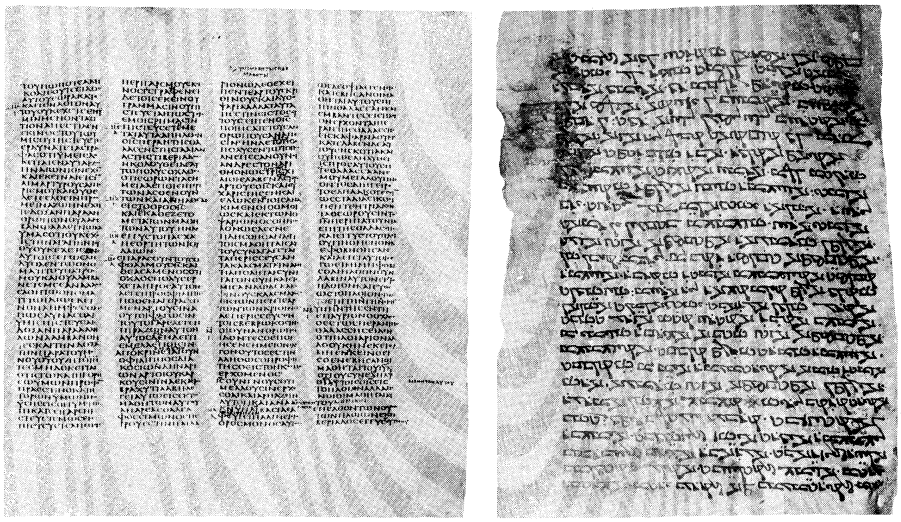
E^a, Codex Laudianus, is a bilingual Greco-Latin text of Acts presented in 1636 by Archbishop Laud, an Anglican churchman, to the Bodleian Library at Oxford. It is a late-6th- or early-7th-century manuscript often agreeing with D^{ea} and its Western readings but also having a mixture of text types, often the Byzantine.

C, Codex Ephraemi Syri rescriptus, is a palimpsest. Originally written as a biblical manuscript in the 5th century, it was erased in the 12th century, and the treatises or sermons of Ephraem Syrus, a 4th-century Syrian Church Father, were written over the scraped text. It was found *c.* 1700 by the French preacher and scholar Pierre Allix; and Tischendorf, using chemical reagents, later deciphered the almost 60 percent of the New Testament contained in it, publishing it in 1843. The text had two correctors after the 5th century but is, on the whole, Byzantine and reflects the not too useful common text of the 9th century.

Although there are numerous minuscules (and lectionaries), their significance in having readings going back to the first six centuries AD was not noted until textual criticism had become more refined in later centuries.

Newer uncial codices

The main uncials and some significant minuscules that were discovered and investigated in the 19th century changed the course of the textual criticism and led the way to better manuscript evidence and methods of dealing with it. This has continued into the 20th century. The



(Left) Gospel According to John 5:38–6:24, from the Codex Sinaiticus. In the British Museum. (Right) Palimpsest in which verses from chapter 5 of the Gospel According to John are overwritten with the story of St. Pelagia in Syriac, from the Codex Palimpsestus Sinaiticus.

By courtesy of the (left) trustees of the British Museum, (right) Cambridge University Press; photograph, John Ray

main new manuscript witnesses are designated ℵ or S, B, W, and Θ.

ℵ or S, Codex Sinaiticus, was discovered in 1859 by Tischendorf at the Monastery of St. Catherine at the foot of Mt. Sinai (hence, Sinaiticus) after a partial discovery of 43 leaves of a 4th-century biblical codex there in 1844. Though some of the Old Testament is missing, a whole 4th-century New Testament is preserved, with the *Letter of Barnabas* and most of the *Shepherd of Hermas* at the end. There were probably three hands and several later correctors. Tischendorf convinced the monks that giving the precious manuscript to Tsar Alexander II of Russia would grant them needed protection of their abbey and the Greek Church. Tischendorf subsequently published ℵ (S) at Leipzig and then presented it to the Tsar. The manuscript remained in Leningrad until 1933, during which time the Oxford University Press in 1911 published a facsimile of the New Testament from photographs of the manuscript taken by Kirsopp Lake, an English biblical scholar. The manuscript was sold in 1933 by the Soviet regime to the British Museum for £100,000. The text type of ℵ is in the Alexandrian group, although it has some Western readings. Later corrections representing attempts to alter the text to a different standard probably were made about the 6th or 7th century at Caesarea.

B, Codex Vaticanus, a biblical manuscript of the mid-4th century in the Vatican Library since before 1475, appeared in photographic facsimile in 1889–90 and 1904. The New Testament lacks Hebrews from chapter 9, verse 14, on, the Pastorals, Philemon, and Revelation. Because B has no ornamentation, some scholars think it slightly older than ℵ. Others, however, believe that both B and ℵ, having predominantly Alexandrian texts, may have been produced at the same time when Constantine ordered 50 copies of the Scriptures. As an early representation of the Alexandrian text, B is invaluable as a most trustworthy ancient Greek text.

W, Codex Washingtonianus (or Freerianus), consists of the four Gospels in the so-called Western order (Matthew, John, Luke, and Mark, as D^{ca}). It was acquired in Egypt by C.L. Freer, an American businessman and philanthropist (hence, the Freer-Gospels), in 1906 and is now in the Freer Gallery of Art of the Smithsonian Institution, in Washington, D.C. Codex Washingtonianus is a 4th–5th-century manuscript probably copied from several different manuscripts or textual families. The Byzantine, Western (similar to Old Latin), Caesarean, and Alexandrian text types are all represented at one point or another. One of the most interesting variant readings is a long ending to the Gospel According to Mark following a reference to the risen Christ (not found in most manuscript traditions).

Θ, Codex Koridethianus, is a 9th-century manuscript taking its name from the place of the scribe's monastery, Koridethi, in the Caucasus Mountains, near the Caspian Sea. Θ contains the Gospels; Matthew, Luke, and John have a text similar to most Byzantine manuscripts, but the text of Mark is similar to the type of text that Origen and Eusebius used in the 3rd–4th centuries, a Caesarean type. The manuscript is now in Tiflis, capital city of the Georgian S.S.R.

Minuscules. Although there are many minuscules, most of them come from the 9th century on; a few, however, shed significant light on earlier readings, representing otherwise not well attested texts or textual "families." In the early 20th century, the English scholar Kirsopp Lake (hence, Lake group) discovered a textual family of manuscripts known as Family 1:1, 118, 131, and 209 (from the 12th to 14th centuries) that have a text type similar to that of Θ, a 3rd–4th-century Caesarean type. At the end of the 19th century, W.H. Ferrar, a classical scholar at Dublin University (hence, the Ferrar group), found that manuscripts 13, 69, 124, and 346—and some minuscules discovered later (from the 11th to 15th centuries)—also seemed to be witnesses to the Caesarean text type. Manuscript 33, the "Queen of the Cursives," is a 9th–10th-century manuscript now at the Bibliothèque Nationale, in Paris; it contains the whole New Testament except Revelation and is a reliable witness to the Alexandrian text (similar to B) but, in Acts and the Pauline Letters, shows influence of the Byzantine text type.

Textual families

Lectionaries range from the 5th to the 6th century on; some early ones are uncials, though many are minuscules. Scholarly work with lectionary texts is only at its beginning, but the textual types of lectionaries may preserve a textual tradition that antedates its compilation and serves to give examples of the various text forms.

Papyri. The earliest New Testament manuscript witnesses (2nd–8th centuries) are papyri mainly found preserved in fragments in the dry sands of Egypt. Only in the recent decades of this century have the relatively recently discovered New Testament papyri been published. Of those cataloged to date, there are about 76 New Testament manuscripts with fragments of various parts of the New Testament, more than half of them being from the 2nd to 4th centuries. All the witnesses prior to 400 are of Egyptian provenance, and their primitive text types, though mainly Alexandrian, establish that many text types existed and developed side by side. One of the most significant papyrus finds is p^{52}, from *c.* 130 to 140, the earliest extant manuscript of any part of the New Testament. P^{52} consists of a fragment having on one side John 18:31–33 and on the other John 18:37–38, indicating that it was a codex, of which the text type may be Alexandrian. It is now in the John Rylands Library at Manchester.

In the early 1930s, British mining engineer A. Chester Beatty acquired three 3rd-century papyri from Egypt; they were published in 1934–37. Known as p^{45}, p^{46}, and p^{47}, they are, for the most part, in his private library in Dublin.

The Chester Beatty papyri

P^{45}, Beatty Biblical Papyrus I (and some leaves in Vienna), contains 30 leaves of an early- or mid-3rd-century codex of Matthew, Mark, Luke, John, and Acts. Each Gospel is of a different text type, and, although the leaves are mutilated, the Alexandrian text appears to predominate (particularly in Acts, in which a short non-Western text prevails); the whole may be thought of as pre-Caesarean.

P^{46}, Beatty Biblical Papyrus II (and Papyrus 222 at the University of Michigan), consists of 86 leaves of an early-3rd-century (*c.* 200) codex quire containing the Pauline Letters in the following order: Romans, Hebrews, I and II Corinthians, Ephesians, Galatians, Philippians, Colossians, and I Thessalonians. Although some of the leaves are quite mutilated, the text type of p^{46} appears to be Alexandrian. P^{47}, Beatty Biblical Papyrus III, is from the late 3rd century. It contains Rev. 9:10–17:2. It is the oldest, but not the best, text of Revelation and agrees with A, C, and ℵ.

Other early significant papyri are: p^{66}, p^{48}, p^{72}, p^{75}, and p^{74}. P^{66}, also known as Papyrus Bodmer II, contains in 146 leaves (some of which have lacunae) almost all of the Gospel According to John, including chapter 21. This codex, written before 200, is thus merely one century removed from the time of the autograph, the original text. Its text, like that of p^{45}, is mixed, but it has elements of an early Alexandrian text. P^{66} and the other Bodmer papyri, which Martin Bodmer, a Swiss private collector, acquired from Egypt, were published 1956–61. They are in the private Bodmer library at Cologny, near Geneva. P^{48} is a late-3rd-century text of Acts now in a library in Florence. It contains Acts 23:11–17, 23–29 and illustrates a Greek form of the Western text in Egypt in the 3rd century. The papyri of p^{72}, Papyri Bodmer VII and VIII, are also from the 3rd century. VII contains a manuscript of Jude in a mixed text, and VIII contains I and II Peter. In I Peter the Greek was written by a scribe whose native language was Coptic; there are numerous examples of misspellings and itacisms that when corrected leave a text similar to the Alexandrian witnesses.

The papyri of p^{75}, Papyri Bodmer XIV and XV, are 2nd–3rd-century codices containing most of Luke and of John, with John connected to Luke on the same page (unlike the Western order of the Gospels). The text coincides most with B but also has affinities with p^{66} and p^{45} as a predecessor of Alexandrian form.

P^{74}, Bodmer Papyrus XVII, is a 6th–7th-century text of Acts and the Catholic Letters. Acts show affinities with ℵ and A and no parallels with the Western text.

These and other papyri witness to the state of the early text of the New Testament in Egypt, indicating that no *one* text dominated and that text types of different origin flourished side by side.

Versions. *Early versions.* Even with all these witnesses, there remain problems in the Greek text. These include variants about which there is no settled opinion and some few words for which no accurate meaning can be found because they occur only once in the New Testament and not in prior Greek works. Very early translations of the New Testament made as it spread into the non-Greek-speaking regions of the missionary world, the so-called early versions, may provide evidence for otherwise unknown meanings and reflections of early text types.

The Koine, Latin, Syriac, Coptic, and other versions

In the Eastern half of the Mediterranean, Koine (common, vernacular) Greek was understood, but, elsewhere, other languages were used. Where Roman rule dominated, Latin came into use—in North Africa, perhaps in parts of Asia Minor, Gaul, and Spain (*c.* 3rd century). Old Latin versions had many variants, and these translations, traditionally known as the *Itala,* or Old Latin (O.L.), are designated in small letters of the Roman alphabet. The African versions were further from the Greek than were those made in Europe.

In dealing with the New Testament, Jerome prepared a Latin recension of the Gospels using a European form of the Old Latin and some Greek manuscripts. Though the completed Latin translation at the end of the 4th century was produced by no one editor or compiler—a commonly accepted Latin text, the Vulgate, emerged. A reworked official critical edition was a concern of the Council of Trent (1545–63), and in 1592 the Clementine Vulgate, named after Pope Clement VIII, became the authoritative edition. Since Vatican II (1962–65), an ecumenical group of biblical scholars using the best available manuscript witnesses has been engaged in the preparation of a critically sound revision of the Vulgate.

At Edessa (in Syria) and western Mesopotamia neither Latin nor Greek was understood. Therefore, Syriac (a Semitic language related to Aramaic) was used. Old Syriac was probably the original language of the *Diatessaron* (2nd century), but only fragments of Old Syriac manuscripts survive. The Peshitta (common, simple) Syriac (known as syr^{pesh}) became the Syrian 22-book Vulgate of the New Testament, and, at the end of the 4th century, its text was transmitted with great fidelity. The Philoxenian (syr^{phil}) and Harclean (syr^{harc}) versions followed in the 6th–7th centuries and contained all 27 of the New Testament books. The Palestinian (similar to Palestinian Aramaic) Syriac (syr^{pal}) may date to the 5th century but is known chiefly from 11th- to 12th-century lectionaries and is quite independent of other Syriac versions, reflecting a different text type.

In Egypt, in the later Hellenistic period, the New Testament was translated into Coptic—in the south (Upper Egypt) the Sahidic (cop^{sah}), and in the north (Lower Egypt) the Bohairic (cop^{boh}), the two principal dialects. By the 4th century, the Sahidic version was known, and the Bohairic somewhat later. The Coptic versions are fairly literal and reflect a 2nd–3rd-century Alexandrian Greek text type with some Western variants.

A Gothic version was made from the Byzantine text type by a missionary, Ulfilas (late 4th century); an Armenian version (5th century) traditionally was believed to have been made from the Syriac but may have come from a Greek text. Related perhaps to the Armenian was a Georgian version; and an Ethiopic version (*c.* 6th–7th century) was influenced by both Coptic and later Arabic traditions. In the various versions there is evidence of geographical spread, of the history of the underlying text traditions used, and of how they were interpreted in the early centuries.

The many readings in the Greek, Latin, and Syriac Fathers, who can be dated and located, can, to some extent, shed light on the underlying New Testament texts they quoted or used.

Another use both of the versions and of the patristic quotations is elucidation of the meaning of hitherto unknown Greek words in the New Testament.

Elucidation of meanings of unknown Greek words

An example is *epiousios* in the Lord's Prayer as given in verse 11 of chapter 6 of Matthew and verse 3, chapter 11, of Luke. The traditional translation in the Western Church is "daily" (referring to bread). From the Old Latin, Jerome, the early Syriac versions, and a retroversion of the Lord's Prayer into a proposed Aramaic substratum, the meaning is either "daily" or, more likely, "for the morrow"; and modern translations include this meaning in footnotes, including the suggestion that it may refer to eucharistic bread. The Greek is possibly a coined compound word that, on the basis of its component parts, yields "for the morrow" or "that which is coming soon." Such latter treatment is not conjectural emendation but rather creative analysis in context, where no Greek variants help. The biblical scholar, in possession of many variants, usually uses conjecture only as a means of last resort, and any conjecture must be both intrinsically suitable and account for the reading considered corrupt in the transmitted text.

Later and modern editions. New Testament editions in the 18th century did not question the *Textus Receptus (T.R.),* despite new manuscript evidence and study, but its limitations became apparent. E. Wells, a British mathematician and theological writer (1719), was the first to edit a complete New Testament that abandoned the *T.R.* in favour of more ancient manuscripts; and English scholar Richard Bentley (1720) also tried to go back to early manuscripts to restore an ancient text, but their work

Modern criteria and classifications of texts

was ignored. In 1734 J.A. Bengel, a German Lutheran biblical theologian, stressed the idea that not only manuscripts but also families of manuscript traditions must be differentiated, and he initiated the formulation of criteria for text criticism. J.J. Wettstein's edition (1730–51) had a wealth of classical and rabbinic quotations, but his theory on text was better than the text itself. A German Lutheran theologian, J.S. Semler (1767), further refined Bengel's classification of families.

J.J. Griesbach (1745–1812), a German scholar and student of Semler, adapted the text-family classification to include Western and Alexandrian text groups that preceded the Constantinopolitan groupings. He cautiously began to alter texts according to increasingly scientific canons of text criticism. These are, with various refinements, still used, as, for example, that "the difficult is to be preferred to the easy reading," and "the shorter is preferable to a longer"—both of which reason (with many other factors) that correction, smoothing, or interpretation leads to clearer and longer readings.

In the 19th century, classical philologist Karl Lachmann's critical text (1831) bypassed the *T.R.*, using manuscripts prior to the 4th century. C. von Tischendorf's discovery of ℵ (S) and his New Testament text (8th edition, 1864) collated the best manuscripts and had the richest critical apparatus thus far.

Two English biblical scholars, B.F. Westcott and F.J.A. Hort of Cambridge, using ℵ and B, brought out an edition in 1881–82 and classified the text witnesses into four groupings: Neutral (B, ℵ, the purest and earliest Eastern text); Alexandrian (a smoothed Neutral text as it developed in Alexandria); Western (D, Old Syrian, O.L., the Western Fathers with glosses that caused many readings to be rejected); and Syrian (A^c and the Byzantine tradition as it later developed). Such a "family tree" clearly showed the *T.R.* (Syrian) and, hence, the King James Version based upon it as an inferior text type; and the Revised Standard Version is based on such superior text types as B and ℵ.

Another critical edition (1902–13) was made by H. von Soden, a German biblical scholar who presupposed recensions to which all manuscripts can lead back. The importance of his work is in his enormous critical apparatus rather than in his theoretical groupings. B.H. Streeter, an English scholar, revised Westcott and Hort's classification in 1924. Basically, he challenged the concept of any uncontaminated descent from originals and made the observation (already alluded to in the evolution of papyrus evidence) that even the earliest manuscripts are of mixed text types. Yet, Streeter grouped texts in five families: Alexandrian, Caesarean, Antiochene, European Western, and African Western—parts of which all led into the Byzantine text and had become the *T.R.*

Unlikelihood of reconstructing an autograph

Despite grouping, it is clear that no reading backward from text families can reach an autograph. A strictly local text theory is useless in view of the papyrus evidence that there were no "unmixed" early texts. The use of external evidence cannot push beyond the boundary of the 3rd century. This insight brought about a new perspective. Only by using the canons of the internal evidence of readings can the best texts be determined, evaluating the variants from case to case—namely, the eclectic method. In modern times, therefore, the value of text families is primarily that of a step in the study of the history of the texts and their transmission. The eclectic method of reconstruction of an earliest possible New Testament text will yield the closest approximation of the historical texts put together into the New Testament canon. (For other, later and modern versions, see above *Old Testament canon, texts, and versions.*)

New Testament history

THE JEWISH AND HELLENISTIC MATRIX

The historical background of the New Testament and its times must be viewed in conjunction with the Jewish matrix from which it evolved and the Hellenistic (Greek cultural) world into which it expanded during a period of Jewish religious propaganda. It is difficult, however, to separate the phenomena of the Jewish and Hellenistic backgrounds, because the Judaism out of which the church arose was a part of a very Hellenized world. The conquests of Alexander the Great culminated in 331 BC, and the subtle but strong influence of Greek culture, language, and customs that was spread by his conquests united his empire. Jews in both Palestine and the Diaspora (Dispersion) were, however, affected by Hellenism, as in ideas of cosmic dualism and rich religious imagery derived in part from Eastern influence as a result of the Greek conquests. Greek words were transliterated into Hebrew and Aramaic even in connection with religious ideas and institutions as, for example, synagogue (religious assembly), Sanhedrin (religious court), and paraclete (advocate, intercessor). It could be argued that the very preoccupation with ancient texts and tradition and the interpretation thereof is a Hellenistic phenomenon. Thus, what may appear as the most indigenous element in the activity of the Jewish scribes, sages, and rabbis (teachers)—*i.e.*, textual scholarship—has its parallels in Hellenistic culture and is part of the general culture of the times. The thought worlds merged, confronted each other, and communicated with each other.

The Hasmonean kingdom. After Alexander's death the empire was split, and first the Ptolemies, an Egyptian dynasty, and then the Seleucids, a Syrian dynasty, held Palestine. Antiochus IV Epiphanes, a 2nd-century-BC Seleucid king, desecrated the Temple in Jerusalem; a successful Jewish revolt under the Maccabees, a priestly family, resulted in its purification and in freedom from Syrian domination in 164 BC. This began the Hasmonean (Maccabean) dynasty, which appropriated the powers both of king and of high priest. This reign, which created dissatisfaction on the part of other groups who considered their own claims falsely usurped, lasted until internecine strife brought it to an end. John Hyrcanus II, a 1st-century-BC Hasmonean king, appealed to Rome for help, and Pompey, a Roman general, intervened, bringing Palestine under Roman rule in 63 BC. John Hyrcanus, given the title of ethnarch, was later executed for treason (30 BC), thus ending the Hasmonean line, but Jewish independence had come to an end by Roman occupation.

Roman intervention in Palestine

Rule by the Herods. The Herods who followed were under the control of Rome. Herod the Great, son of Antipater of Idumaea, was made king of Judaea, having sided with Rome, and he ruled with Roman favour (37–4 BC). Though he was a good statesman and architect, he was hated by the Jews as a foreigner and semi-Jew. Jesus was born a few years before the end of his reign, and "the slaughter of the innocents," young children of Bethlehem who were killed as possible pretenders to Herod's throne, was attributed to Herod. After his death, Palestine was divided among three of his sons: Philip was made tetrarch of Iturea (the northeast quarter of the province) and ruled from 4 BC until AD 37. Herod Antipas became tetrarch of Galilee and Peraea until AD 39 and, like his father, was a builder, rebuilding Sepphoris and Tiberias before he was banished. Herod Antipas had John the Baptist beheaded and treated Jesus with contempt at Jesus' trial before him, before sending him back to Pontius Pilate, the Roman procurator (AD 26–36) at the time of Jesus' Crucifixion. Archelaus was made ethnarch of Judaea, Samaria, and Idumaea but was removed by AD 6 for his oppressive rule, and Judaea then became an imperial province, governed by procurators responsible to the emperor.

Two other Herods are mentioned in the New Testament: Agrippa I (called "Herod the king," AD 37–44) had James, the brother of John, killed and had Peter arrested; and the last of the Herods, Agrippa II, king of Trachonitis (*c.* AD 50–100), welcomed the procurator Festus (*c.* AD 60–62), who replaced Felix (*c.* AD 52–60) for the trial of Paul.

Roman occupation and Jewish revolts. In AD 66–70 there was a Jewish revolt while Nero was emperor of Rome (54–68). When he died and was succeeded by Vespasian, his former army commander (69–79), the siege and final destruction of Jerusalem occurred (AD 70). Before this event, Jewish Christians had fled, perhaps to Pella, and Yohanan ben Zakkai, a leading Jewish rabbi, with a group of rabbinical scholars, fled to Yavneh, where they established an academy that gave leadership to the Jews. Under the emperors Trajan (98–117) and Hadrian (117–

The effects of the fall of Jerusalem

138), Jews in Egypt and Mesopotamia rebelled and again fought unsuccessfully against Rome in Palestine for forbidding the practice of religious rites, and, under Simeon Bar Kokhba (or Bar Koziba), a Jewish revolutionary messianic figure, the final Jewish war was waged (132–135). After this defeat Jerusalem became a Roman colony; a temple to Jupiter was erected there, and Jews were prevented from entering the city until the 4th century.

When the Romans had entered Palestine in 63 BC, they practiced a relatively humane occupation until *c.* AD 66–70. They did not interfere with religious practices unless they considered them a threat to Rome, and their rights of requisition were precise and limited.

JEWISH SECTS AND PARTIES

From both the New Testament and extrabiblical material the main religious groups or parties in Palestinian Judaism may be discerned. Such descriptions, however, may be somewhat biassed or apologetic. Philo, an Alexandrian Jewish philosopher (died *c.* AD 40), Josephus, a Jewish apologist to the Romans (died *c.* 100), and sectarian writings found at Qumrān near the Dead Sea in 1947 that date back to about *c.* 200 BC and end about AD 70 all provide data about the respective Jewish religious groups in Palestine in the 1st century BC and the 1st century AD. The Pharisees (typically Jesus' opponents, although his ideas may have been close to their own), the Sadducees, and the Zealots are mentioned in the New Testament. The Essenes were described by Philo and Josephus, but new evidence from their own writings makes their group better understood (*i.e.,* the Dead Sea Scrolls from Qumrān).

The Pharisees. The Pharisees (possibly spiritual descendants of the Ḥasidim [Pious Ones], who were the exponents of Maccabean revolt) were strict adherents to the Law. Their name may come from *parush*—*i.e.,* "separated" from what is unclean, or what is unholy. They were deeply concerned with the Mosaic Law and how to keep it, and they were innovators in adapting the Law to new situations. They believed that the Law was for all the people and democratized it—even the priestly laws were to be observed by all, not only by the priestly class—so that they actually had a belief in a priesthood of all believers. They included Oral as well as Written Law in their interpretations. Though they did not accept the Roman occupation, they kept to themselves, and by pious acts, such as giving alms and burying the dead, they upheld the Law. Their interpretations of Law were sometimes considered casuistic because they believed they must find interpretations that would help all people to keep the Law. Their underlying hope was eschatological: in the day when Israel obeyed the Torah, the Kingdom would come. The Pharisees were called "smooth interpreters" by their opponents, but their hope was to find a way to make the living of the Law possible for all people. In their meal fellowship *(ḥavura)* they observed the laws strictly and formed a nucleus of obedient Israel. The Pharisees believed in the resurrection of the dead and had a developed angelology.

Pharisaic interpretations of the Law

The Sadducees. The Sadducees, more conservative and static, consisted mainly of the old priesthood and landed aristocracy and, perhaps, some Herodians. They were collaborators with Rome. They did not believe in resurrection because they found no Old Testament enunciation of such a doctrine. In a way, they seemed to respect the Pharisees in legal matters; but both the Pharisees—because they were a bourgeois rather than a popular movement—and the Sadducees—because they were aristocrats—rejected the *'am ha-aretz* (People of the Land), who were no party but simply the poor, common people whom they considered ignorant of the Law.

The Zealots. The Zealots were revolutionaries who plotted actively against the Roman oppression. That the Pharisees did not react in this way was perhaps because of their belief in Providence: what happens is the will of God, and their free will is expressed in the context of trust and piety in conjunction with an eschatological hope of winning God's Kingdom through obedience to Law.

The Essenes. Though the Essenes of the Dead Sea Scrolls are not mentioned in the New Testament, they are described by Philo, Josephus, and Eusebius, a 4th-century Christian historian. With publication of the Essenes' own sectarian writings since the 1950s, however, they have become well known. They did not have any really new ideas, but their founder, the Teacher of Righteousness, believed that he knew the interpretation of the prophets for his time in a way that was not even known to the prophets of their own day. Their withdrawal into desert seclusion was in opposition to the ruling powers in the city and the Temple of Jerusalem. They lived apart from society in constant study of the Scriptures and with a firm belief that they were the elect of Israel living in the end of days and to whom would come messianic figures—a messiah of David (royal) and a messiah of Aaron (priestly). Membership in their group and acceptance or rejection of its founder determined their place in the age to come. After a long period of probation and initiation, a man became a member of this elect community that had strict rules of community discipline that would seal or destroy his membership in their New Covenant. Ritual lustrations preceded most liturgical rites, the most important one of which was participation in a sacred meal—an anticipation of the messianic banquet, to which only the fully initiated members in good standing were admitted and which was presided over by representatives of the Davidic and Aaronic messiahs. From what is known of them, their communities were celibate, living "in the presence of the angels" and thus required to be in a state of ritual purity. Their laws were strict, their discipline severe, and—unlike Pharisees, Sadducees, and Zealots—they were not simply different parties within Judaism but a separate eschatological sect. The Pharisees did have lodges and a common meal, but membership in the Pharisaic party did not, as it did with the Essenes, guarantee a place in the age to come; and the attitude of the Pharisees to a leader or founder was not, as it was to the Essenes, one of the bases on which such place could be attained. Thus, the Essenes—as the early Jewish Christians—were an eschatological Jewish sect. They believed that they alone, among those living in the end time, would be saved. The apocalypticism of the Essenes and the early Christians had many similarities, but the Christians had a higher eschatological intensity because they already knew who the Messiah would be when he came in the future at the Parousia (the "Second" Advent), and they also had a recollection of the earthly Jesus, knowledge of the risen Lord, and the gift of the Spirit upon the church. Both communities lived in an era wherein the cosmic battle of God versus Satan-Belial was taking place, but the Christian community already had the traditions of Jesus' victory over Satan and the experience of his Resurrection. Both Essenes and Christians were sects with tightly knit organizations, but the church had a historically based messiah. The Essenes probably were killed or forced to flee from their wilderness community *c.* AD 68, yet some of their ideas can still be traced in the ministry of John the Baptist (who might have been an Essene) and in the thought world of the New Testament (see also JUDAISM).

Eschatological views of the Essenes

Similarities and differences between Essenes and Christians

THE RELIGIOUS SITUATION IN THE GRECO-ROMAN WORLD OF THE 1ST CENTURY AD

Hellenistic religions. With the expansion of Christianity into the Hellenistic world either to Jews or increasingly to Gentiles, there were various reasons why the Christian message that spread, for example by Paul, met the needs of the Hellenistic Age and world. There was no lack of religions, but there was a crisis of upheaval, unrest, and uncertainty and a desire to escape from mortality and the domination of unbending fate. There was also a desire to win personal knowledge of the universe and a dignified status within it—*i.e.,* a religious identity crisis. City-states with their cults of civic gods were unstable, because men changed from place to place and the gods of the city were distant from individual needs and anxieties. After Alexander's conquests, the resulting religious syncretism did not meet individual needs and longings that were increasingly becoming conscious. Many Gentiles turned to Judaism, at least as "god fearers," and later to Christianity. There were also "mystery religions," the secrets of which were known only to the initiate, which may have arisen from Eastern

The religious crisis of the 1st century AD

fertility cults with their dying and rising gods and were transformed in the Hellenistic Age to cults of a saviour god whose dying and rising gives personal immortality. Such mystery cults often provided meaningful relationships with fellow initiates.

Astrology. There were elements in the Greek world that may have come from the East, partly Egyptian and Babylonian, which gave rise to astrology. The basic conviction of astrology was that the heavenly bodies were deities that in a direct way control life and events on earth. An older idea of *tychē*, or "fate," originally signified the chance element in the universe, a capriciousness that increased insecurity. Astrology transformed this into a fate or destiny in which everything is strictly regulated by celestial deities. Man's problem, then, is that of finding security from overwhelming powers outside human control. One way is to "read a horoscope." Because the heavenly deities are systematic and orderly according to astronomic observation, this order and regularity can be exploited to see how and in what way events will happen and can perhaps be used or avoided. Another way is to deal with such forces through magic. From the Hellenistic period many magical papyri with formulas for dealing with sicknesses, demons, and other adverse forces have been found. Magic attempts to manipulate and control what affects the world by a kind of participation in the event.

Philosophical solutions. Solutions were also sought in philosophy. Socrates, a 5th-century-BC Greek philosopher, was largely concerned with the search for the "good," the good life. After Plato and Aristotle, however, philosophical systems sought to supply man's longing for inward security and stability. These were sought not by an in-depth understanding of reality but by ad hoc constructions—a new dogmatism for providing infallible plans and attaining immediate security—that the age demanded. Those philosophies were crude constructions that gave shelter and were defended by an unyielding dogmatism as absolute truths; if they were proved false, they would remove their promised security. Epicureanism, founded by the Greek philosopher Epicurus (341–270 BC), was basically a philosophy of escape, and its goal was serenity and tranquillity, a negative concept characterized by absence of fear, pain, and struggle. Fate, providence, and the afterlife were eliminated to deny the anxieties they provoked in terms of control, reward, or judgment. Epicurus attempted to meet this crisis by adopting a completely material view of the universe, including the soul, and thereby eliminating interference by deities both in life and after death. He did believe in the gods; but they, too, lived in their own perfect tranquillity, away from the universe. The Epicurean was both self-reliant and at peace with the absence of pain. There was also emphasis on friendship and the development of close communities.

Goals of escape or acceptance

Zeno, a 3rd-century-BC philosopher, was the founder of Stoicism. Stoicism was a rule of life that held that all reality was material but was animated by a rational principle that was at the same time both the law of the universe and of the human soul. The wise man then could accept and learn to live a life in conformity to this permeating reason without letting anything affect him. He responded to duty and accepted it.

Cynicism was a philosophy that maintained a cosmic view of life with a method of dealing with crisis by reducing man's needs to a minimum. Later in the Hellenistic period, a group of Stoic–Cynic preachers arose and, in New Testament times, wandered around calling men to repent and change their lives from sin to virtue.

ADAPTATION OF THE CHRISTIAN MESSAGE TO THE HELLENISTIC RELIGIOUS SITUATION

The Christian message adapted itself to this Hellenistic situation of crisis and proved a successful answer: Jesus was proclaimed as Lord and Saviour, Baptism was practiced as a form of initiation and a passage from death to new life, and the Lord's Supper was celebrated as a sacral meal. The obvious difference between Christianity and the mystery religions is that a historical person, Jesus, forms the center of cult and devotion; his titles came from his Jewish background. Adaptation took place out of the Jewish matrix of Christianity—and Hellenistic terms that were meaningful were also used, such as illumination and regeneration. Such terms are not to be found in the earliest origins of Christianity but in the communication of the Christian message to a new environment. Among the religious and philosophic needs of the time was that of a cult that provided for the needs of the individual along with a community of worship. Christ as Lord was viewed as universal, and his teachings made the universe understandable, as well as providing a basis for ethics. In a period of expansion, all religions are to some extent syncretistic, as is the case of Christianity in the 2nd century. Such a phenomenon belongs to a religion in a time of strength. Though universal, however, Christ was believed to have an exclusive claim, and in this there was security and relief for the anxieties of the period. The church was more than a philosophy; it had a social and enduring structure. It also reached out to all men—not only to those regarded as the best of men. It called them to a new life and gave them a new home and community, the church.

THE LIFE OF JESUS

Though the fact that Jesus was a historical person has been stressed, significant, too, is the fact that a full biography of accurate chronology is not possible. The New Testament writers were less concerned with such difficulties than the person who attempts to construct some chronological accounts in retrospect. Both the indifference of early secular historians and the confusions and approximations attributable to the simultaneous use of Roman and Jewish calendars make the establishment of a chronology of Jesus' life difficult. That the accounts of Matthew and Luke do not agree is a further problem. Thus, only an approximate chronology may be reconstructed from a few somewhat conflicting facts. The points of reference are best taken from knowledge of the history of the times reflected in the passages.

Difficulties in establishing a chronology of the life of Jesus

According to Matthew, Jesus was born near the end of the reign of Herod the Great, thus before 4 BC. In Luke, chapter 2, verses 1 to 2, Jesus is said to have been born at the time of a census when Quirinius was governor of Syria. Such a census did occur, but in AD 6–7. Because this was after Herod's death and not in agreement with a possible date of Jesus' baptism, this late date is unlikely. There may have been an earlier census under another governor; an inscription in the Lateran Museum records an unnamed governor who twice ruled Syria, and the suggestion has been made that this was, indeed, Quirinius and that in an earlier time a reported census according to Roman calculation might have been carried out *c.* 8 BC, one of a series of such. With such speculation and the combined evidence of Matthew and Luke, an approximate year of birth might be 7–6 BC.

In Luke, chapter 3, verse 23, it is stated that Jesus' ministry began when he was about 30 years of age. This would not come within the dates of the procuratorship of Pontius Pilate (AD 26–36), and the age might simply approximate a term for Jesus' having arrived at maturity. In Luke several dates are implied to assist in dating the Baptism of Jesus: the 15th year of Tiberius (*c.* 29, according to his accession as co-emperor with Augustus), while Pontius Pilate was in office (during 26–36), while Herod Antipas was tetrarch (4 BC–AD 39) and Philip tetrarch (4 BC–AD 37). These limits make a speculation of Jesus' Baptism and the start of his ministry *c.* AD 27/28.

The duration of Jesus' ministry can be an average of the one year, as indicated in the Synoptic Gospels (Matthew, Mark, and Luke) or about three years as indicated in John, based on various cycles of harvests and festivals. This would be about two years. Because Jesus was crucified before 36 and his ministry started about 27/28, he then was crucified about AD 30 (see also JESUS).

THE CHRONOLOGY OF PAUL

For the chronology of Paul's ministry, there are also some extra-biblical data: According to Josephus, Herod Agrippa I was made ruler of all Palestine by the emperor Claudius in AD 41 and reigned for three years. His death was thus in AD 44. A famine in Claudius' reign took place when

Tiberius Alexander was procurator of Judaea (*c.* 46–48), and Egyptian papyri suggest (by reference to high wheat prices) that the date of the famine was about 46. The Gallio inscription at Delphi (in Greece) gives a date for Gallio, proconsul of Achaia when Paul was at Corinth. It notes that Claudius was acclaimed emperor for the 26th time. This would bring the date of being declared emperor to about 52 and Gallio's term of office (about one year) to about 51–52.

The chronology of Paul's missionary journeys

The chronology of Paul's missionary journeys and the dates of his letters have been the object of an investigation made difficult by the fact that the account in Acts does not agree with Paul's own letters, which are, of course, more reliable.

With the help of external references, some degree of absolute chronology might be sought—with several years' margin both because of uncertainty as to extra-biblical dating and much ambiguity about internal evidence. Although Paul would be in a better position to know his own situation, often his letters are, in their present form, combined fragments from various times (see below *The Second Letter of Paul to the Corinthians; The Letter of Paul to the Phillipians*). A chronology can be reached by comparing Paul's accounts of his journeys and sojourns with those reported in Acts. Given references in Acts and the Gallio inscription, it is possible to place Paul in Corinth in AD 51, and, since he was there for 18 months, it can be assumed that he began his missionary work sometime in 49 (he had previously been in Thessalonica and Philippi and in Troas and Asia Minor). This probably fits in with the "expulsion" of Jews from Rome about AD 49, thus indicating that Paul met Priscilla and Aquila, two Roman Jewish Christians, in Corinth at this time. This indicates that he was at an "apostolic conference" at Jerusalem sometime shortly before this (a comparison of chapters 13 and 15 of Acts with chapters 1 and 2 of Galatians shows that the author of Acts made two visits out of the one recorded by Paul), which was either in 49 or 48.

Though the dates in Galatians 1 and 2 are uncertain—not indicating whether they refer to 17 years *in toto* or only 14 years, because half years were equated with whole ones—they do establish the call of Paul to become a Christian in 31 or about 34–35. Working in the other direction, it is known that Paul wrote to the Thessalonians from Corinth, thus indicating a date of about 50 as probable for the writing of I Thessalonians.

From Corinth, Paul went to Ephesus, where, according to Acts, he remained (probably in prison) for three years. This would place him in Ephesus during the period 52–55, thus allowing time for a journey from Corinth via Ephesus to Antioch and then back to Ephesus. A sequence given in Acts, chapters 16 and 18, shows two possibilities for Paul to have been in Galatia that work in agreement with Galatians, chapter 4, verse 13, demonstrating that Galatians was written from Ephesus about 53–54. Ephesus can also be the location from which came I Cor., Phil., and probably Philem.

II Corinthians appears to have been written from Macedonia during 55. From the dating of the periods of Felix and Festus in office at Caesarea (mid-50s) and from the events in Felix' time of office, it is probable that Paul was in prison under Felix by 56.

Thus, data of Acts 18 and 20 regarding the journey and sojourn at Corinth can be correlated with data in Romans 15 to place the epistle to the Romans in about the year 56, before the journey back to Jerusalem, ending in the arrest of Paul in 56. The two years of Acts 24:27 can then be explained as the time during which Paul was in prison at Caesarea, so that in 58 Paul was before Festus and was sent to Rome.

That Paul was then in Rome for two more years is established in Acts chapter 28, verse 30. It can be concluded that Paul died sometime after 60, possibly during or before the Neronian persecution of 64 (*cf.* I Clem. 5). All this does not resolve the question of a possible Spanish journey nor give precise dates and locations for II Thessalonians, Colossians, Ephesians, or the Pastoral Letters (see also PAUL).

New Testament literature

INTRODUCTION TO THE GOSPELS

Meaning of the term gospel. From the late AD 40s and until his martyrdom in the 60s, Paul wrote letters to the churches that he founded or guided. These are the earliest Christian writings that the church has, and in them he refers to "the gospel" (*euangelion*). In Romans, chapter 1, verse 1, he says: "Paul, a servant of Jesus Christ, called to be an apostle, set apart for the gospel of God . . ." and goes on to describe this "gospel" in what was already by that time traditional language, such as: "promised beforehand through his prophets in the holy scriptures, the gospel concerning his Son, who was descended . . . our Lord" (Rom. 1:1–4). This gospel is the power of God for salvation to everyone who has faith ". . . for in it the righteousness of God is revealed through faith for faith . . ." (1:17). In I Corinthians Paul had reminded his congregation in stylized terms of "the gospel" he had brought to them. It consisted of the announcement that Jesus had died and risen according to the Scriptures.

Earliest concept of "gospel"

Thus, the "gospel" was an authoritative proclamation (as announced by a herald, *kēryx*), or the kerygma (that which is proclaimed, *kērygma*). The earthly life of Jesus is hardly noted or missed, because something more glorious—the ascended Lord who sent the Spirit upon the church—is what matters.

In the speeches of Peter in Acts, the transition from kerygma to creed or vice versa is almost interchangeable. In Acts 2 Jesus is viewed as resurrected and exalted at the right hand of God and made both Lord and Christ. In Acts 3 Peter's speech proclaims Jesus as the Christ having been received in heaven to be sent at the end of time as judge for the vindication and salvation of those who believe in him. Here the proclaimed message, the gospel, is more basic than an overview of Jesus' earthly life, which in Acts is referred to only briefly as "his acting with power, going about doing good, and healing and exorcising" (10:38ff.). Such an extended kergyma can be seen as a transition from the original meaning of gospel as the "message" to gospel meaning an account of the life of Jesus.

The term gospel has connotations of the traditions of Jesus' earthly ministry and Passion that were remembered and then written in the accounts of Matthew, Mark, Luke, and John. They are written from the post-Resurrection perspective and they contain an extensive and common Passion narrative as they deal with the earthly ministry of Jesus from hindsight. And so the use of the term gospel for Matthew, Mark, Luke, and John has taken the place of the original creedal–kerygmatic use in early Christianity. It is also to be noted that, in the Evangelists' accounts, their theological presuppositions and the situations of their addressees molded the formation of the four canonical Gospels written after the Pauline Letters. The primary affirmations—of Jesus as the Christ, his message of the Kingdom, and his Resurrection—preceded the Evangelists' accounts. Some of these affirmations were extrapolated backward (much as the Exodus event central in the Old Testament was extrapolated backward and was the theological presupposition for the patriarchal narratives in Genesis). These stories were shaped by the purpose for their telling: religious propaganda or preaching to inspire belief. The kerygmatic, or creedal, beginning was expanded with material about the life and teaching of Jesus, which a reverence for and a preoccupation with the holy figure of Jesus demanded out of loving curiosity about his earthly ministry and life.

Theological presuppositions that helped to shape the Gospels

The English word gospel is derived from the Anglo-Saxon *godspell* ("good story"). The classical Greek word *euangelion* means "a reward for bringing of good news" or the "good news" itself. In the emperor cult particularly, in which the Roman emperor was venerated as the spirit and protector of the empire, the term took on a religious meaning: the announcement of the appearance or accession to the throne of the ruler. In contemporary Greek it denoted a weighty, authoritative, royal, and official message.

In the New Testament, no stress can be placed on the etymological (root) meaning of *eu* ("good"); in Luke, chap-

ter 3, verse 18 (as in other places), the word means simply authoritative news concerning impending judgment.

Form criticism. In the Pauline writings, as noted above, gospel, kerygma, and creed come close together from oral to written formulas that were transmitted about the Christ event: Jesus' death and Resurrection. In the apostolic Fathers (early 2nd century), the transition was made from oral to written tradition; the translation of the presumed Aramaic traditions had taken place before the Gospel material had been committed to writing. By the time of Justin Martyr (*c.* 155), these writings were called Gospels and referred to in the plural; they contain the words, deeds, and Passion narratives—*i.e.,* the present four Gospels compiled and edited by the Evangelists according to their various needs and theological emphases. Justin also referred to these as "memoirs of the Apostles."

Patterns noted in the Gospels

Such a Gospel began with a missionary announcement concerning a cosmic divine figure, a man with divine characteristics who would bring salvation and hope to the world. The earthly historical Jesus, however, was the criterion of the proclamation—being both the content of the church's proclamation and the object of its faith.

The identification of basic patterns in the history of oral and written traditions—the stage of tradition prior to any literary form and particularly as the traditions passed from an oral to a written form—and the determination of their creative milieu, or their situations and functions in various places and under various circumstances, are tasks of form criticism. Through such study, small independent units may be isolated in a postulated more primitive form than they were before being incorporated into more extended accounts. The term *Sitz-im-Leben* refers to the "Sitz im Leben der Kirche"—*i.e.,* the situation in the life of the church in which the material was shaped and adjusted to the needs at hand. Only through such studies is it possible to progress tentatively to an assessment of a "Sitz im Leben Jesu."

Both Jews and Gentiles could use "biographies," often for propaganda purposes. Philo and Josephus recounted the wonderful lives and deeds of Old Testament heroes such as Moses; and there are miraculous tales of the prophets Elijah and Elisha told in order that faith might be inspired or justified. A miracle worker (*theios anēr,* "divine man") and stories about him comprised an aretalogy (from *aretē,* "virtue"; also manifestation of divine power, miracle). Aretalogies were frequently used to represent the essential creed and belief of a religious or philosophical movement. *The Life of Apollonius of Tyana,* a Neo-Pythagorean philosopher and wonder-worker (transmitted by the Greek writer Philostratus), was widely read. He was depicted as having performed miracles and as being possessed of divine cosmic power not as an exception but as an example to men who have the possibility of sharing such power (*cf.* Matt. 9:8). There were tales of Heracles, the Greek hero, and a whole literature of Alexander the Great as wonder-workers, divine men.

Types of forms and genres

Though the pericopes (small units) of which the Gospels are constituted include many forms, or genres, they are mainly divided into narratives (including legends, miracle stories, exorcisms, healings, and tales) and sayings (prophetic and apocalyptic sayings, proverbs and wisdom sayings, parables, church discipline and rules for the community, Christological sayings, such as the socalled "I am" sayings [*e.g.,* "I am the bread of life"] in John, revelations, and legal sayings). Some stories may simply be the background for a pithy saying; these latter are sometimes called paradigmatic sayings, and the pronouncement stories are their vehicles of transmission. The forms have many different names, but form criticism started with Homeric form analysis (taking oral tradition into account), which was applied to Old Testament studies by Hermann Gunkel, a German biblical scholar, and applied to the New Testament, on the basis of the German classical philologist Eduard Norden's stylistic studies, by such biblical scholars as Rudolf Bultmann and Martin Dibelius.

Form criticism asks and answers questions about what shaped the preliterary tradition and the earliest written traditions into blocks as they are found in the Gospels. This may be a historical context (as a missionary situation), a need for admonition (as church-discipline sections), or for the transmission of teaching in a faithful way (as in a "school," be it Matthean, Pauline, or Johannine). One large block of the material, however, is to all intents and purposes the same (although differing in details) in all four canonical Gospels: the Passion narrative. In the Synoptic Gospels there is also a basic nucleus in the sayings about Jesus that are mysterious, prophetic, and apocalyptic and that point to the significance of Jesus as the Christ who has come in history in the person of Jesus of Nazareth.

Such form-critical studies were centred on the smaller units of tradition (pericopes) that make up the Gospels, and their intention was partly to assess relative age and authenticity of such traditions. In more recent times the tools of form criticism have been applied to a more synthetic method that could be used to determine the relation between a genre of literature and the Christological and theological perspectives that made such genres natural. A presentation of Jesus material in the form of more or less disconnected sayings (as in the so-called Q Source, composed of independent sayings, behind Matthew and Luke, and in the *Gospel of Thomas;* see below *The two- and four-source hypotheses*) tends to fit a Christology in which Jesus is viewed as a teacher of Wisdom, an envoy of Wisdom, or as Wisdom herself. The collections of wonder stories (aretalogies) grew out of a Christology of Jesus as the divine man. Another type of Jesus material with independent existence seems to have been "revelations," or "apocalypses," in which Jesus Christ speaks to his followers. This is seen, for example, in Mark 13, I Thessalonians, chapter 4, the canonical book of Revelation to John, and the noncanonical *Didache* 16.

These genres of material now represented in the canonical Gospels are amply represented also in the noncanonical writings from the first Christian centuries. The discovery of a Gnostic library of Coptic writings at Naj̱ Ḥammādī, in Egypt, in the 1940s gave scholars a new opportunity to compare the canonical Gospels with the Jesus material of these various types, some of them having been called and used as gospels (such as the *Gospel of Thomas*). In the light of such a wider spectrum of material, it appears that the gospel form for which Mark is the earliest witness became a criterion for the orthodox transmission of the Christian message about Jesus. By making the confession of Jesus as the crucified and risen Lord (the earliest kerygma and "gospel" as found in Paul and Acts) the form of an extensive Passion account prefaced by a limited amount of narrative and teaching, Mark set the stage for a faith that anchored faith in Jesus Christ in the events of the earthly life of Jesus. This form of the "gospel" became the standard within which the other commonly accepted Gospels grew. It became the criterion for later creedal statements concerning Jesus Christ as true God and true man. By such a criterion, gospels that seemed to disregard his humanity (*e.g., Gospel of Thomas,* the *Gospel of Peter*) were judged heretical.

Significance of the Passion account

THE SYNOPTIC PROBLEM

Early theories about the Synoptic problem. Since the 1780s, Matthew, Mark, and Luke have been referred to as the Synoptic Gospels (from *synoptikos,* "seen together"). The extensive parallels in structure, content, and wording of Matthew, Mark, and Luke make it even possible to arrange them side by side so that corresponding sections can be seen in parallel columns. John Calvin, the 16th-century Reformer, wrote a commentary on these Gospels as a harmony. Such an arrangement is called a "synopsis," or Gospel harmony, and, by careful comparison of their construction, compilation, and actual agreement or disagreement in wording or content, literary- or source-critical relationships can be seen. Augustine, the great 4th–5th-century Western theologian, considered Mark to be an abridged Matthew, and, until the 19th century, some variation of this solution to literary dependency dominated the scene. It still recurs from time to time.

The Synoptic problem is one of literary or of source criticism and deals with the written sources after compilation and redaction. Matthew was the Gospel most used for the selections read in the liturgy of the church, and other

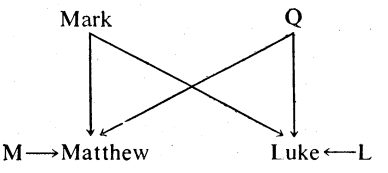
Gospels were used to fill in the picture. One attempted solution to the problem of priority was the proposed existence of an Aramaic primitive gospel, which is now lost, as the first Gospel from which a later Mark in Greek was translated and arranged. The Greek Mark would thus be first based on a prior Semitic Matthew, and later both Mark and Matthew would be translations dependent on Matthew, and Luke dependent on both. The preservation of an ecclesiastical priority of Matthew breaks down because of the literary word-for-word agreement in Matthew, Mark, and Luke. This agreement occurs to far too great an extent to be accounted for in translations and revisions, not to mention the agreement in the order of the various pericopes as they are viewed in a synoptic parallel arrangement.

For similar reasons, a fragment theory holding that the Gospels were constructed of small written collections brought together in varying sequences cannot stand the test of actual structure—but it has the merit of stressing compilation of sources.

In 1789 J.J. Griesbach, a German biblical scholar, hypothesized that the Synoptics had not developed independently, but in his "usage-hypothesis" he recognized that there must be literary dependency. He thought that Mark used Matthew as well as Luke, but this could not account for the close relationship of Matthew and Luke. His basic concept of literary dependency, however, paved the way for K. Lachmann, who observed in 1835 that Matthew and Luke agree only when they also agree with Mark and that, where material is introduced that is not in Mark, it is inserted in different places. This, it is held, can only be explained on the basis of the priority of Mark and its use as the patterning form of Matthew and Luke. This insight led to a so-called two-source hypothesis (by two German biblical scholars, Heinrich Holtzmann in 1863, and Bernhard Weiss in 1887–88), which, with various modifications and refinements of other scholars, is the generally accepted solution to the Synoptic problem.

The two- and four-source hypotheses. The two-source hypothesis is predicated upon the following observations: Matthew and Luke used Mark, both for its narrative material as well as for the basic structural outline of chronology of Jesus' life. Matthew and Luke use a second source, which is called Q (from German *Quelle,* "source"), not extant, for the sayings (logia) found in common in both of them. Thus, Mark and Q are the main components of Matthew and Luke. In both Matthew and Luke there is material that is peculiar to each of their Gospels; this material is probably drawn from some other sources, which may be designated M (material found only in Matthew's special source) and L (material found only in Luke's special source). This is known as the four-document hypothesis, which was elaborated in 1925 by B.H. Streeter, an English biblical scholar. The placement of Q material in Luke and Matthew disagrees at certain points according to the needs and theologies of the addressees of the gospels, but in Matthew the Marcan chronology is the basic scheme into which Q is put. Mark's order is kept, on the whole, by Matthew and Luke, but, where it differs, at least one agrees with Mark. After chapter 4 in Matthew and Luke, not a single passage from Q is in the same place. Q was a source written in Greek as was Mark, which can be demonstrated by word agreement (not possible, for example, with a translation from Aramaic, although perhaps the Greek has vestiges of Semitic structure form). A diagram might thus be:

The source called Q



In approximate figures, Mark's text has 661 verses, more than 600 of which appear in Matthew and 350 in Luke. Only *c.* 31 verses of Mark are found nowhere in Matthew or Luke. In the material common to all three Synoptics, there is very seldom verbatim agreement of Matthew and Luke against Mark, though such agreement is common between Matthew and Mark or Luke and Mark or where all three concur.

The postulated common saying source of Matthew and Luke, Q, would account for much verbatim agreement of Matthew and Luke when they include sayings absent from Mark. The fact that the sayings are used in different ways or different contexts in Matthew and Luke is an indication of a somewhat free way in which the editors could take material and mold it to their given situations and needs. An example of this is the parable in Matthew and Luke about the lost sheep (Matt. 18:10–14, Luke 15:3–7). The basic material has been used in different ways. In Matthew, the context is church discipline—how a brother in Christ who has lapsed or who is in danger of doing so is to be gently and graciously dealt with—and Matthew shapes it accordingly (the sheep has "gone astray"). In Luke, the parable exemplifies Jesus' attitude toward sinners and is directed against the critical Pharisees and scribes who object to Jesus' contact with sinners and outsiders (the sheep is "lost").

Variations in the use of certain sayings of Jesus

Another example of two passages used verbatim in Luke and Matthew is Jesus' lament over Jerusalem. In Luke (13:34–35; the lament over Jerusalem) Jesus refers to how they will cry "Blessed be the King who comes in the name of the Lord" when he enters Jerusalem (Lk. 19:38). In Luke, the passage is structured into the life of Jesus and refers to his triumphal entry into Jerusalem, "Blessed is he who comes in the name of the Lord"). In Matthew (23:37–39) this same lament is placed after the entry into the city (21:9) and thus refers to the fall of Jerusalem and the Last Judgment. Apparently, Luke has historicized a primarily eschatological saying.

Since the 1930s, scholars have increasingly refined sources, postulated sources behind sources, and many stages of their formation. The premise of the two- (or four-) source hypothesis is basic and provides information as to literary sources; further refinement is of interest only to the specialist. Another movement in synoptic research—and also research including John—is that which concentrates rather on the treatment of gospels as a whole, formally and theologically, with patterns or cycles to be investigated. It may be significant that the latest and best regarded Greek synopsis is that of the German scholar Kurt Aland, *Synopsis Quattuor Evangeliorum* (1964; *Synopsis of the Four Gospels,* 1972), which includes the Gospel According to John and, as an appendix, the *Gospel of Thomas,* as well as ample quotations from noncanonical gospels and Jesus' sayings preserved in the Church Fathers.

THE SYNOPTIC GOSPELS

The Gospel According to Mark. The Gospel According to Mark is the second in canonical order of the Gospels and is both the earliest gospel that survived and the shortest. Probably contemporaneous with Q, it has no direct connection with it. The Passion narrative comprises 40 percent of Mark, and, from chapter 8, verse 27, onward, there is heavy reference forward to the Passion.

Though the author of Mark is probably unknown, authority is traditionally derived from a supposed connection with the Apostle Peter, who had transmitted the traditions before his martyr death under Nero's persecution (*c.* 64–65). Papias, a 2nd-century bishop in Asia Minor, is quoted as saying that Mark had been Peter's amanuensis (secretary) who wrote as he remembered (after Peter's death), though not in the right order. Because Papias was from the East, perhaps the Johannine order would have priority, as is the case in the structure of the Syrian scholar Tatian's *Diatesseron* (harmony of the Gospels).

Traditional connection with Peter

Attempts have been made to identify Mark as the John Mark mentioned in Acts 12 or as the disciple who fled naked in the garden (Mark 14). A reference to "my son, Mark," in I Peter is part of the same tradition by which Mark was related to Peter; thus the Evangelist's apostolic guarantor was Peter.

The setting is a Gentile church. There is no special interest in problems with Jews and little precision in stating Jewish views, arguments, or terminology. Full validity is given the worship of the Gentiles. In further support of

a Gentile setting and Roman provenance is the argument that Mark uses a high percentage of so-called Latinisms—*i.e.,* Latin loanwords in Greek for military officers, money, and other such terms. Similar translations and transliterations, however, have been found in the Jerusalem Talmud, a compendium of Jewish law, lore, and commentary, which certainly was not of Roman provenance. The argument from Latinisms must be weighed against the fact that Latin could be used anywhere in the widespread Roman Empire. In addition, for the first three centuries the language of the church of Rome was Greek—so the Gentile addressees might just as well have been Syrian as Roman. The Latinisms—as well as the Aramaisms—are rather an indication of the vernacular style of Mark, which was "improved" by the other Evangelists.

Simplicity and directness of style

Mark is written in rather crude and plain Greek, with great realism. Jesus' healing of a blind man is done in two stages: first the blind man sees men, but they look like trees walking, and only after further healing activity on Jesus' part is he restored to see everything clearly. This concrete element was lost in the rest of the tradition. It is also perhaps possible that this two-stage healing is a good analogy for understanding Mark theologically: first, through enigmatic miracles and parables in secret, and only later, after recognition of Jesus as the Christ, is there a gradual clarification leading to the empty tomb. In chapter 3, verse 21, those closest to Jesus call him insane ("he is beside himself"), a statement without parallel in the other Gospels.

In Mark, some Aramaic is retained, transliterated into Greek, and then translated—*e.g.,* in the raising of Jairus' daughter (5:41) and in the healing of the deaf mute (7:34). The well-known *abba,* Father, is retained in Mark's account of Jesus' prayer in Gethsemane. In the two miracle stories, the Aramaic may have been retained to enhance the miracle by the technique of preserving Jesus' actual words. And a cry of Jesus on the Cross is given in Aramaized Hebrew.

The stories in Mark are woven together with simple stereotyped connectives, such as the use of *kai euthus* ("and immediately," "straightway"), which may be thought of as a Semitic style (as a typical simple connective in the Old Testament narrative style). More likely, however, this abruptness indicated that the compiler-redactor of Mark has used geography and people simply as props or scenes to be used as needed to connect the events in the service of the narrative.

Except for the Passion narrative, there is little chronological information. References in chapters 13 and 14 appear to presuppose that the Jerusalem Temple (destroyed in AD 70) still stood (in Matthew and Luke this is no longer the case); but the context of chapter 13, the "Little Apocalypse," is so interwoven with eschatological traditions of both the Jewish and Christian expectations in the 1st century that it cannot serve with certainty as a historical reference. To some extent, however, chapter 13 does help to date Mark—the priority of which has already been established from literary criticism—because it is in good agreement with the traditions that Mark was written after the martyrdom of Peter. Mark may thus be dated somewhere after 64 and before 70, when the Jewish war ended.

Structure of the Gospel

The organization and schematizing of Mark reveals its special thrust. It may be roughly divided into three parts: (1) 1:1–8:26—the Galilean ministry—an account of mighty deeds (an aretalogy); (2) 8:27–10:52—discussions with his disciples centred on suffering; and (3) 11:1–16:8—controversies, Passion, death, the empty tomb, and the expected Parousia in Galilee.

"The beginning of the Gospel" in the first words of Mark apparently refers to John the Baptist, who is clearly described as a forerunner of the Messiah who calls the people to repentance. Jesus never calls himself the Messiah (Christ). After Jesus' Baptism by John, the heavens open, the Spirit descends, and a heavenly voice proclaims Jesus as God's beloved son with whom He is well pleased. Already in this account there is a certain secrecy, because it is not clear whether the onlookers or only Jesus witnessed or heard. Jesus was then driven by the Spirit into the wilderness, the place of demons and struggle, to be tempted by Satan, surrounded by wild beasts (the symbols of the power of evil and persecution) and ministered to by angels. Here again he is in secret, alone. The opening of the struggle with Satan is depicted, and the attendance by angels is a sign of Jesus' success in the test.

Many references to persecution in Mark point toward Roman oppression and a martyr church that was preoccupied with a confrontation with the Satanic power behind the world's hostility to Jesus and his message. There was stress on the underlying fact that the church must witness before the authorities in a hostile world. Much of the martyrological aspect of Mark's account is grounded in his interpretation of the basic function of Jesus' Passion and death and its implication that the Christian life is a life of suffering witness.

The proximity of the Kingdom of God

What Jesus preached in Galilee at the beginning of his ministry was that the time is fulfilled and the Kingdom of God is "at hand"; *i.e.,* very very near—therefore repent! (1:15). In Matthew this same message is that of both John the Baptist (3:2) and Jesus (4:17). This sets the stage; and the miraculous ministry in Galilee about which the followers are enjoined to secrecy points not so much to Jesus as the wonder-worker as to the great scheme of pushing back the frontier of Satan. Toward the end of this first section, the Pharisees ask Jesus for a sign, and he answers in no uncertain terms that no sign will be given (8:12). In the Synoptic Gospels the miracles are never called "signs" (as in John); and no sign is to be given prior to the cosmological, eschatological signs from heaven that belong to the end: darkening of the Sun and Moon and extreme tribulations that in postbiblical Jewish eschatology—the mood of the first Christian century—is a sign of the coming of the heavenly Son of man to judge the world.

Parables are a revelatory mode of expression; they are not just illustrations of ideas or principles. Jesus, the revealer, tells his disciples that the secret of the Kingdom of God is given to them but that to the outsider everything is in parables (or riddles) *in order that* they may not hear and understand lest they repent and be forgiven (4:10–12). This mystery and hiddenness is particularly related to the parables about the coming of the kingdom. Yet, even Jesus' disciples did not recognize him as the Messiah, although his miracles were such that only a messianic figure could perform them: forgiving sins on earth, casting out demons, raising the dead, making the deaf hear and the stammerer (the dumb) speak, and the blind to see—all fulfillments of Old Testament prophecy concerning the Messiah. Only the demons, supranatural beings, recognize Jesus. There is a constant campaign against Satan from the temptation after Jesus' Baptism until his death on the Cross, and, in each act of healing or exorcism, there is anticipated the ultimate defeat of Satan and the manifestation of the power of the new age. In all this Mark stresses the need for secrecy and Peter's confession of Jesus as the Christ (8:29) is told in Mark as the opportunity to motivate an acceptance of the admonition "not to tell" by reference to the necessity of suffering.

Emphasis on suffering and the messianic secret

This strong emphasis on the necessity of suffering—in the life of Jesus and in the life of the disciples—before the hour of victory gives the best explanation to what scholars have called the secrecy motif in Mark—*i.e.,* the constant stress on not telling the world about Jesus' messianic power.

According to William Wrede, a German scholar, the messianic secret motif was a literary and apologetic device by which the Christological faith of the early church could be reconciled with the fact that Jesus never claimed to be the Messiah. According to Wrede, Mark's solution was: Jesus always knew it but kept it a secret for the inner group. After Peter's confession at Caesarea Philippi, Jesus began to speak of a *suffering* Son of man. The Son of man in Jewish apocalyptic was a glorious, transcendent, heavenly figure who would come victorious on clouds of glory to judge the world at the end of time. Suffering was not part of this picture. E. Sjöberg (1955) has interpreted the messianic secret not as a literary invention but as an understanding both that the Messiah would appear without recognition except by those who are chosen and to whom he reveals himself and that he must suffer. For

outsiders, then, he remains a mystery until the age to come. Even his disciples did not understand the necessity of suffering. Only in the light of Resurrection faith—the hope of the Parousia and final victory over Satan—could they understand that he had to suffer and die to fulfill his mission and how they, too, must suffer.

Martyrological aspects in Mark can be noted from the beginning. Already according to 2:20 Jesus' disciples are not to fast until "when the bridegroom is taken away from them and then they will fast . . . ." In Mark 8 to 10, there is great concentration on discussions with the disciples. The theme is suffering, and repeatedly they are reminded that there is no way of coming to glory except through suffering. Three Passion predictions meet either with rejection, fear, or confusion. In the Transfiguration (9:2–13; in which three disciples—Peter, James, and John—see Jesus become brighter and Elijah and Moses, two Old Testament prophets, appear) there is the same emphasis. The tension between future glory and prior suffering is the more striking when the Transfiguration is recognized as a Resurrection appearance, placed here in an anticipatory manner. The disciples are reminded of an association of Elijah with John the Baptist and his fate. This is also a hidden epiphany (manifestation)—the triumphal enthroned king closely juxtaposed with suffering and death.

After the third Passion prediction, in chapter 10, two of the disciples ask for places of honour when Jesus is glorified. He reminds them that suffering must precede glory for "The Son of man also came not to be served but to serve, and to give his life as a ransom for many." It is worth noting that this is the only reference to the death of Christ as a ransom or sacrifice but that Mark does not dwell on the Christological implications, but uses the saying for ethical purposes. Even so, the Marcan text gives one of the important building blocks for Christological growth and reflection on the suffering Son of man.

Just as Jesus' public ministry in Mark started with the calling of disciples, so the central part of the Gospel calls them to participate through suffering in his own confrontation with the power of Satan.

Emphasis on the Passion

In the last section of the Gospel, the scene is shifted to Jerusalem, where Jesus is going to die. His entry is described as triumphal and openly messianic and is accompanied by acted-out parables in a judgment of a barren fig tree, casting money changers out of the Temple, and in a parable of a vineyard in which the beloved son of the owner is killed. There is an increasing conflict and alienation of the authorities. Chapter 13, the "Little Apocalypse," made up of a complex arrangement of apocalyptic traditions, serves as instruction to the disciples and thence to the church that they must endure through tribulation and persecution until the end time. Thus, although the setting is Jerusalem, the orientation is toward Galilee, the place where the Parousia is expected. The Holy Spirit will come to those who must witness in the situation of trial before governors and authorities (13:11); in the final eschatological trials only by God's intervention can anyone endure unless the time be shortened for the elect. Because this chapter is shaped as a discourse that precedes the Passion narrative, it serves as a farewell address, a type of testament including apocalyptic sayings and warnings to the messianic community at the end of the "narrative" before the Passion—as do most testament forms (admonitions given before death to those beloved who will remain behind).

The Cross is both the high point of the Gospel and its lowest level of abject humiliation and suffering. A cry of dereliction and agony and the cosmic sign of the rending of the Temple veil bring from a Gentile centurion acknowledgment of Jesus as Son of God. The disciples reacted to the scandal of the Cross with discouragement, although already the scene is set for a meeting in Galilee. There are no visions of the risen Lord, however, in the best manuscripts (verses 9–20 are commonly held to be later additions), and Mark thus remains an open-ended Gospel. The Resurrection is neither described nor interpreted. Not exultation but rather involvement in the battle with Satan is the inheritance until the victorious coming in glory of the Lord—a continual process with the empty tomb pointing to hope of the final victory and glory, the Parousia in Galilee. The Gospel ends on the note of expectation. The mood from the last words of Jesus to the disciples remains: What I say to you, I say to all: Watch!

The Gospel According to Matthew. Matthew is the first in order of the four canonical Gospels and is often called the "ecclesiastical" Gospel, both because it was much used for selections for pericopes for the church year and because it deals to a great extent with the life and conduct of the church and its members. Matthew gave the frame, the basic shape and colour, to the early church's picture of Jesus. Matthew used almost all of Mark, upon which it is to a large extent structured, some material peculiar only to Matthew, and sayings from Q as they serve the needs of the church. This Gospel expands and enhances the stark description of Jesus from Mark. The fall of Jerusalem (AD 70) had occurred, and this dates Matthew later than Mark, *c.* 70–80.

Matthew as the "ecclesiastical" Gospel

Although there is a Matthew named among the various lists of Jesus' disciples, more telling is the fact that the name of Levi, the tax collector who in Mark became a follower of Jesus, in Matthew is changed to Matthew. It would appear from this that Matthew was claiming apostolic authority for his Gospel through this device but that the writer of Matthew is probably anonymous.

The Gospel grew out of a "school" led by a man with considerable knowledge of Jewish ways of teaching and interpretation. This is suggested by the many ways in which Matthew is related to Judaism. It is in some ways the most "Jewish" Gospel. Striking are 11 "formula quotations" ("This was to fulfill what was spoken by the prophet . . .") claiming the fulfillment of Old Testament messianic prophecies.

The structure of Matthew

The outstanding feature of Matthew is its division into five discourses, or sermons, following narrative sections with episodes and vignettes that precede and feed into them: (1) chapters 5–7—the Sermon on the Mount—a sharpened ethic for the Kingdom and a higher righteousness than that of the Pharisees; (2) chapter 10—a discourse on mission, witness, and martyrological potential for disciples with an eschatological context (including material from Mark 13); (3) chapter 13—parables about the coming of the Kingdom; (4) chapter 18—on church discipline, harshness toward leaders who lead their flock astray and more gentleness toward sinning members; and (5) chapters 23–25—concerned with the end time (the Parousia) and watchful waiting for it, and firmness in faith in God and his Holy Spirit. Each sermon is preceded by a didactic use of narratives, events, and miracles leading up to them, many from the Marcan outline. Each of the five sections of narrative and discourse ends with a similar formula: "now when Jesus had finished these sayings. . . ." The style suggests a catechism for Christian behaviour based on the example of Jesus: a handbook for teaching and administration of the church. This presupposes a teaching and acting community, a church, in which the Gospel functions. The Greek word *ekklēsia,* ("church") is used in the Gospels only in Matthew (16:18 and 18:17).

The discourses are preceded by etiological (sources or origins) material of chapters 1–2, in which the birth narrative relates Jesus' descent (by adoption according to the will of God) through Joseph into the Davidic royal line. Though a virgin birth is mentioned, it is not capitalized upon theologically in Matthew. The story includes a flight into Egypt (recalling a Mosaic tradition). Some "Semitisms" add to the Jewish flavour, such as calling the Kingdom of God the Kingdom of the Heaven(s). The name Jesus (Saviour) is theologically meaningful to Matthew (1:21). Chapter 2 reflects on the geographical framework of the Messiah's birth and tells how the messianic baby born in Bethlehem came to dwell in Nazareth.

After the five narrative and discourse units, Matthew continues from chapter 26 on with the Passion narrative, burial, a Resurrection account, and the appearance of the risen Lord in Galilee, where he gives the final "great commission," with which Matthew ends.

The setting of Matthew

Matthew is not only an original Greek document, but its addressees are Greek-speaking Gentile Christians. By the time of the Gospel According to Matthew, there had been a relatively smooth and mild transition into a Gentile

Christian milieu. The setting could be Syria, but hardly Antioch, where the Pauline mission had sharpened the theological issues far beyond what seems to be the case in Matthew. Matthew has no need to argue against the Law, or Torah, as divisive for the church (as had been the case earlier with Paul in Romans and Galatians, in which the Law was divisive among Gentile Christians and Jewish Christians), and, indeed, the Law is upheld in Matthew (5:17–19). For Matthew, there had already been a separation of Christianity from its Jewish matrix. When he speaks about the "scribes and the Pharisees," he thinks of the synagogue "across the street" from the now primarily Gentile church. Christianity is presented as superior to Judaism even in regard to the Law and its ethical demands.

The Matthean church is conscious of its Jewish origins but also of a great difference in that it is permeated with an eschatological perspective, seeing itself not only as participating in the suffering of Christ (as in Mark) but also as functioning even in the face of persecution while patiently—but eagerly—awaiting the Parousia. The questions of the mission of the church and the degree of the "coming" of the Kingdom with the person and coming of Jesus are handled by the Evangelist by a "timetable" device. The Gospel is arranged so that only after the Resurrection is the power of the Lord fully manifest as universal and continuing. Before the Resurrection the disciples are sent nowhere among the Gentiles but only to the lost sheep of the house of Israel; and the end time is expected before the mission will have gone through the towns of Israel. Even in his earthly ministry, however, Jesus proleptically, with a sort of holy impatience, heals the son of a believing Roman centurion and responds to the persistent faith of a Canaanite woman—whose heathen background is stressed even more than her geographical designation, Syro-Phoenician, given in the parallel in Mark—by healing her daughter. The Jewish origins of Jesus' teaching and the way the Evangelist presents them do not deny but push beyond them. The prophecies are fulfilled, the Law is kept, and the church's mission is finally universal, partly because the unbelief of the pious Jewish leaders left the gospel message to the poor, the sick, the sinner, the outcast, and the Gentile.

Jesus as teacher, Messiah, King, and Judge

In Matthew, because of the use of Q and Matthew's theological organization, there is stress on Jesus as teacher, his sharpening or radicalizing of the Law in an eschatological context; and Jesus is presented not in secret but as an openly proclaimed Messiah, King, and Judge. In the temptation narrative Jesus refuses Satan's temptations because they are of the devil, but he himself later in the Gospel does feed the multitude, and after the Resurrection he claims all authority in heaven and on earth. By overcoming Satan, Jesus gave example to his church to stand firm in persecution. Messianic titles are more used in Matthew than in Mark. In the exorcism of demoniacs, the demons cry out, calling him Son of God and rebuking him for having come "before the time" (8:29). Again, this shows that Jesus in his earthly ministry had power over demons, power belonging only to the Messiah and the age to come; and he pushed this timetable ahead. Yet, as in Mark, the miracles are not to be interpreted as signs. When asked for a sign, the Matthean account gives only the sign of Jonah, an Old Testament prophet—*i.e.,* the preaching of the gospel—which in later tradition took on an added interpretation as presaging the Son of man (Jesus) being three days and nights in the tomb (12:40, a later addition to Matthew).

Even the antitheses in the Sermon on the Mount are not new but demonstrate a higher ethic—one that is sharpened, strict, more immediate because the end time is perceived as coming soon. People who took this intensification of the Law upon themselves dared to do it as an example of "messianic license"—*i.e.,* to use the ethics of the Kingdom in the present in a church still under historical ambiguity and in constant struggle with Satan.

At such points the peculiar nature of Matthew comes into focus. The sharpening of the Law and the messianic license for the disciples are clearly there. At the same time Matthew presents the maxims of Jesus as attractive to a wider audience with Hellenistic tastes: Jesus is the teacher of a superior ethic, beyond casuistry and particularism. Similarly, in chapter 15, he renders maxims about food laws as an example of enlightened attitudes, not as rules for actual behaviour.

According to Matthew, the "professionally" pious were blind and unhearing, and these traits led to their replacement by those who are called in Matthew the "little ones"; in Final Judgment the King-Messiah will judge according to their response to him who is himself represented as one of "the least of these." The depiction of Jesus as Lord, King, Judge, Saviour, Messiah, Son of man, and Son of God (all messianic titles) is made in a highly pitched eschatological tone. The Lord's Prayer is presented in this context, and, for example, the "temptation" (trial, test) of "Lead us not into temptation" is no ordinary sin but the ordeal before the end time, the coming of the Kingdom for which the Matthean church prays. Martyrdom, though not to be pursued, can be endured through the help of the Spirit and the example of Jesus.

The Passion narrative is forceful and direct. Pilate's part in sentencing Jesus to be crucified is somewhat modified, and the guilt of the Jews increased in comparison with the Marcan account. In Matthew the Resurrection is properly witnessed by more than one male witness so that there can be no ambiguity as to the meaning of the empty tomb. The risen Lord directs his disciples to go to Galilee, and the Gospel According to Matthew ends with a glorious epiphany there and with Jesus' commission to the disciples—the church—to go to the Gentiles, because the risen Jesus is Lord of heaven and earth for all time.

The Gospel According to Luke. Luke is the third in order of the canonical gospels, which, together with Acts, its continuation, is dedicated by Luke to the same patron, "most excellent" Theophilus. Theophilus may have been a Roman called by a title of high degree because he is an official or out of respect; or he may have been an exemplification of the Gentile Christian addressees of the Lucan Gospel. The account in Luke–Acts is for the purpose of instruction and for establishing reliability by going back to the apostolic age. The very style of this preface follows the pattern of Greek historiography, and thus Luke is called the "historical" Gospel. Historically reliable information cannot be expected, however, because Luke's sources were not historical; they rather were embedded in tradition and proclamation. Luke is, however, a historian in structuring his sources, especially in structuring his chronology into periods to show how God's plan of salvation was unfolded in world history. That he uses events and names is secondary to his intention, and their historical accuracy is of less importance than the schematization by which he shows Jesus to be the Saviour of the world and the church in its mission (Acts) to be part of an orderly progress according to God's plan.

The purpose of Luke

The sources of the Gospel are arranged in the service of its theological thrust with definite periodization of the narrative. Approximately one-third of Luke is from Mark (about 60 percent of Mark); 20 percent of Luke is derived from Q (sometimes arranged with parts of L). Almost 50 percent is from Luke's special source (L), especially the infancy narratives of John the Baptist and Jesus, and parables peculiar to Luke (*e.g.,* the prodigal son, the good Samaritan, the rich fool). L material is also interwoven into the Passion narrative. While Matthew structured similar teaching materials in his five discourses, Luke places them in an extensive travel account that takes Jesus from Galilee to Judaea via Jericho to Jerusalem. This is similar to the ways in which Acts is structured on the principle of bringing the word from Jerusalem to Rome (see below).

The author has been identified with Luke, "the beloved physician," Paul's companion on his journeys, presumably a Gentile (Col. 4:14 and 11; cf. II Tim. 4:11, Philem. 24). There is no Papias fragment concerning Luke, and only late-2nd-century traditions claim (somewhat ambiguously) that Paul was the guarantor of Luke's Gospel traditions. The Muratorian Canon refers to Luke, the physician, Paul's companion; Irenaeus depicts Luke as a follower of Paul's gospel. Eusebius has Luke as an Antiochene physician who was with Paul in order to give the Gospel apostolic authority. References are often made to Luke's

medical language, but there is no evidence of such language beyond that to which any educated Greek might have been exposed. Of more import is the fact that in the writings of Luke specifically Pauline ideas are significantly missing; while Paul speaks of the death of Christ, Luke speaks rather of the suffering, and there are other differing and discrepant ideas on Law and eschatology. In short, the author of this gospel remains unknown.

The setting of Luke

Luke can be dated *c.* 80. There is no conjecture about its place of writing, except that it probably was outside of Palestine because the writer had no accurate idea of its geography. Luke uses a good literary style of the Hellenistic Age in terms of syntax. His language has a "biblical" ring already in its own time because of his use of the Septuagint style; he is a Greek familiar with the Septuagint, which was written for Greeks; he seldom uses loanwords and repeatedly improves Mark's wording. The hymns of chapters 1 and 2 (the Magnificat, beginning "My soul magnifies the Lord"; the Benedictus, beginning "Blessed be the Lord God of Israel"; the Nunc Dimittis, beginning "Now lettest thou thy servant depart in peace") and the birth narratives of John the Baptist and Jesus either came from some early oral tradition or were consciously modelled on the basis of the language of the Septuagint. These sections provide insight into the early Christian community, and the hymns in particular reflect the Old Testament psalms or the *Thanksgiving Psalms* from Qumrān. Though on the whole Matthew is the Gospel most used for the lectionaries, the Christmas story comes from Luke. The "old age" motif of the birth of John to Elizabeth also recalls the Old Testament birth of Samuel, the judge. All the material about John the Baptist, however, is deliberately placed prior to that of Jesus. When Mary, the mother of Jesus, visits Elizabeth, Jesus' superiority to John is already established. The Davidic royal tradition is thus depicted as superior to the priestly tradition.

Writing out of the cultural tradition of Hellenism and that of Jewish *'anawim* piety—*i.e.,* the piety of the poor and the humble entertaining messianic expectations—Luke has "humanized" the portrait of Jesus. Piety and prayer (his own and that of others) are stressed. Love and compassion for the poor and despised and hatred of the rich are emphasized, as is Jesus' attitude toward women, children, and sinners. In the Crucifixion scene, the discussion between the robbers and Jesus' assurance that one of them would be with him in Paradise, as well as the words, "Father, into thy hands I commit my spirit!"—which are in contrast to the cry of dereliction in Mark and Matthew—all point toward the paradigm of the truly pious man. Parables peculiar to Luke—among which are those of the good Samaritan, the importunate friend, the lost coin, and the prodigal son—have an element of warmth and tenderness. Thus, Luke "civilizes" the more stark eschatological emphasis of Mark (and Matthew), leading the way, perhaps, to a lessening of eschatological hopes in a time in which the imminent Parousia was not expected but pushed into the distant future.

The structure of the Gospel

The interplay between Luke and Acts reveals Luke's answer to the coming of the Kingdom. Once the church has the Holy Spirit, the delay of the Parousia has been answered for a time. Thus, Luke divides history into three periods: (1) the end of the prophetic era of Israel as a preparation for revelation, with John the Baptist as the end of the old dispensation; (2) the revelation of Jesus' ministry as the centre of time—with Satan having departed after the temptation and, until he once again appears, entering into Judas to betray Jesus; and (3) the beginning of the period of the church after Jesus' Passion and Resurrection.

Consistent with this schematization, John the Baptist's arrest occurs before Jesus' Baptism, though it is placed later in Mark and Matthew. From the beginning, the rule of the Spirit is a central theme, important in healing, the ministry, the message, and the promise of the continued guidance of the Spirit in the age of the church, pointing toward part two of Luke's work, the book of Acts of the Apostles, in which Pentecost (the receiving of the Holy Spirit by 120 disciples gathered together the 50th day after Easter) is a decisive event.

Just as Luke arranges his Gospel to show the divine plan of salvation in historical periodization, so he orders its structure in accordance with a geographical scheme. Chapter 1 (verse 8) of Acts provides the framework: after the coming of the Spirit, the church will witness in Jerusalem, in all Judaea and Samaria, and then to the end of the inhabited world. These places foreshadow the church's mission. The end of the old dispensation takes place in Jerusalem and its environs. The Resurrection appearances in Luke are placed in Jerusalem (Mark, Matthew, and John point toward Galilee). Jerusalem is also the place of the beginning of the church, and the old holy place thus becomes the centre of the new holy community. The necessity of suffering was made clear and interpreted as the fulfillment of prophecy. Rejection by people from his old home, Nazareth, and by Jewish religious leaders corresponds to the beginning of the ministry to the Gentiles—to the end of the earth.

Luke's interpretation of the Passion

Luke's account of the Crucifixion heightens the guilt of the Jews, adding a trial and mockery by Herod Antipas. The Crucifixion in Luke is interpreted as an anticipatory event: that the Christ must suffer by means of death before entering into glory. Jesus' death, therefore, is not interpreted in terms of an expiatory redemptive act. The centurion who saw the event praised God and called Jesus a righteous man, thus describing his fate as that of a martyr, but with no special meaning for salvation. The link between past salvation history and the period of the church is through the Spirit; salvation history continues in Acts.

THE FOURTH GOSPEL: THE GOSPEL ACCORDING TO JOHN

John is the last Gospel and, in many ways, different from the Synoptic Gospels. The question in the Synoptic Gospels concerns the extent to which the divine reality broke into history in Jesus' coming, and the answers are given in terms of the closeness of the new age. John, from the very beginning, presents Jesus in terms of glory: the Christ, the exalted Lord, mighty from the beginning and throughout his ministry, pointing to the Cross as his glorification and a revelation of the glory of the Father. The Resurrection, together with Jesus' promise to send the Paraclete (the Holy Spirit) as witness, spokesman, and helper for the church, is a continuation of the glorious revelation and manifestation (Greek *epiphaneia*).

Irenaeus calls John the beloved disciple who wrote the Gospel in Ephesus. Papias mentions John the son of Zebedee, the disciple, as well as another John, the presbyter, who might have been at Ephesus. From internal evidence the Gospel was written by a beloved disciple whose name is unknown. Because both external and internal evidence are doubtful, a working hypothesis is that John and the Johannine letters were written and edited somewhere in the East (perhaps Ephesus) as the product of a "school," or Johannine circle, at the end of the 1st century. The addressees were Gentile Christians, but there is accurate knowledge and much reference to Palestine, which might be a reflection of early Gospel tradition. The Jews are equated with the opponents of Jesus, and the separation of church and synagogue is complete, also pointing to a late-1st-century dating. The author of John knows part of the tradition behind the Synoptic Gospels, but it is unlikely that he knew them as literary sources. His use of common tradition is molded to his own style and theology, differing markedly with the Synoptics in many ways. Yet, John is a significant source of Jesus' life and ministry, and it does not stand as a "foreign body" among the Gospels. Confidence in some apostolic traditions behind John is an organic link with the apostolic witness, and, from beginning to end, the confidence is anchored in Jesus' words and the disciples' experience—although much has been changed in redaction. Traces of eyewitness accounts occur in John's unified Gospel narrative, but they are interpreted, as is also the case with the other Gospels. Clement of Alexandria, a late-2nd-century theologian, calls John the "spiritual gospel" that complements and supplements the Synoptics. Although the Greek of John is relatively simple, the power behind it (and its "poetic" translation especially in the King James Version) makes it a most beautiful writing. Various backgrounds for John have been suggested: Greek philosophy (especially the Stoic concept

Theories of various backgrounds of the Gospel According to John

of the *logos,* or "word," as immanent reason); the works of Philo of Alexandria, in which there is an impersonal *logos* concept that can not be the object of faith and love; Hermetic writings, comprising esoteric, magical works from Egypt (2nd–3rd centuries AD) that contain both Greek and Oriental speculations on monotheistic religion and the revelation of God; Gnosticism, a 2nd-century religious movement that emphasized salvation through knowledge and a metaphysical dualism; Mandaeanism, a form of Gnosticism based on Iranian, Babylonian, Egyptian, and Jewish sources; and Palestinian Judaism, from which both Hellenistic and Jewish ideas came. In the last source there is a Wisdom component and some ideas that possibly come from Qumrān, such as a dualism of good versus evil, truth versus falsehood, and light versus darkness. Of these backgrounds, perhaps, all have played a part, but the last appears to fit John best. In the thought world of Jewish Gnosticism, there is a mythological descending and ascending envoy of God. In the prologue of John, there is embedded what is proclaimed as a historical fact: The Logos (Word) took on new meaning in Christ. The Creator of the world entered anew with creative power. But history and interpretation are always so inextricably bound together that one cannot be separated from the other.

In John there is a mixture of long meditational discourses on definite themes and concrete events recalling the structure of Matthew (with events plus discourses); and, although the source problem is complex and research is still grappling with it, there can be little doubt that John depended on a distinct source for his seven miracles (the sign [or *sēmeia*] source): (1) turning water to wine at the marriage at Cana; (2) the healing of an official's son; (3) the healing of a paralytic at the pool at Bethzatha; (4) the feeding of the multitude; (5) Jesus walking on water; (6) the cure of one blind from birth; and (7) the raising of Lazarus from the dead. In chapter 20, verse 30, the purpose of the signs is stated: "Jesus did many other signs in the presence of the disciples, which are not written in this book; but these are written that you may believe that Jesus is the Christ, the Son of God, and that believing you may have life in his name."

The sign source for John

A major part of John is in the form of self-revelatory discourses by Jesus. Some would assign these to a distinct source, but they may rather be the work of the author.

Jesus' coming "hour"—the hour of his glorification—could not come about at any bidding but only according to a divine plan, and Jesus is obedient to it. The Paraclete is promised to come to the disciples, and it is necessary that Jesus go away in order that the Paraclete may come to the church. In John, Christ is depicted as belonging to a higher world, and his kingship is not of this world. He is said to have come into this world to his own people, and they rejected him, but this is but another example of the church's mission having passed both historically and theologically to the Gentile milieu.

The Christology in John is heightened: though the Synoptics have Jesus speaking about the Kingdom, in John, Jesus speaks about himself. This heightened Christology can be seen in many of the "I am" sayings of Jesus *(e.g.,* "I am the bread of life") in the context of their discourses and accompanying signs. This type of discourse is a concentration in terms and titles of the way in which the Messiah openly reveals his identity by a striking phenomenon: in the Old Testament the association with "I am" is the revelation of the name of God in the theophany (manifestation of God) to Moses (Exodus), and this theophanic interpretation carries over in John. Jesus says "I am" with regard to his function as Messiah, as divine. These sayings are self-revelatory pronouncements: (1) bread of life, (2) light of the world, (3) door of the sheepfold, (4) good shepherd, (5) resurrection and life, (6) way, truth, and life, and (7) true vine. Such theophanic expressions are heightened in other sayings: "I and the Father are one"; "Before Abraham was, I am"; "He who has seen me has seen the Father"; and Thomas' cry after the Resurrection "My Lord and my God."

The "I am" pronouncements

John 14 is a farewell speech, one of a series, before the Passion. In testament form, it is the bidding of farewell by one who is dying and giving comfort to those he loves. In John, however, the eons (ages) overlap. The significance of the farewell address, thus, is in the teaching that Jesus is God's representative. The fact that he must go to the Father means that the eschatological era already started in Jesus' presence as the Christ and will be intensified at his death and manifested further in the coming of the Spirit to the church. The times shift; the eschatology—here and still to come—also shifts but remains on the whole realized in John, although there is still a tension between the "already" and the "not yet."

John's allegorical thought is shown by his ending of the miracle of Jesus' walking on the sea. The frightened disciples took him into their boat, "and immediately the boat was at the land." This fits the pattern of John's Gospel, namely that, when Jesus is with his church, the new era has already arrived, and, where Jesus is, there is the Kingdom fulfilled. Similarly, the raising of Lazarus in chapter 11 is to demonstrate that the power of the Resurrection, of the fulfilled "eschaton" (last times), is already present in Jesus as Christ now, not only in some future time. Thus, there would appear to be a "realized eschatology" in John; *i.e.,* the last times are realized in the person and work of Jesus. The coming of the Spirit, the Paraclete, however, is still to come, so, even in this most eschatological Gospel, there is a building up, a crescendo, of glorification. In chapter 12, verse 32, Jesus is depicted as saying, "I, when I am lifted up . . . will draw all men to myself"—again an exaltation and glorification that points to the Cross. At the point of death on the Cross, Jesus' words "It is finished" are interpreted to mean that part of the "eschaton" is consummated, fulfilled. After the finding of the empty tomb, there is a Resurrection appearance to the disciples. This includes the "doubting Thomas" pericope, which teaches that those who have to depend on the witness of the Gospel are at no disadvantage.

In an appended chapter, 21, there is a touching story of the Apostle Peter, who, having denied his Lord thrice, is three times asked by Jesus if he loves him. Peter affirms his knowledge that Jesus knows what love is in his heart and is given the care of the church and a prediction that he himself will be persecuted and crucified.

The numerous differences between the Synoptics and John can be summed up thus: in John eternal life is already present for the believer, while in the Synoptics there is a waiting for the Parousia for the fulfillment of eschatological expectations. This Johannine theology and piety has great similarities to the views that Paul criticizes in I Cor. 15 (see below). The contrast between Paul and John is even more striking if one accepts the most plausible theory that John as we have it includes passages (added later) by which the realized eschatology has been corrected so as to fit better into the more futuristic eschatology that was stressed in defense against the Gnostics. John 5:25–28 is such a striking correction.

Differences between John and the Synoptics

The Johannine chronology also differs from the Synoptic. John starts the public ministry with the casting out of the money changers: the Synoptics have this as the last event of the earthly ministry leading to Jesus' apprehension. The public ministry in John occupies two or three years, but the Synoptics telescope it into one. In John Jesus is crucified on 14 Nisan, the same day that the Jewish Passover lamb is sacrificed; in the Synoptics Jesus is crucified on 15 Nisan. The difference in the chronologies of the Passion between John and the Synoptics may be because of the use of a solar calendar in John and a lunar calendar in the Synoptics. Nevertheless, the actual dating is of less importance than the fact that John places the Crucifixion at the time of the Passover sacrifice to emphasize Jesus as the Paschal lamb. There is no celebration of the Last Supper in John, but the feeding of the multitude in chapter 6 gives the opportunity for a eucharistic discourse. Because Jesus is regarded as the Christ from the very beginning of John, there is no baptism story—John the Baptist bears witness to Jesus as the Lamb of God—no temptation, and no demon exorcisms. Satan is vanquished in the presence of Christ. Each of the four Gospels presents a different facet of the picture, a different theology. Although in all the Gospels there is warning about persecution and the danger of discipleship, each has the retrospective comfort

of having knowledge of the risen Lord who will send the Spirit. In John, however, there is a triumphant, glorious confidence: "In the world you have tribulation; but be of good cheer, I have overcome the world."

THE ACTS OF THE APOSTLES

The relationship between Luke and Acts

As indicated by both its introduction and its theological plan (see *The Gospel According to Luke*), Acts is the second of a two-volume work compiled by the author of Luke. Both volumes are dedicated to Theophilus (presumably an imperial official), and its contents are divided into periods. In the Gospel, Luke describes first the end of the old dispensation and then the earthly life of Jesus. Near the end of the Gospel, the stage is set for the next period: the "new dispensation" of the church as presented in Acts. After the Ascension of the risen Lord in Jerusalem (Acts 1), there is Pentecost, called Shavuot in Hebrew (*i.e.*, "the 50th day" after Passover). This Jewish festival of the revelation of the Law on Mt. Sinai becomes the day when the Spirit is poured out. For Acts this event marks the beginning of a new era (Acts 2): as in Luke, Jesus, endowed by the Spirit, was led from Nazareth to Jerusalem, so in Acts, the outpouring of the Spirit at Pentecost leads the church from Jerusalem to Rome.

The purpose and style of Acts. Although the title, Acts of the Apostles, suggests that the aim of Acts is to give an account of the deeds of the Apostles, the title actually was a later addition to the work (about the end of the 2nd century). Acts depicts the shift from Jewish Christianity to Gentile Christianity as relatively smooth and portrays the Roman government as regarding the Christian doctrine as harmless. This book is the earliest "church history," viewing the church as guided by the Spirit until a future Parousia (coming of the Lord).

Probably written shortly after Luke (*c.* 85) as a companion volume, in no manuscripts or canonical lists is Acts attached to the Gospel.

Luke edited his history as a series of accounts, and thus Acts is not history in the sense of accurate chronology or of continuity of events but in the ancient sense of rhetoric with an apologetic aim. The author weaves strands of varying traditions and sources into patterns loosely clustered around a nucleus of past events viewed from the vantage point of later development.

The structuring of the material by time and geography may account for the unique way in which both the Ascension of Christ to heaven (40 days after the Resurrection) and the outpouring of the Spirit at Pentecost (50 days after the Resurrection) became fixed and dated events.

The redactor (editor) of Acts composed speeches with primary primitive material within them; about one-fifth of Acts is composed in this way. This manner of using speeches was part of the style and purpose of the work and was not unlike that of other ancient historians such as Josephus, Plutarch, and Tacitus.

The use of speeches and the "we-passages"

In the latter part of Acts are several sections known as the "we-passages" (*e.g.*, 16:10, 20:5, 21:1,8, 27:1, 28:16) that appear to be extracts from a travel diary, or narrative. These do not, however, necessarily point to Luke as a companion of Paul—as has been commonly assumed—but are rather a stylistic device, such as that noted particularly in itinerary accounts in other ancient historical works (*e.g.*, Philostratus' *Life of Apollonius of Tyana*). Though the pronoun changes from "they" to "we," the style, subject matter, and theology do not differ. That an actual companion of Paul writing about his mission journeys could be in so much disagreement with Paul (whose theology is evidenced in his letters) about fundamental issues such as the Law, his apostleship, and his relationship to the Jerusalem church is hardly conceivable.

Acts was written in relatively good literary Greek (especially where it addresses the Gentiles), but it is not consistent, and the Koinē (vernacular) Greek of the 1st century was apparently more natural to the writer. There are some Semitisms, especially when stressing Jewish backgrounds; thus, Paul is called Saul in accounts of his conversion experience on Damascus road. In chapter 17, Paul's speech on the Areopagus, a hill in Athens that traditionally was the meeting place of the city's council, for an intellectual Athenian audience is in good Greek, assimilating Gentile thought patterns, but is expressed in Old Testament universalistic terms.

The content of Acts. The outline of Acts can be roughly divided into two parts: the mission under Peter, centred in Jerusalem (chapters 1–12); and the missions to the Gentiles all the way to Rome (*cf.* chapter 1, verse 8), under the leadership of Paul (chapters 13–28). The earlier sections deal with the Jerusalem church under Peter and the gradual spread of the gospel beyond Jewish limits (in chapters 10–11, for example, Peter is led by the Spirit to baptize the Roman centurion, Cornelius). References to Peter are abruptly ended in chapter 12; James, the brother of the Lord, has become the head of the Jerusalem church, and Philip, a Greek-speaking missionary, is commanded by the Spirit to baptize an Ethiopian eunuch.

Paul's missionary journeys and primitive traditions about Jesus

Paul's missionary journeys are traditionally separated into three: (1) 13:1–14:28; followed by the Council of Jerusalem *c.* AD 49 (15:1–35); (2) 15:36–18:22 with a stop at Antioch; and (3) 18:23–21:14. After that, Paul is imprisoned and sent to Rome where Acts leaves him witnessing openly and unhindered in the capital of the Empire. These journeys may be seen as a part of the writer's "theological geography," because they form one continuous circuit—with stops on the way—between the geographical poles of Jerusalem and Rome. After the Council of Jerusalem *c.* AD 49, the situation was changed, and Paul became the spokesman for the whole Christian mission.

The earliest chapters of Acts contain some primitive traditions important both for any study of the early church and its preaching and for the church's own development of its understanding of itself and of Jesus. After Peter healed a lame man, he made a speech, in chapter 3, in which Jesus is proclaimed as the one appointed but who is now in heaven and who will come as the Christ at the Parousia (Second Coming). In his Pentecost speech in chapter 2, Peter preached that God made Jesus Lord and Christ at his Resurrection.

The titles used for Jesus show both a preservation of primitive tradition and theology and a clear differentiation made by the writer between Jesus in his earthly life (in Luke) and reflection on him in Acts. Christ (Messiah) is consciously used as the title of Jesus; the title Son of man, used frequently in Luke, is used only once in Acts, at the death of the martyr Stephen, when he is granted a vision of the Lord in glory. Early titles, "servant" and "righteous one," reflect the Old Testament background of God's "suffering servant." The Hellenistic term saviour (*sōtēr*) is used in Acts in chapters 5 and 13. The more primitive Christologies and titles show not only a flexibility of traditions but also the functional nature of New Testament Christology.

Differences between views of Paul in Acts and in Paul's Letters

Acts presents a picture of Paul that differs from his own description of himself in many of his letters, both factually and theologically. In Acts, Paul, on his way to Damascus to persecute the church, is dramatically stopped by a visionary experience of Jesus and is later instructed. In his letters, however, Paul stated that he was called by direct revelation of the risen Lord and given a vocation for which he had been born (recalling the call of an Old Testament prophet, such as Jeremiah) and was instructed by no man.

The account of Paul's relation to Judaism in Acts also differs from that in his letters. In Acts, Paul is presented as having received from the Jerusalem apostolic council the authority for his mission to the Gentiles as well as their decision—the so-called apostolic decree (15:20; *cf.* 15:29)—as to the minimal basis upon which a Gentile could be accepted into fellowship with Jewish Christians. According to this decree, Gentile converts to Christianity were to abstain from pollutions of idols (pagan cults), unchastity, from what is strangled, and from blood (referring to the Jewish cultic food laws as showing continuity with the old Israel). Circumcision, however, was not required, an important concession on the part of the Jewish Christians.

In Acts Paul is not called an Apostle except in passing, and the impression is given, contrary to Paul's letters, that he is subordinate to and dependent upon the twelve Apostles. When Paul entered a new city, he went first to the synagogue. If his message of the gospel was rejected,

he turned to the Gentiles. According to Paul's missionary practice and theology, the message had first to be spoken to the Jews as a reminder that Christianity is grounded in redemptive history; this prevents the connection with the old Israel from being forgotten. Because most Jews rejected Paul's message, the author proclaimed that salvation thus passed to the Gentiles.

Paul's relationship to the Romans and the dominance of the Holy Spirit

Roman authorities are depicted as treating Paul (and other Christians) in a just manner. The author repeatedly stressed that the Roman authorities did not find fault with the Christians but rather viewed Christian–Jewish antagonisms merely as one problem among Jewish factions. While in Corinth, during a conflict with the Jews, the Roman proconsul of Achaea in Greece, Gallio, refused to hear the charges brought against Paul because, according to Roman law, they were extralegal. On a later occasion in Ephesus, during a conflict with the silversmiths who derived their income from selling statuettes of the goddess Diana, Paul was protected from local antagonisms and a riot by Roman authorities. Toward the end of his career, after having been in the protective custody of the Judaean procurator Felix, Paul was heard by Felix's successor, Festus, and the Jewish king Agrippa II, and, had he not appealed to Caesar as a Roman citizen, he could have been set free. He thus had to go to Rome to be tried, and that is the last that is heard about him in Acts.

The doctrine of the Holy Spirit is a dominant theme in Acts, as it is in the Gospel According to Luke. Just as Jesus started his public ministry in Luke by reading from the Book of Isaiah: "The Spirit of the Lord is upon me . . ." so also in Acts the new age of the Spirit began at Pentecost, which is viewed as the fulfillment of the prophecy of Joel that in the new age the Spirit would be poured out on all men. That persons from many nations heard in their own tongues the mighty works of God has been viewed as a reversal of the Tower of Babel narrative, with languages no more confused and people no longer scattered.

Although Peter, Stephen, and Paul are central figures in Acts, the piety of the humbler members of the church also permeates the book. Church structure and organization, with apostles, disciples, elders, prophets, and teachers, exhibits great fluidity. Paul, in bidding farewell at Miletus to the elders from Ephesus, exhorted them to "take heed . . . to all the flock in which the Holy Spirit made you guardians (bishops) to feed the church. . . ." Offices may be conveyed by prayer and laying on of hands but there is little stress on distinction of office or succession, thus indicating a very early period in the life of the church.

Because Peter "departs and goes to another place" and Paul is left under house arrest awaiting trial, the readers appear to be left in suspense concerning the fates of these two leaders. The readers, however, probably knew what had happened to them—*i.e.,* that these Apostles had eventually been martyred sometime in the 60s before Acts was written. What is more, the interest in Acts is not in the fates of Peter and Paul; the gospel has finally reached Rome, the center of the *oikoumenē* ("the inhabited world"), and thus the ending is suitable to the book—Paul is left "preaching the kingdom of God and teaching about the Lord Jesus Christ quite openly and unhindered."

THE PAULINE LETTERS

In the New Testament canon of 27 books, 21 are called "letters," and even the Revelation to John starts and ends in letter form. Of the 21, 13 belong to the Pauline corpus; the Letter to the Hebrews is included in the Pauline corpus in the East but not, however, in the West. Three letters of this corpus, the Pastoral Letters, are pseudonymous and thus are not considered here. Of the remaining 10, the Letters to the Colossians and Ephesians are from the hand of a later Pauline follower and II Thessalonians is spurious. How this Pauline corpus was collected and published remains obscure, but letters as part of Holy Scripture were an early established phenomenon of Christianity.

The church was poor and widespread, and, in the early stages, expected an imminent Parousia. More formal sacred writings were thus superseded in importance by letters (*e.g.,* those of bishop Ignatius of Antioch) that answered practical questions of the early churches.

The letters of Paul, written only about 20–30 years after the crucifixion, were preserved, collected, and eventually "published." In general, they answered questions of churches that he had founded. When all the Pauline Letters as a corpus were first known is difficult to determine. Because Pauline theology and some quotations and allusions were certainly known at the end of the 1st century, the Pauline Letters probably were collected and circulated for general church use by the end of the 1st century or soon thereafter. A disciple of Paul, possibly Onesimus, may have used Ephesians as a covering letter for the whole collection.

The letters Galatians and Romans both contain an extensive discussion about the Law (Torah) and justification (in language not found in the other letters) to solve the problem of the relation of Christianity to Judaism and of the relationship of Jewish Christians with Gentile Christians. Galatians is older and differs from Romans in that it deals with Judaizers—*i.e.,* Gentile Christians who were infatuated with Jewish ways and championed Jewish ceremonial law for Gentile Christians. On the other hand, Romans speaks to the question of the Jews and the Christian faith and church in God's plan of salvation.

The significance of the Pauline Letters to the addressees

In I and II Corinthians (which may include fragments of much Corinthian correspondence preserved in a somewhat haphazard order), there is no preoccupation with either Jews or Judaizing practices. They deal with a church of Gentile Christians and are therefore the best evidence of how Paul operated on Gentile territory.

The earliest book in the New Testament is I Thessalonians, which is concerned with the problem of eschatology. Though II Thessalonians is obvious in its imitation of the style of I Thessalonians, it reflects a later time, elaborates on I Thessalonians, and is thus not viewed as genuine.

Philippians may be a composite letter in which various themes of Pauline teaching are held together by a testament form. Thus, it is a compendium without too specific a focus on the Philippian situation. Philemon, although addressed to a house church, is uniquely concerned with the fate of a slave being returned to his master, with the hope that he will be forgiven and be sent back to help Paul in prison, an example of manumission in Paul's name.

Ephesians appears to be dependent on Colossians, and both, although using the Pauline style, reflect a time and imagery sometimes different from and later than Paul's genuine letters. Ephesians covers the content of Colossians in more compact form and may be a covering letter for the entire Pauline corpus by a disciple or other later Paulinist.

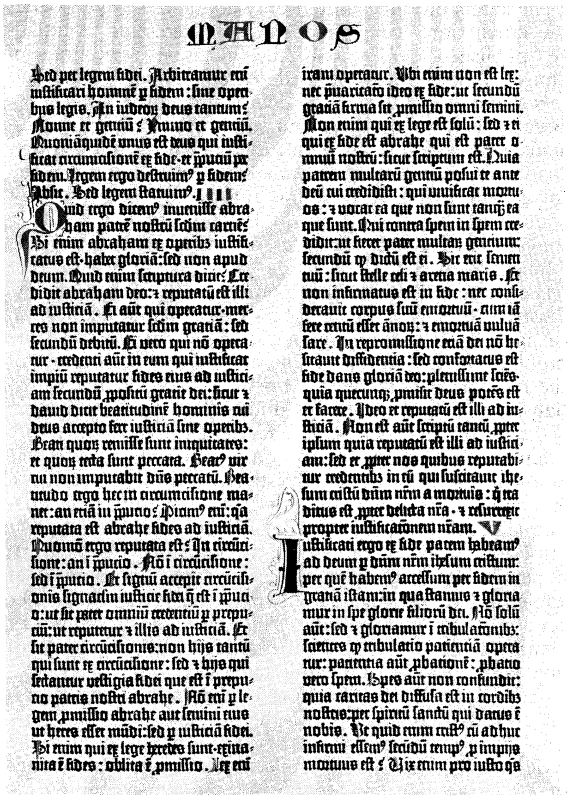
The style and form of Paul's letters

The style of Paul's letters is an admixture of Greek and Jewish form, combining Paul's personal concern with his official status as Apostle. After his own name, Paul names the addressees or congregation being addressed and adds "grace and peace." This is often followed by thanksgivings and intercession that are significantly adapted to the content and purpose of the letter. Doctrinal material usually precedes advice or exhortation (*parenesis*), and the letters conclude with personal news or admonition and a blessing: "The grace of our Lord Jesus Christ be with you." Paul's letters were probably dictated to an amanuensis (who might be named, for example, Sosthenes, I Cor. 1:2), and some greetings were written at the end of the letters in his own hand. They were obviously meant to be read aloud in the church, however, and thus their style is different from that of purely personal letters.

The reason for Paul's Letter to the Romans

The Letter of Paul to the Romans. Romans differs from all the other Pauline letters in that it was written to a congregation over which Paul did not claim apostolic authority. He stressed that he was merely going to Rome in transit, because it was his principle not to evangelize where others had worked. Because his apostolic ministry appeared to be completed in Asia Minor and Greece, Paul planned to go to Spain via Rome, a city that he had never visited. Before going westward, however, he first had to go to Jerusalem to deliver to the church there a collection of money.

Because Paul was going to a church he had not founded, his writing to the Roman Christians offered him an opportunity to present his theological views in a systematic way, which he had not done in other letters. Paul reflected



The end of chapter 3, chapter 4, and the opening of chapter 5 of the Letter of Paul to the Romans from a facsimile of Gutenberg's 42-line Bible. In the British Museum.

By courtesy of the trustees of the British Museum; photograph, J.R. Freeman & Co. Ltd.

on how his special mission fitted into God's plan for the salvation of mankind, of both Jews and Gentiles—a theme that reached its climax in chapters 9–11. Chapters 1–8 unfold with great specificity how the coming of Jesus the Messiah has made it possible for the Gentiles to become heirs to God's promises. His argument is at first negative, stating that neither Gentile nor Jew could effect his own salvation. He then shows a new way in which eventually both can be delivered from the bondage of sin by being justified—*i.e.,* made "right with God"—not through acceptance of the Law but by faith in the crucified Lord.

The theological section (chapters 1–11) is followed (as is often the case in Pauline letters) by ethical instructions. There is little doubt about the integrity of Romans 1–15; the letter was written from Corinth *c.* 56. Chapter 16, however, seems to be a later addition. It contains numerous salutations to individuals (which is unusual in that Paul had never been to Rome) and an antinomian (antilegalistic) tone that would be more appropriate to the situation in Asia Minor. The doxology (16:25–27) is rhetorical and its vocabulary is not in keeping with that of Paul's usual thought. Because the doxology occurs in different manuscripts in varying positions in the course of textual transmission, it is probably secondary. Chapter 16 may thus preserve portions of a letter or letters from some other time or to some place other than Rome, possibly Ephesus.

The main theme of the Letter to the Romans

In chapter 1, verses 1–17, there are greetings and thanksgivings leading to the main theme of the letter: the gospel is

> the power of God for salvation to every one who has faith (*i.e.,* that Jesus is the Messiah), to the Jew first and also to the Greek. For in it the righteousness of God is revealed through faith for faith; as it is written, "The righteous shall live by faith."

Paul took this sentence from the Old Testament Book of Habakkuk, chapter 2, verse 4, not as a principle but as a prophecy now fulfilled. Thus, the translation should read "will live" rather than "shall live." This does not refer to God's faithfulness but rather to the believer's trust. Justification by faith is not, however, the answer to the question of man, plagued by conscience, about his salvation nor is it deep theology. It is rather an argument totally grounded in the problem of the relationship of Jews and Gentiles—*i.e.,* how it will be possible for the Gentiles to be fellow heirs with Jews and how both Jews and Gentiles can be members of the church. In chapters 2–3 both Gentiles and Jews are demonstrated to have fallen short of the glory of God and to be under condemnation. A turning point, however, is emphasized in chapter 3: "But now the righteousness of God has been manifested apart from law. . . ." Justification is a gift through Jesus Christ and his expiating death for the salvation and vindication of all who believe in him. Because all this is through Christ and not by works of the Law, salvation is equally available to the Gentiles as well as to the Jews. For both, the means is the same: faith in Jesus the Christ.

The central problem after chapter 8, which describes the glory of the new dispensation in Christ and the Spirit (presented in chapters 9–11), centres on the mystery revealed to Paul, namely, that the Gentiles should be incorporated and be fellow heirs with the Jews. This is what Paul yearned for with respect to his fellow Jews. What makes it equally possible for Jew or Gentile to come to Christ is justification by faith, with the Law viewed as obsolete because Christ is the end of the Law (chapter 10, verse 4). Thus, there are, in effect, no distinctions between Gentile and Jew. Paul viewed his ministry as having made possible the inclusion of the Gentiles; as an apostle to the Gentiles he never urged them to carry on a mission to the Jews. He envisaged the Jewish acceptance of Christ as a mystery beyond human planning and effort, a divine event that will be the climax of history.

The ethical section (12:1–15:13) has no special reference to a situation in Rome. A close analysis shows that Paul here repeats thoughts and admonitions that are more specific in other letters. A metaphor of the church as a body (12:5), for example, is stylized and compressed as compared with the fuller use of the same in I Corinthians, chapter 12, and the pattern of weakness and strength in matters of food is best understood in the light of the fuller exposition in I Corinthians, chapters 8 and 10.

The First Letter of Paul to the Corinthians. This letter is part of Paul's correspondence with the Corinthian congregation founded by him and composed of Gentile Christians. The problems of Galatians and Romans, written to Christians with Jewish and Roman legal concepts, are different from those of I Corinthians, and, thus, the justification language is absent.

Except for the brief communication with Philemon (see below), I Corinthians is the most specifically practical, situation-oriented of Paul's letters. No other Pauline letter is so directly devoted to the consideration of practical and theological problems, many of them apparently communicated by the congregation through correspondence or by delegations. The letter, therefore, does not tend to stand as a unit and it is not uniform in its treatment of the varying situations.

Emphasis on practical and theological problems

Literary criticism—or redaction—has traditionally split the letter into several fragments with a presumed historical development within a relatively short period in the Corinthian church. Paul's reference to a previous letter of his in chapter 5, verse 9, has been the object of scholarly efforts to restore the earlier letter. The fragmentary and not-too-uniform nature of both I and II Corinthians, however, precludes much probability of success in such searches.

Writing from Ephesus *c.* 53 or 54 upon hearing from a certain Chloe's people that the church was rent by party factions, Paul tried to bring unity to the congregation. Whether these factions actually represented outside interference (*e.g.,* Cephas [Peter], Apollos, or others) or were factions of the congregation under the influence of a widespread heresy of the time is a question perhaps best answered by the fact that the factions do not come up again after I Corinthians, chapter 1, and that I Corinthians, chapter 3, reduces the factions to Apollos and Paul, who claims he is head of no party. The Christ "party"—*i.e.,* those who claim no party at all—(1:12; *cf.* 3:23) may be the only "party" Paul advocated because Christ is not divided. Paul warned that Christians should not fashion

themselves into parties under various leaders, because all these leaders are servants of Christ and stewards of the mysteries of God through whom Christians come to belief. The church is not a society with competitive philosophical schools.

The cause of the difficulties in Corinth: proto-Gnosticism

The letter is a response to difficulties caused or increased by a relatively strong group in Corinth that may be described as "enthusiasts." This group of enthusiasts may have been proto-Gnostics (early religious dualists not yet organized into definite sects). The Corinthian enthusiasts did, however, have some characteristics that would later be found in 2nd–3rd-century Gnosticism: a belief in salvation through spiritual knowledge or wisdom communicated by a revealer (not a redeemer); an otherworldliness that could lead either to licentiousness (scorn) or asceticism (withdrawal); and a basically dualist and deliberately syncretistic system of beliefs using the mythical speculations and magical ideas of their time.

The Corinthian problems might well be traced to such enthusiasts. Their *gnōsis* ("esoteric knowledge") was a religious knowledge that gave them the feeling of superiority over more pedestrian Christians. This *gnōsis* Paul identified as false wisdom. In chapter 14 Paul describes the views and related practices of those maintaining that they have spiritual gifts of inspiration, especially speaking in tongues (glossolalia) and *gnōsis.* Such enthusiasts prized eloquent or secret wisdom; they sought a revealer who had come into the world hidden from the evil powers and known only to those, the *pneumatikoi,* or the spiritual elite, who recognize him; and they tolerated gross immorality by claiming anything to be lawful for them (especially their slogan quoted by Paul: "for me all things are lawful"). These enthusiasts also rejected marriage because it furthered the propagation of the present evil world; they claimed to possess knowledge that made them indifferent to the world; and they believed that their salvation was guaranteed by ritual and rites. Though they prized spiritual gifts, they scorned the ordinary Christian services for the community; and they did not believe in a future resurrection of the dead, which in their system had no place or was nonsense.

The concept of love and views on the resurrection

The main Pauline answer (*e.g.,* as emphasized in chapter 13) was that love, namely concern for the building up of the community, surpasses all knowledge or spiritual gifts and that love is a corrective because it demands service, edification (*i.e.,* building up) of the church, and involves Christians with one another. Those Corinthians whom Paul viewed as opponents emphasized *gnōsis* over against love. The discussion of the resurrection in chapter 15 sheds further light on this. The opponents did not deny the Resurrection of Jesus Christ about which there was common agreement, but rather they debated about the future resurrection of Christians from the dead. Their view was perhaps similar to that reported as heresy in II Timothy, chapter 2, verse 18—*i.e.,* the believer already had eternal life and that a future resurrection of the body was meaningless. In holding such a view, Paul's opponents claimed they were faithful to the received kerygma (proclamation).

Another indication that some Corinthians had no disagreement with tradition but interpreted it too enthusiastically is found in I Corinthians, chapter 11. The liturgical formula pertaining to the Lord's Supper is sound:

> The Lord Jesus on the night when he was betrayed took bread, and when he had given thanks, he broke it, and said "This is my body which is for you. Do this in remembrance of me." In the same way also the cup, after supper, saying, "This cup is the new covenant in my blood. Do this, as often as you drink it, in remembrance of me." (11:23–25.)

In a discussion of the sacraments in chapter 10, however, the enthusiasts probably believed in a rather magical efficacy of Baptism and the Eucharist, though Paul qualified such an interpretation and took exception to it. The misunderstanding of the enthusiasts points to a special reinterpretation of Scripture and tradition (which resembles that of the 1st-century Jewish philosopher Philo and also the later Gnostics)—taking Scripture, tradition, and liturgical practices as effectively bringing about an otherworldly, spiritual reality immediately for those who really understand (*i.e.,* those who have *gnōsis*). Paul also criticized these spiritualists for their disregard of the poor members of the congregation, who found no food left when they came from their work.

The question of freedom and unity

Discussions about Christian and apostolic freedom (in chapters 5, 6, 7, 9, and 11) and also a discussion about being free to eat meat that had been sacrificed to idols and leftovers of pagan sacrifices sold in the marketplace were caused by conflicts with the enthusiasts who paraded their spiritual freedom, strength, and superiority at the expense of their weaker brothers in the faith, who were not ready for this freedom. A shift in the discussion in chapter 12 (the body and its members are equal in Christ)—from a very speculative idea of the body of Christ to a more metaphorical one that is reminiscent of Stoic philosophical ideas about society as an organism—can best be understood if it is assumed that the enthusiasts actually pressed for a mythical understanding of Christianity, in which one became literally incorporated into Christ, otherworldly, and divine. Paul added some qualifications that brought the church into concrete everyday life and even provided a source of political reality. A somewhat drastic understanding of spiritual gifts that was presupposed and criticized by Paul in chapters 12–14 fits well into such a pattern.

Permeating all the discussion of individual topics in I Corinthians is the theme of Christian unity and edification, a topic introduced and underscored in the preface and thanksgiving of this letter and in its introduction. Such unity is defended as being very inclusive, real, and concrete—as over against the enthusiastic attempt to speak in terms of spiritual reality and achievement, in which the true life of the spirit is only for the few (*i.e.,* the Gnostic elitists).

Paul viewed the necessity of unity in the wisdom of God as it is evinced in the scandal of the cross. In order to deflate the exalted and to make foolish the destructive (speculative) wisdom established by men, God showed his wisdom in the "foolishness" of Jesus' crucifixion. Here, although hidden, is God's true wisdom. The opponents hailed their ideal teachers as bringers of hidden wisdom. To this Paul said that it is Christ who is the Wisdom.

Ethical questions and views on marriage

In chapters 5 and 6 Paul dealt with certain ethical scandals and difficulties in the congregation: incest and fornication; the use of pagan courts for settling disputes among Christians; traffic with prostitutes—all for the demonstration of Christian "freedom." These wrongs might have been the direct or indirect consequences of the spiritual "powers" of the enthusiasts. According to Paul, however, such immorality was impossible for the Christian because of the concreteness of his allegiance to Christ and of inspiration (with the idea of the body as the temple of the Holy Spirit).

Because Paul expected an imminent Parousia (Second Coming of Christ), he suggested (chapter 7) the unmarried state as the preferable one, but conceded that marriage can prevent fornication. Paul even advised against breaking up mixed marriages between baptized Christians (both Jews and Gentiles) and unbaptized Gentiles. He advocated the practice of ascetics living together as "virgins," male and female, although he took this as a strain that is hard to bear and thus suggested marriage in unbearable cases. Not only the imminence of the Parousia but also radical change ("the form of this world is passing away") caused Paul, on the whole, to affirm the social status quo—whether it concern circumcision, slavery, or other matters. Everybody is advised to remain—for the short time ahead—in the state in which he finds himself. Such eschatological fervour caused Paul to argue against any worldly anxiety, fear, or worries stemming from them. This is reflected in the ethical criterion of possessing things as though one did not have them.

In chapter 9, Paul used his own conduct, in contrast to that of the enthusiasts who flaunted their freedom in such a way that it often had destructive influences, as a paradigm for an understanding of responsible freedom. Here he showed by various examples from his own lifestyle that he had never made use of his rightful privileges to the fullest, that he has, rather, been guided by what serves the weaker brothers and sisters. It is in this sense

that he subdued his body and that he urged the spiritual "snobs" to imitate him.

Discussions on corporate worship

In chapters 11–14, Paul turned to problems of corporate worship. Paul did not question the right and ability of prophetically gifted women to make inspired statements in Christian worship, but he pointed out that women need protection. Arguments about a veil or long hair for a woman are in the context of the church's worship before God himself, in which the congregation worships in the presence of the angels. Paul stressed the subordination of women in chapters 11 and 14; they are forbidden to speak in worship. In chapter 14 Paul stated (perhaps) a general principle that would allow for exceptions in cases of clear prophetic inspiration of women (*cf.* however, Galatians, chapter 3, verse 28).

In discussion of proper restraint and mutual regard in celebrating the Lord's Supper, Paul seemed to presuppose a prior common meal (possibly an agape meal) as part of the eucharistic celebration. This common meal, however, had apparently been devalued because of the interest of the enthusiasts in the sacrament itself. As a result, the communal aspect showed up social differences in the community; and some brought ample food, whereas others, of lower station, had nothing. In view of this, Paul again used the criterion of love and suggested that people eat their meal at home and then come together, being sensitive to each other's needs. The Lord's Supper would then be what it is, a proclamation of the death of Christ in anticipation of his return; mutual and corporate concern and responsibility thus become a part of the Eucharist.

Similarly, mutual edification and love are linked in chapter 13 as the appropriate centre of the discussion of spiritual gifts, manifested particularly in public worship (chapter 14).

The emphasis on the communal aspect of the church is continued in chapter 15. Paul did not dwell on his own vision of Christ nor on his role in founding the church at Corinth but rather argued for the resurrection of all as a future experience, not as though each person had already had this experience. Paul viewed the resurrection as a collective phenomenon in the expectation of an end-time resurrection from the dead, with Christ as the first fruits of those who have died.

That love is to extend beyond the immediate community and be shared with all the saints (members of the church) is demonstrated in chapter 16, the closing chapter, by the collection for the Jerusalem church. The keynote might be: "Let all that you do be done in love." The final passage—including the cry: "Our Lord, come!"—may reflect or repeat a eucharistic formula or setting.

The Second Letter of Paul to the Corinthians. This letter, as is I Corinthians, is composed of a collection of fragments of Paul's correspondence with the Corinthians about a year later (*i.e., c.* 55) from Macedonia. The diversity of I Corinthians was caused by the variety of problems discussed, but the diversity of II Corinthians was the result of a reflection of the underlying, rather turbulent history of Paul and his congregation. A pattern of fragments that make up II Corinthians can be understood in terms of a development that can be reconstructed. Gaps and editorial seams in this pattern are more recognizable and abrupt than those in I Corinthians, and a more original order for II Corinthians can be restored by fitting together blocks of material that obviously belong with one another in terms of context and unity of thought.

Though historical settings can be reconstructed with a high degree of validity to account for the fragments of II Corinthians, later editorial processes account for the order in which the fragments appear in the letter as it is now written. Based on both internal and external evidence, II Corinthians probably was later than I Corinthians, which was written after Paul's first trip to Corinth. Not long before the composition of II Corinthians, Paul was in mortal danger in Asia and travelled to Macedonia, where he remained.

New apostles and heresies had apparently invaded the Corinthian congregation and Paul sent his companion Timothy to try to bring them back to the true gospel as Paul had preached it. This mission was apparently unsuccessful, and Paul, in chapters 2 to 7, wrote to the church with a defense of his apostolic office, still counting on the loyalty of the Corinthians. His letter apparently did not change things, and there is some dispute as to whether Paul himself made an intermediate second visit to Corinth that was abruptly cut short by conflict with a member of the Corinthian church who violently opposed him. He considered such a second visit, but, according to chapter 2, verse 4, and chapters 10 to 13, he sent Titus to Corinth with a strongly polemical "letter of tears" and anxiously awaited his return, going from Troas to Macedonia to meet him.

New problems within the Corinthian congregation

Paul had almost been in despair over the Corinthians, but Titus and the letter seemed to have restored the Corinthian church to order. Titus and some of his companions were then sent to take up the collection for the church at Jerusalem, a sign of Christian mutual love and unity. He took with him Paul's "letter of reconciliation," which was written from Macedonia and which can be noted in chapter 1, verse 1, to chapter 2, verse 3; chapter 7, verses 5 and 6; and chapter 8. In chapter 8 the Macedonians are held up as an example of generosity. A similar section regarding the collection is in chapter 9, and the Achaeans (and probably their capital city, Corinth) were cited as an example to the Macedonians for generous giving. This was probably sent shortly before Paul's third (and last) visit to Corinth. From Corinth Paul wrote to the Roman church a letter that shows no sign of difficulties with the Corinthians and that presumed the conveying of the collection to Jerusalem.

The problem of the "letter of tears" and the "letter of reconciliation"

If the Corinthian controversy had been smoothed out, a question is raised as to why II Corinthians ends in the "letter of tears" rather than in the "letter of reconciliation." This may be understood if the literary order of the several sections was arranged by a redactor who collected the fragments probably in the last decade of the 1st century. The redactor may have used a "form" amply illustrated in Christian writings of the late 1st and early 2nd century; one of the end-time expectations was that "false prophets would show signs and wonders to lead the elect astray," and chapters 10–13 deal with "false prophets" and "servants of Satan." Such warnings were placed at the end of writings of that time.

Several abrupt editorial seams that resulted from an arrangement of a letter of reconciliation, an apology on the nature of Paul's apostolic authority, a polemic against opponents, two letters concerning the collection, and a possible non-Pauline insertion (in chapter 6, verse 14, to chapter 7, verse 1) can thus be understood. The reconciliation of chapters 1 and 7 is hardly in agreement with Paul's elaborate defense of his ministry in chapter 2. Even more jarring to such a reconciliation is the polemic of chapters 10–13. These latter chapters are viewed as a substantial fragment of Paul's "letter of tears," after which the Corinthians disengaged themselves from outside agitators and caused them to leave. Such opponents, who are mentioned in chapter 11, verse 4, and who tried to attract the congregation away from Paul's ideas, were probably Hellenized Jewish Christians from Palestine.

The problem of outside agitators

The outside agitators (who provoked the response of chapters 10–13) probably were Christians who imitated the Hellenistic-Jewish missionaries and had developed an elaborate propagandizing missionary theology and practices analogous to the missionary movements in the pagan world. Their goal was to prove the spiritual power of their own religion in conscious and aggressive competition with other religions, thus hoping to attract others and convert them to Christianity.

The major criteria for successful competition were affinity or identity with the ancient Mosaic traditions and objective manifestations of the current power of that tradition in the form of miraculous demonstrations. The link between the ancient traditions and the current careers of the itinerant missionaries was the record of Jesus as understood from the miracle stories of the Gospels—a demonstrated epiphany of the powers of the Spirit. These missionaries were seen as "divine men," as were the heroes of old. Their miracles were to be imitated. Such traditions about Jesus as a wonder-worker might have been used by

Paul's opponents, with over-emphasis on such works as criteria of power.

Paul's attacks on the "super-apostles"

That which Paul attacks as "bragging" or "boasting," particularly the preaching of the so-called "super-apostles," in chapter 11, verse 5, was probably understood by his opponents as no more than faithful testimony to, and a demonstration of, the spiritual powers of tradition as they perceived it in their own experiences. To them faithfulness to Jesus was primarily the acknowledgment of Jesus' being the most powerful "divine man" and, secondarily, their establishment and maintenance of relationship to him through imitation in their powerful demonstrations and wondrous acts.

Paul (who in I Corinthians, chapter 1, had advocated the dialectic of the cross) would thus be discredited by miracle-working men like the opponents in II Corinthians. Paul's credibility and validity as an Apostle came into question along with his Christology, which was a "theology of the cross." Confronted with the challenge of the powerful "super-apostles," Paul's message could be distorted as hiding his own inability or incapacity—an apostle who dared not take money because, being an ineffective speaker and a weak person, he had nothing for which to ask payment. His defense was Paul's first attempt to deal with these new problems caused by invading opponents who had undercut his authority.

Paul's defense of his own apostleship

Paul centred his defense around the issue most debated; true apostleship and his own sufficiency. Because he derived his ministry from God himself as a servant preaching not himself but Jesus Christ as Lord, no "peddler of God's word or selling or recommendation is called for, but only the living record—*i.e.,* the people brought to believe in Christ. Paul quickly alluded to his own weakness and "carrying in the body the death of Jesus, so that the life of Jesus may also be manifested . . ." (chapter 4, verse 10). Paul found his weakness one of the things that made him one with the Lord and that made his ministry a true ministry of Jesus Christ, who was crucified through weakness but lives by the power of God—as does his true apostle. This weakness seems to refer to a physical handicap of Paul's (epilepsy?), the "thorn in the flesh" that interfered with his travel plans.

Paul placed his own apparent weakness, in which he proclaimed that God had manifested himself, against the boastings of the "super-apostles." Unlike them, he strikes a non-heroic note. It is confidence in the power of Jesus' Resurrection that produces glory for the Gospel message and final (eschatological) reward and recognition for the Apostle.

Though Paul may himself sound "enthusiastic," his statements are made with a realistic assessment of the world, as demonstrated not least in the sufferings of Paul himself. Emphasis on God's act of grace, however, makes Paul urge the Corinthians to accept him and to reach out to the promise of God's salvation even in the present.

Paul's defense of his apostleship and a following visit did not succeed. Agitation from outside opponents apparently increased and solidified. The "letter of tears" reflects this situation. Paul revealed himself personally, coming close to autobiographical statements. Paul spoke of himself only with theological purpose and as part of his tactical argument with his opponents concerning attitudes and conduct. His point was that a style of life is a reflection of an underlying theology. He demonstrated to his opponents that his work for the church is constructive, and that though he boasted of his ministry, he boasted only "of the Lord," of the work Christ had done through him.

Paul's use of the technique of the "fool's speech"

In his so-called fool's speech, in which he blatantly asked the Corinthians to "bear with me in a little foolishness," Paul adopted the technique of the mime of the street theatres of his times, consciously drawing on the laughter and mockery of his audience, but then he successfully reversed the scene and made his audience realize that in laughing at him they mocked themselves, thus revealing the perversion of their criteria of superiority. Paul used metaphorical images, identifying the congregation with the bride, Jesus as the bridegroom, himself as the best man, and Satan (the opponents) as the adulterer. The plot assumed a successful seduction, and the best man who recommended the bride stands disproven. Paul then pretended to try to shift this balance by bragging about himself and scolding both seducers and the seduced. He accepted no inferiority to the opponents—the seducers ("super-apostles")—and claimed that they preached another Christ than the true Christ and brought another spirit and that he would accept no support from the church that was led astray.

In chapter 11, Paul continued to boast "as a fool," claiming to have all the qualifications of his opponents, but that he was more truly a representative of Christ. This he explained ever more intensely in an ironic and almost sarcastic trend in the dialectic of the so-called fool's speech. He boasted not of strength but of weakness—though he could boast of ecstatic experience as his opponents had—and that he had learned through bitter experience (possibly a chronic illness) that he must not exalt himself, but rather that he has been told through a word of Christ that his power is made perfect in weakness. In the enumeration of his qualifications, Paul has jested "as a fool" concerning his suffering, visions, miraculous heavenly travels, and oracles. Yet, it is clear that through Christ these modes of experience and communication have been transformed. Thus, Paul establishes that he is a true apostle and not inferior to the "super-apostles."

Paul's intention to visit the congregation again

Paul expressed his intention of visiting the congregation and told them that he desired to come not as a judge but as a father. Neither he nor Titus had or would deceive or take advantage of them. At this, the end of the "letter of tears," Paul announced his possible third visit and revealed a definite fear that he might be forced to act as a judge of the congregation, which was increasingly falling away from the apostolic gospel. Paul, however, still hoped that reconciliation might be accomplished, that truth would prevail, and that his authority could be used for building up rather than destruction. He exhorted the community to keep peace and blessed them.

The "letter of reconciliation," found in chapters 1, 2, and 7, assumed that Titus had returned with good news of the Corinthians, their eagerness to prove that they had amended their ways. Paul responded with a report of the consolation this had brought him and of the grave danger he had escaped (in prison in Ephesus). He exhorted the church at Corinth to remember the Christian message in love—of Paul for them and of the congregation for him. The shadow between Paul and the Corinthians had been dispersed, and Paul reaffirmed his constant and continuous concern for them and God's love in Christ manifest in Baptism and the gift of the Spirit. Paul interceded for a man who had offended him and forgave him. Paul then told the Corinthians of his eagerness for Titus' news of them that occasioned his special trip to Macedonia. This news brought joy and consolation; therefore, Paul urged the Corinthians again to forgive the man who had offended him.

Fragments of two letters concerning the collection for Jerusalem, a sign of unity of the church (chapter 8 especially being close to the "letter of reconciliation" and chapter 9, a fragment probably later than chapter 8), are signs that Paul's relation to the Corinthians again became close and joyful. The collection was a bond of mutual and reciprocal relationship that reached its climax in thanksgiving and praise of God. For the whole church he exclaimed: "Thanks be to God for His inexpressible gift!"

The Letter of Paul to the Galatians. Paul's Letter to the Galatians is a forceful and passionate letter dealing with a very specific question: the relation of Jewish Christians and Gentile Christians in the church, the problem of justification through faith not works of the Law, and freedom in Christ. Paul probably wrote from Ephesus *c.* 53–54 to a church he had founded in the territory of Galatia in Asia Minor.

The question of the relationships between Jewish and Gentile Christians

This congregation had been "unsettled" since his last visit to Galatia. Gentile Christians, Judaizers who were fascinated with Jewish customs and festivals and who asserted that Gentiles must adhere to the Law, the Torah, had attempted to undermine Paul's message and effectiveness. The Judaizers believed that Gentile Christians should be circumcised and keep the Jewish food laws. There were

probably some Jewish Christians in this church, but the majority were Gentile Christians. Paul attacked the Judaizers vigorously by defending his own call and the independence of the revelations of his personal apostolate. This is supported by reports of agreement between him and the Jerusalem church and by argument from Scripture. In these, he proved that the Law was given only a limited role in the total history of salvation. The letter ends with Paul pointing out that through the Spirit the Christian in faith is admonished to good behaviour and brotherly love. He admonishes faith in the cross of Christ, wishes peace upon his followers, and prays for mercy on Israel.

This Pauline letter is the only one without either kindly ingression, thanksgiving, or personal greetings appended to the final blessing. It is very specific in dealing with the problems concerned. In chapter 1, an account of Paul's call, he defended his apostolic office, having received it directly from God in the revelation of Christ. He provided autobiographical data concerning his former persecution of the church and zeal in his Jewish tradition. He referred to his call on the model of that of the Old Testament prophets called by God in order that they may serve him and said that his mission had been revealed to him to be the apostle to the Gentiles. Paul viewed himself as being chosen to be an instrument to take the message of God and Christ to the Gentiles, a call rather than a "conversion experience." Handpicked as God's servant (slave), he received a revelation—not from men but by secret knowledge from God—that the Gentiles will come to the Christian faith without the Law, the Torah of the Jews. He himself could bear the Law, but he was told that the Gentiles do not need the Law in order to be accounted righteous. The conviction that the Gentiles stand equal before God was reinforced by his visit to James, Cephas (Peter), and John in Jerusalem, who confirmed his mission, enjoining him only to remember the poor (probably reference to the Jerusalem collection). Faith in Christ has thus superseded righteousness of works, and the Law is no longer needed.

The freedom of the gospel

The freedom of the gospel is the theme developed in chapters 3–4 in a series of allegorical-typological interpretations based on the Law. Paul first recalled the covenant promise to Abraham: that he "believed God and it was reckoned to him as righteousness" and that through Abraham all nations would be blessed.

In chapter 3 there is a complex line of thought: Christ has redeemed men from the curse of the Law by becoming a "curse" for men; Christ has taken away this curse by accepting it himself in order that all men by faith might receive the Spirit that was promised. But the promise had already been made to Abraham and his seed (singular), the Messiah, Christ; the Law had come only 430 years later, a sign that it is not eternal. In this chapter, Paul constructed arguments against the Law. First, the Law was added because of transgressions committed first by the people who caused Moses to shatter the first tablets of the Law and was thus not ultimate but rather time-bound, limited, and tainted by the evil reality it had to counteract; secondly, the Law was given only for a restricted time, from Moses "till the offspring should come to whom the promise had been made" (*i.e.,* Christ); thirdly, the Law came "ordained by angels through an intermediary," who is not God and thus is neither something glorious in itself nor the absolute manifestation of the salvation of God. Paul expanded on the Law in the image of a *paidagōgos* (instructor or custodian). Such a custodian is now not needed and served only as a restraint so that in God's timetable of salvation the Gentiles could be delivered after the Law has been "outgrown." Paul then showed the reasoning behind his statement that the Law was obsolete: in Christ (*i.e.,* in the church) there are no divisions between Greek and Jew, slave or free, male or female—all divisions or partitions are broken down.

Freedom from the elemental powers of the universe

Paul's arguments are bold. He even claimed that, as heirs through Christ, men were no longer bound under the elemental powers of the universe, which were apprehended as negative, as was the Law, in Paul's mind. In chapter 4 the Judaizers are said to keep themselves, like many Greeks, under astrological powers—not unlike the Jewish calendar of feasts—which kept man, according to Paul, enslaved by cosmic order. But to those free from the Law and possessing the Spirit, sonship and inheritance can come by adoption. Thus, Paul was negative in Galatians concerning the Law, and taught that freedom from it brings unity and the fruits of the Spirit.

In chapters 5–6 Paul listed catalogs of virtues and vices, fruits of the Spirit or the flesh, and stressed mutual forgiveness in the church. This is an exhortatory section that leads to the closing of the letter in Paul's own hand and to his stress on seeing his only glory in the cross of Christ.

The Letter of Paul to the Ephesians. The authenticity of Ephesians as a genuinely Pauline epistle has been doubted since the time of the Dutch Humanist Erasmus in the 16th century. It is most reasonable to consider it as "deutero-Pauline"—*i.e.,* in the tradition of Paul but not written by him. The problem of Ephesians cannot be solved apart from that of Colossians, because many similarities are noted in the style and development of Pauline thought into cosmic imagery; yet they treat different problems. In both, the heritage of Paul is preserved by a "Paulinist," and it is on this basis that Ephesians and Colossians were accepted into the canon. Both are "captivity epistles," ostensibly written by Paul from prison. Of the 155 verses in Ephesians, 73 have verbal parallels with Colossians; and when parallels to genuine Pauline letters are added, 85 percent of Ephesians is duplicated elsewhere. It would appear that Ephesians is dependent on an earlier, more specifically oriented Colossians, and it may be that Ephesians uses, combines, and condenses the material of Colossians for its own needs.

The universal appeal of Ephesians

Though Colossians is directed explicitly and strongly against a particular Judaizing proto-Gnostic heresy—*i.e.,* an incipient form of a religious dualistic system that emerged as a very attractive heretical movement in the 2nd century—Ephesians is not polemically oriented and is not clearly connected to a particular congregation, its problem, or its individuals. Though Ephesians uses a letter style with an introduction, greeting, and closing benediction, the only person mentioned in it is Tychicus, already mentioned in the same context in Colossians. The doctrinal section shows that the whole world—not only the Jews—is in a cosmic sense subjected to Christ, and Jew and Gentile are reconciled and united through him. This is the mystery of God's plan revealed to the church through Paul but expanded in scope. All are saved and reconciled through Christ, who has made both Jew and Gentile one and has "broken down the dividing wall of hostility," bringing peace and unity. The author of Ephesians continues Pauline language and makes it more Pauline than Paul himself.

After the address—which, according to the best manuscripts, lacks a reference to Ephesus—there is a hymn of praise to God in terms of a cosmic plan of redemption. Through the ascended Christ, salvation is for all, and he is the head of the body, his church. Because the address and thanksgiving are to the church in general (the place name, Ephesus, being an early gloss), it is possible that Ephesians was meant as an encyclical, to be distributed, perhaps, as a covering letter for the whole Pauline collection. The "mystery of God's will" (chapter 1, verse 9) is spelled out in chapter 2 as the reconciling act of Christ for both Gentile and Jew. In chapter 3 Paul's role in giving knowledge of this mystery in his ministry leads to a doxology. After this semi-epistolary form, the general admonitions follow in terms of gifts of grace with stress on unity: one hope, one Lord, one faith, one baptism, one God for all. A warning against a heathen way of life is given in contrast with the Christian's old nature as opposed to his new being in Christ. In chapter 6, verses 10–20, the Christian is enjoined "to put on the whole armor of God" as defense against evil and Ephesians ends as a letter, with a blessing.

The Christology and ecclesiology imply a background of a Christianized, mythological proto-Gnosticism, or a strongly Hellenized Judaism. Perhaps one of the best clues to the lateness and pseudonymity of Ephesians in comparison with the genuine Pauline letters, however, is the phrase "revealed to his (Christ's) *holy* apostles and prophets by the Spirit." Such an expression is certainly

later than Paul and looks back on the apostolic age as a time in the past.

The date and style of Ephesians

A possible date is shortly after Colossians, in the early 2nd century. Because there are so many similarities to Colossians, Asia Minor might be the place of composition, but this is merely conjecture. The non-Pauline use of the term mystery to denote that Gentiles are fellow heirs with Jews, the uniting of all in Christ, and an analogy between marriage and Christ's relation to the church, all point to a different and later time than that of Paul. The style of Ephesians builds up long, almost unmanageable, unpunctuated, excited, and abundant sentences, even longer than those of Paul when he is most provoked or, perhaps, absentminded and does not finish sentences that he begins. A comparison of the table of duties of Colossians 3 and Ephesians 5 and 6 also shows a strong development in the direction of making the relationship of Christ and his church the basis for all other relationships.

The eschatology of Ephesians is attenuated, if not far in the background, and a continuation of the church is implied. In chapter 1, verse 13, the writer sees the Spirit as the guarantee (down payment) of the Christian's inheritance—a present indication through the Spirit that the Christian can live in faith in the world looking for the Kingdom but already sure he can draw on the powers thereof without an imminent expectation of the end-time. Ephesians gives hope for universal salvation, grace as a gift of God, strength in patience, and an example of unity for the church as well as freedom in the Spirit to attain maturity as a Christian.

The Letter of Paul to the Philippians. In its present canonical form Philippians is, according to several scholars, a later collection of fragments of the correspondence of Paul with the congregation in Philippi that was founded by Paul himself. The first of the two major difficulties leading to this conclusion concerning redaction of the letter is created by a discrepancy between chapters 2 and 3—*i.e.,* an entirely unexpected polemic in chapter 3 after a calm second chapter. Another major difficulty is the relationship of chapter 4, verses 10 and following, with Paul's joyful acceptance of his suffering, and the remainder of the present letter that deals with the collection the Philippians had made and sent to Paul in prison. The place of the expression of Paul's gratitude at the end of the letter is odd, particularly because Epaphroditus, the Philippian delegate conveying the gift, is thanked as though he had just arrived; yet he has already been described as ill when he was with Paul (who apologized in chapter 2 for not having told about Epaphroditus' illness sooner and the delay in sending him back). Yet, Epaphroditus is obviously back and the sequence of events is, indeed, confusing.

The fragmentary condition of the letter in its present form

The following rearrangement of the parts of the letter is probably acceptable. Chapter 4, verses 10–20, shows Paul reacting to the gift of the Philippians and the arrival of its bearer, Epaphroditus, and seems to be the earliest fragment, written probably during Paul's imprisonment (*c.* 53–54). The portions of the letter that treat of the theme of mutual joy (1:1–3, 4:4–7, and probably 4:21–23 that refers back to chapter 1) are best taken together as fragments of a second and somewhat later letter. The third section is 3:2–4:3 and possibly 4:8–9, which addresses the danger caused by outsiders and opponents who had started to penetrate the Philippian congregation with a theology Paul considered heretical and against which he aimed his polemic. Because this is an entirely new situation, it is probably a third letter, of which only the preface is missing. This arrangement also attempts properly to account for the fact that chapter 4 actually comprises endings of several letters. Thus, chapter 3, verse 1, which is itself a summation and ending, fits in.

The reference to frequent visits between Paul and the Philippians referred to in the correspondence makes its origin in Rome unlikely and points rather toward Ephesus as the place of imprisonment. Paul's reaction to the gift of the Philippians is almost rude (although he accepted gifts from no other congregation but preferred to support himself during his apostolic mission). He actually avoided expressing direct gratitude and attempted to divert the significance of the gift from its material side to its spiritual meaning. He emphasized the sympathy proven by the Philippians, the importance of the value of the gift for them as a spiritual sacrifice for God.

The "letter of joy"

The "letter of joy" section describes Paul's enthusiasm in his mission efforts—and their success—and his joy in the energy and growth of the mission in Philippi, which Paul shared with his congregation. Paul's address to "bishops and deacons," terms unique in Paul's letters except here, are, perhaps, circumlocutions for missionaries active in Philippi, a congregation that had become a strong and stable Christian community. Paul had traditionally remained there about one week and, in chapters 1 and 2, encouraged and praised the Philippians for continuing in their faith in his absence. This is part of the thanksgiving in Philippians—an emphasis on the participation, cooperation, collaboration, and empathy of the Philippians with respect to the preaching of the gospel. Thus, the terms bishop and deacon may belong to the language of a self-supporting mission church with its own overseers (bishops) and workers (deacons) and does not carry the connotations of later ecclesiastical structures. Paul expressed his confidence in the fine beginning of this young church that sought "to become pure and blameless for the day of Christ," the final judgment.

Paul then turned to his own experience of imprisonment, which he viewed as advancing the gospel. Though he considered that not all preachers of Christ preach on the basis of selfless motives, the fact that Christ is proclaimed is a most important cause for rejoicing. Paul then exhorted the Philippians to work hard for the sake of the gospel, not minding any opposition, and to do this in a sense of unity and mutual support.

This exhortation toward a strong and active sense of community was reinforced by quoting an early Christian hymn that described the humiliation (*kenōsis*) and exaltation of Jesus who is made the Lord of the universe and confessed by all cosmic powers. A part of Jesus' humiliation, his death on the cross, can be taken as part of his manifest glorification. The verses following the hymn make clear that the incorporation of the hymn with its triumphal ending also has a missionary purpose, because Paul emphasized again the need to responsibly act out one's own calling even before non-Christians. Thus, active responsibility continuously exercised in the perspective of the approaching Parousia merges with Paul's own readiness to sacrifice himself.

The missionary emphasis

In chapters 3–4 the situation may be totally different. Paul reacted to the threat of the appearance of Jewish-Christian missionaries who are rather close in theology to the Galatian Judaizers. Paul's polemic indicates that in addition to Jewish tradition, they must have emphasized the Law in particular. Reference is made to circumcision, and Paul emphatically claimed that he could compete with heretics boasting of their Jewish tradition and, in elaborating on that, emphasized his former pious righteousness under the Law, in which he was blameless. He then stressed categorically that for him the experience of Christ has terminated his former piety completely and that he has left it behind as of no value. Such a polemic implies that for his opponents such was not the case. Paul also argued against libertinistic tendencies, which indicates that his opponents were not legalists in an ordinary sense but combined faithfulness to the Law with a strong and fanatical enthusiasm that could lead toward "mysticism" and easily be misinterpreted as libertinism. Paul's emphasis on true Christian experience as not being completed but rather still being in the state of expectation might be a further polemic against overenthusiasm. In chapter 4, verse 8, Paul reaffirms his own example, making it, in imitation of the teaching of popular philosophy, the epitome of all positive ethical values and virtues, and thus the pattern to be imitated. This tendency toward the paradigmatic, together with warnings and autobiographical material in chapter 3, verse 2, to chapter 4, verse 3, can be seen as a "testament" of Paul, consciously written with an awareness of impending death or martyrdom. Thus Paul presents himself—his life, ideas, admonitions, and an eschatological section—as his heritage and as an incorporation of the message he preached and its value.

The Letter of Paul to the Colossians. Colossians presents the problem of having, on the one hand, numerous (though superficial) affinities with the circumstances of the Letter of Paul to Philemon while, on the other hand, being addressed mainly to a different situation. In this new situation he uses ideas and expressions that seem to be rather a development of Pauline ideas about the cosmic realm than genuinely Pauline argumentation. In this latter aspect, Colossians and Ephesians share the heritage of Paul, but a later "Paulinist" changed details to meet different situations.

The purpose of Colossians

Colossians was written ostensibly by Paul from prison (in Ephesus) to a predominantly Gentile Christian congregation founded by his co-worker, Epaphras, at Colossae. The Colossian congregation was endangered by a heresy involving a "philosophy" that was connected with the elemental spirits of the universe to which men seemed to be bound, with circumcision, feast days and food laws, visions, and an asceticism that was not only false in its piety but foreign to the Christian faith.

To combat these proto-Gnostic, syncretistic, and Judaizing tendencies, the Paulinist appealed to the authority of Paul's apostolate and his thought but accented his theology in a new way, enlarging Paul's theological dimensions, so that they included the whole universe, the fate of the entire cosmos. This whole world is depicted as subject to Christ and has its meaning, aim, and goal in the church, which is Christ's body and over which he is the head. This transformation of Paul's theology would appear to be somewhat later than Paul, yet not so much later than Philemon, and its import has been forgotten. Colossians cannot be dated or placed with certainty, but the end of the 1st century or the beginning of the 2nd century has been suggested.

In a first edition, before the Paulinist changed or added to it, Colossians seems close to the situation of Philemon. In both letters Paul is in prison. Onesimus appears in Colossians, chapter 4, and the readers of Colossians are asked to transmit a special injunction through the church of the Laodiceans to Archippus—possibly that the former slave, Onesimus, now referred to as a "beloved brother," be freed for service of the gospel. The same five names appear in Philemon and Colossians (Col. 4:10 ff.; *cf.* Philem. 23), which is unusual because the church at Colossae is strange to Paul. The lost letter to the Laodiceans may possibly be the Letter to Philemon, and the request to the slave owner would, by being read aloud in a neighbouring large church (Colossae), reinforce Paul's request that the slave be freed.

Later substantial redaction has obviously taken place, however, and it is the heresy at Colossae rather than the situation of Philemon that is mainly addressed in Colossians. Though Paul asserted that he did not preach and exhort where another has founded a church, here the Paulinist, using and amplifying Pauline theology, taught, gave thanks, and interceded for a church that he did not found and that was in danger of accepting heretical Judaizing teachings, thus falling away from Christ. The doctrinal section of Colossians sets forth in a hymn Christ's preeminence over the whole cosmos, all principalities and powers, to bring redemption through the cross and to be the head of the body, the church.

From this cosmological beginning, the style and imagery differ from the authentic Pauline letters. Colossians is wider and broader in scope, with long, almost breathless sentences. There is a hierarchy in Christ being head of the body, his church, which differs from the Pauline expression of equality of all the members, although with differing functions (*cf.* I Corinthians, chapter 12, and Romans, chapter 12).

Christological and ethical emphases

The Christology is applied to the situation of the church and Paul's role in behalf of the church—his suffering with Christ and knowledge of God's mystery, Christ—is used to bolster his defense against heresy. This polemic is based first on tradition and then proceeds to specific warnings against false teaching, cult, or practice. An admonition "to set your minds on the things that are above," because in Baptism the Christian has died and been raised with Christ, is followed by the conclusion that the Christian's conduct should be ruled by love and be thus free from all wrongdoing.

Another difference from the genuine Pauline letters can be noted in this latter section. When Paul referred to the resurrection of Christians he used the future tense in most cases, but Colossians, chapter 2, verse 12, and chapter 3, verse 1, presuppose that because the Christian is risen with Christ, ethical demands can be made.

In Colossae, such Christian ethics apparently were lacking, thus the inclusion of a table of duties—*i.e.,* a list of household duties and of relations between members of a household. General exhortations to prayer and right conduct are followed by the conclusion of the letter with its list of greetings. There are some similarities in Colossians to Paul's polemic against Judaizers in Galatians, but Colossians seems to reflect a later time and a more developed "cosmic" theology of a later deutero-Pauline writer.

The First Letter of Paul to the Thessalonians. In all probability I Thessalonians is the earliest of Paul's letters, particularly because the memory of the events that led to the founding of that congregation are still fresh in the mind of the Apostle. The letter was written from Corinth. According to I Thessalonians, chapter 3, verse 2, Paul had sent Timothy to Thessalonica from Athens during his brief stay there, had just experienced the delegate's return, and had received reports about the congregation to which he is reacting in this letter. I Thessalonians gives expression to Paul's surprise over the rapid growth of the Christian mission at Thessalonica, which was achieved despite immediate persecutions from pagan contemporaries. Paul acknowledged that the successful development had been wrought in the Thessalonians by their own acceptance, fully recognizing the human frailty of the Apostle, their founder (2:1–12), and not by a mistaken understanding that he himself was divine.

The rapid growth of the church at Thessalonica and Paul's response to eschatological expectations

Paul's surprise results, therefore, in overwhelming gratitude, and the customary Pauline thanksgivings here exceed the usual limits. A second reason for this unusually long thanksgiving—which actually makes thanksgiving the theme of the letter—is Paul's intent to undergird the encouragement he gives in 4:13–5:11. After having dwelt so extensively on his being moved by the change in the Thessalonians, Paul continues to state that therefore they have no reason for giving up faith in the face of the death of some fellow Christians, who had died between their conversion and the expected imminent Parousia of Christ. Apparently, they had expected the Parousia and final salvation as the promise of the Christian message. Paul encouraged his congregation that he had a "word of the Lord" that the dead and the living in Christ will rise together. "Word of the Lord" could refer to a word of Jesus known to Paul but could instead be a direct revelation to Paul.

In chapter 5 there is further thanksgiving, emphasizing the present gift and power of Christian faith and corporate Christian life. This emphasis is linked with ethical applications, with stress on brotherhood, diligence in keeping the faith, and religious industriousness. The difficulties of balancing the expectation of the Christian with God's timetable is outweighed by the hope and joy in what has already been experienced and what is hoped for. Paul's real emphasis is more on the actual description of Christian life in the face of coming salvation and vindication than on the preceding discussion of the fate of those who had died or on the actual circumstances of Christ's appearance from heaven.

The encouragement of the Thessalonians was introduced in chapter 4 by a genuinely ethical exhortation to proceed properly on the way to holiness and sanctification already begun. The brevity of this rather traditional exhortation is most unusual in Paul's letters and supports the observation that it was written in joy and confidence for a new congregation well begun in order to support it against attacks and doubts as it matured in the faith.

The Second Letter of Paul to the Thessalonians. A feature of II Thessalonians that resembles the otherwise most unusual feature of I Thessalonians is its excessively long thanksgiving. Within this thanksgiving there is an excursus dealing with the timing of the Parousia, but in

Corrections about apocalyptic expectations

II Thessalonians Paul aggressively argues against any expectation of an imminent coming of Christ that might be expected from the things he wrote in I Thessalonians. II Thessalonians perhaps presupposes I Thessalonians and intimates that believers had a false understanding of that communication of Paul. In II Thessalonians, much to the surprise of the reader of both letters, the statement is made that a letter "purporting to be from us" is "to the effect that the day of the Lord has come." II Thessalonians then presents a problem as to whether it was a self-correction of Paul or directed to the situation of a later time and thus the writing of a later author in a "Pauline" tradition. II Thessalonians does have more apocalyptically catastrophic language than I Thessalonians. Such a description not only underestimates the positive work of God and Christ for the believer but also says little about the Parousia. II Thessalonians claims that not all the events preceding the Parousia have yet occurred. The "mystery of lawlessness," opposed to the "mystery of godliness," is still at work in the world, and the full activity of Satan has not yet unfolded itself. Emphasis in II Thessalonians is on steadfastness as God's gift and promise in the days of tribulation, which makes the apostle ask for support in prayer. Criticism of people leading disorderly and idle lives follows. The perhaps casual admonition to work is thus elaborated into a major point.

Salvation seems to be sought almost exclusively in futuristic terms. Incipient or actual Gnosticism in the church could account both for the assertion that the fulfillment has already come and for the depiction of disorderly lives (because in "proto-Gnostic" terms the world is evil and provokes a response either of total renunciation or libertinism). II Thessalonians may thus reflect these problems and fit into the late 1st century. Verbal agreements between the two letters may be evidence of deliberate spurious writing, as also the suggestion in II Thessalonians that false letters may be circulating. A later author saw Paul's heritage threatened by too enthusiastic an understanding of Paul in Thessalonians and composed this letter to preserve Paul's meaning.

THE PASTORAL LETTERS: I AND II TIMOTHY AND TITUS

The Pastoral Letters as a unit. The First and Second Letters of Paul to Timothy and the Letter of Paul to Titus, three small epistles traditionally part of the Pauline corpus, are written not to churches nor to an individual concerning a special problem but to two individual addressees in their capacity as pastors, or leaders of their local churches. The purpose of the letters is to instruct, admonish, and direct the recipients in their pastoral office. Since the 18th century they have been referred to as a unit, the Pastoral Letters, and they contain common injunctions to guard the faith, to appoint qualified officials, to conduct worship, and to maintain discipline both personally and in the churches. Their similar peculiarities of style and vocabulary as well as the similarity of the heresies and other problems they faced place them in a common time and allow them to be dealt with as a unit. Their content presents a picture of the post-apostolic church when pastoral offices and tradition came to the fore and the formerly high apocalyptic tension appears attenuated.

Reasons for accepting a non-Pauline authorship

The Muratorian Canon (a list of biblical books from *c.* 180) includes references to the Pastoral Letters and notes that they were written "for the sake of affection and love." They have a place in the canon because "they have been sanctified by an ordination of the ecclesiastical discipline." These letters, however, do not appear among the Pauline letters in P 46, an early-3rd-century manuscript, and there is no clear external attestation in the primitive church concerning them until the end of the 2nd century. Not until the 19th century were doubts expressed about the Pastorals as being authentically Pauline, when German scholars and others noted discrepancies in style and vocabulary, church organization, heresies, biographical and historical situations, and theology from those found in the Pauline letters. The problems of authorship, authenticity, and dating almost paralyze investigation of the Pastorals unless discussion of these problems is seen as connected also with the literary character of the material.

Attempts have been made to apply the tools of statistical analysis in comparing these disputed letters to the rest of the New Testament (particularly to the Pauline corpus) for the purpose of establishing authorship. The studies, utilizing computer technology, point toward non-Pauline authorship with affinities to language and style of a later, possibly 2nd-century, date. More refined and complex analyses, however, are still needed.

Linguistic facts—such as short connectives, particles, and other syntactical peculiarities; use of different words for the same things; and repeated unusual phrases otherwise not used in Paul—offer fairly conclusive evidence against Pauline authorship and authenticity.

Content and problems. Church offices are more developed in the Pastoral Letters than in Paul's time. There are presbyters and bishops, but these are sometimes used interchangeably and the monarchical episcopate is not yet depicted, although church offices appear to be heading in that direction. Requirements for office are strict and leaders are chosen and ordained by laying on of hands. Such leaders must be able to teach true and sound doctrine and guard what has been entrusted to them, the *parathēkē*—*i.e.,* the deposit of teaching or the message to be carried on. They must also be able to stand firm and argue against heresy. Such offices and aims suggest an expectation of future generations of faithful witnesses to carry on the traditions, perhaps particularly necessary as some may be killed for the witness they make.

Internal organizational and theological developments

The heresies referred to appear to be Gnostic and the arguments are rather mild and reasonable, unlike Paul's urgency in combatting heresy with strenuous argumentation. The heresies taught by false teachers are an early partly Encratitic (abstaining) Gnosticism, with "higher knowledge" that emphasizes "godless and silly myth," or are statements that the resurrection has already taken place, which is a denial of future resurrection and a glorification and spiritualizing of resurrection as a rebirth, as, for example, in Baptism.

Biographical notes about Paul's journeys and situations contradict his own letters as well as the accounts in Acts. The Pauline sense of living in a time close to the end of the age is missing in these descriptions of churches; they are viewed as settling down with a succession of tradition with Hellenized expressions of salvation and a replacement of enthusiasm with bourgeois ethics. This indicates a period of de-emphasized eschatology and an expectation of a long community life in which people must live out their lives in Christian responsibility and moral behaviour.

I Timothy and Titus are more similar to each other than they are to II Timothy, but all three mark exhortations to personal lives of exemplary conduct and give rules of conduct for church order and discipline for the group as a whole and for individual parts of it—sometimes in terms of catalogs of virtues and vices recalling the Jewish two-way orders: the way of life being good, the way of death including a list of sins. Each concludes with a final blessing or salutation. They are all pseudonymous, using Paul as an epistolary model and using pseudonymous devices, such as naming individuals known to be Paul's co-workers. The authority of Paul is invoked to lend authority to the teachings contained in the letters: the avoidance of heresy, holding to sound doctrine, and piety of life. The author is anonymous, the place of writing and the addressees are unknown, but they probably are later spiritual children of Pauline teaching. The date of the letters is about the turn of the 2nd century.

II Timothy uses the background of Pauline imagery most fully. It is cast at least in part in the testament form to Timothy as his spiritual heir because Paul is depicted as suffering, fettered in prison, and awaiting the martyr's crown. He exhorts Timothy and through him the church to share in these sufferings as they will eventually share in glory. II Timothy, chapter 2, verses 1–13, is an exhortation to martyrdom with a faith that Christ, triumphant over death, will save his faithful witnesses. Recollection of the creed is followed by a direct application to bearing suffering and its meaning in God's plan of salvation. The words "faithful is the word" occur in 2:11. This "word," unlike Paul or any Christian, cannot be bound. It both

Early liturgical Christological hymns

confirms salvation described in the preceding verses and introduces a hymn that may represent liturgical usage in that it is poetic and balanced.

> Faithful is the word:
> If we have died with him, we shall also live with him;
> if we endure, we shall also reign with him;
> if we deny him, he also will deny us;
> if we are faithless, he remains faithful—for he cannot deny himself
>
> (II Tim. 2:11–13)

The hymn preserves within itself a reflection of sayings of Jesus that those who endure and persevere will reign with the Lord and that even to those who deny him (as did Peter) God will remain faithful because Christ cannot deny his own faithfulness. Even in this hymn there is allusion to a "testament" form, with Paul already martyred, as a pseudonymous device to spur the Christian on to endurance and faithfulness as a member of the redeemed community.

Another small poetic hymnic section serves to demonstrate that the church of the Pastorals, albeit somewhat deeschatologized, retains the "mystery" in God's household, the church—*i.e.,* the gospel and creed alive in the liturgy in the mystery of piety and worship.

> Great indeed, we confess, is the mystery of our religion:
> He who was manifested in the flesh,
> vindicated in the Spirit,
> seen by angels;
> who was proclaimed among the nations,
> believed in throughout the world,
> glorified in high heaven
>
> (I Tim. 3:16)

Here in miniature are creed and gospel somewhat reminiscent of the Gospel According to Matthew.

Paul's views of Christian brotherhood transcending slavery

The Letter of Paul to Philemon. From Ephesus, where he was imprisoned (*c.* 53–54), Paul wrote his shortest and most personal letter to a Phrygian Christian (probably from Colossae or nearby Laodicea) whose slave Onesimus had run away, after possibly having stolen money from his master. The slave apparently had met Paul in prison, was converted, and was being returned to his master with a letter from Paul appealing not on the basis of his apostolic authority but according to the accepted practices within the system of slavery and the right of an owner over a slave. He requested that Onesimus be accepted "as a beloved brother" and that he be released voluntarily by his master to return and serve Paul and help in Christian work. Paul appealed to the owner that Onesimus (whose name in Greek means "useful") is no longer useless because of his conversion and claimed that the owner owed Paul a debt (as he probably was also instrumental in his conversion) and that any debt or penalty incurred by the slave would be paid by Paul. Such manumission is part of Paul's concept of being an ambassador to further the mission of Christianity, rather than a judgment on the social framework of slavery, because in the Lord such social order is transcended.

Philemon, however, is not a purely personal letter, because it is addressed to a house church (a small Christian community that usually met in a room of a person's home), and it ends with salutations and a benediction in the plural form of address. The body of the letter, however, uses "you" (singular) and is addressed to the slave's owner, a man whom Paul himself has not met. Philemon, the first name in the address, is called a "beloved fellow worker," which implies that he knew Paul, and it has been convincingly argued that the slave's owner was Archippus (see above *Colossians*), perhaps Philemon's son, who was called a "fellow soldier," a term usual in business accounts and suitable for a document on the manumission of a slave. The thanksgiving contains the main theme of the whole letter: sharing of faith for the work of promoting knowledge of Christ.

The letter was written from prison, and Paul apparently expected a release in the near future, because he requested a guest room, a suggestion that he was not very far from Colossae or Laodicea, which would be true of Ephesus. Colossae would be reached from Ephesus via Laodicea, and the letter could be addressed to a house church there.

The role of Onesimus in the early 2nd century

In a letter to the Ephesians (*c.* 112) by Ignatius, bishop of Antioch, the language is very reminiscent of Philemon, and the name of the bishop of Ephesus (*c.* 107–117) was Onesimus. It has been suggested that the slave was released to help Paul, that in his later years he might have become bishop of Ephesus, and that his "ministry" or "service" was the collection of the Pauline corpus. This is based not simply on the identity of name, but on similarities to Philemon found in Ignatius' letter to the Ephesians, as well as two possible plays on words in chapter 2, verse 2 (*cf.* Philemon, verse 20), and chapter 4, verse 2 (*cf.* Philemon 11), relating to the bishop and unity of the church. Such a prominent position and role for one of Paul's followers might shed further light on why Philemon, apparently a very personal plea, became a part of the canon and Pauline corpus. Even if this suggestion cannot be proved, Philemon still shows Paul in his apostolic ministry, furthering the message of Christ and seeing beyond the limitations of the social order of his day, in which both slaves and freemen are servants of God.

THE LETTER TO THE HEBREWS

The question of authorship and peculiarities of style and content

The writing called the Letter to the Hebrews, which was known and accepted in the Eastern church by the 2nd century, was included also by the Western church as the 14th Pauline epistle when the canon of East and West was assimilated and fixed in 367. Hebrews has no salutation giving the name of either the writer or the addressees, although it does have a doxology and greeting at the end, which suggest that at some point the writing was sent as a letter to a community known to the author. There are also numerous admonitions in the text that appear to be directed to a definite circle of addressees and some admonitions to the church at large. In chapter 6, verses 4–8, is a severe warning against the sin of apostasy, for which there is no second repentance. Even so, Hebrews is essentially more a theological treatise than a letter. It is homiletical in style and calls itself a *paraklēsis,* which has many meanings: consolation, exhortation, sermon, advocacy, and even intercession.

The thoughts, metaphors, and ideas of Hebrews are distinct from the rest of the New Testament, with closest affinities to Stephen's speech in Acts, chapter 7. It attempts to prove the superiority and ultimacy of the revelation in Christ and the perfection of his offering of himself once and for all supersedes and makes obsolete any other revelation. Hebrews gives strength to its readers through the example of Christ and the hope and promise of free access to God and to eternal rest, an access in which Christ is High Priest and mediator forever. Such promise, on the basis of Christological developments and new covenant hopes, enables endurance in persecution, but its vocabulary is that of the sacrificial language of the Old Testament. Another theme is a typological analogy with the wilderness wanderings of Israel in which, despite their murmurings of unbelief and the hardening of their hearts in their trials, they persevered. Thus, the church, as the pilgrim people of God, travels toward the future place of Sabbath rest with Christ as their pioneer and perfector of faith.

A "word of consolation" is needed to strengthen faith in time of trouble. Actual persecution leading to martyrdom is seen as not yet come, but the church is sharply warned against apostasy, the sin of all sins. Hope during persecution and trial is expressed in the image of Christ as the perfect everlasting high priest, one of whose functions is to stand as intercessor and protector.

Suggested authors of Hebrews

Hebrews was considered a Pauline letter in the early Eastern church. Clement of Alexandria, a theologian of the late 2nd and early 3rd centuries, held that Paul had written it in Hebrew for the Hebrews and that Luke had translated it into Greek. Origen, Clement's successor as leader in the catechetical school at Alexandria, commented that its thoughts reflected Paul but that it was written at a later time with a totally different style and phraseology, and he stated "who wrote the epistle, God knows." Paul, for example, uses the term mediator only once and in a negative sense, in Galatians, chapter 3, verse 19, but Hebrews uses it several times of Christ as mediator

of the new covenant. In the West, Tertullian, a North African theologian of the late 2nd and early 3rd centuries, suggested Barnabas as the author, because Hebrews, called a "word of consolation," might have been written by Barnabas, whose name is translated by Luke as "son of consolation" in Acts, chapter 4, verse 36. After Hebrews' acceptance into the canon in the mid-4th century, it was considered Pauline, but doubts persisted; and because of basically different content and style in contradiction to Paul, various authors have been suggested for Hebrews—*e.g.,* Apollos (a Jewish Christian Alexandrian), or a follower of Stephen and the Hellenists, who had come into conflict with those not sharing his universalistic ideas. Hebrews, however, remains anonymous. The title "To the Hebrews" is secondary and may reflect either an idea as to its addressees or that it was influenced by its extensive Old Testament material.

According to internal evidence, Hebrews was written in a second or later generation of Christians. Persecution references suggest a time after Nero's persecution and about the time of the emperor Domitian but early enough to be quoted or alluded to in the First Letter of Clement (*c.* 96), thus suggesting a date of *c.* 80–90.

The place of the addressees may be Italy, because 13:24 is understood as a greeting sent home from one writing from abroad, but this is not certain. The addressees were probably Gentile Christians who needed instruction in "the elementary doctrines of Christ" and concerning faith in God.

Allegorical or typological interpretive techniques

Hebrews constitutes the first Christian example of a thoroughly allegorical, typological exegesis (critical interpretation) of the Old Testament. There were precursors of such a methodology in Jewish Alexandrian biblical exegesis (*e.g.,* Philo), and Platonic tendencies found in Hebrews can also be found in Jewish-Alexandrian methods of interpretation of the Old Testament. The language of Hebrews is extremely polished, elegant, and cultured Greek, the best in the New Testament. Linguistically and stylistically, it shows only a slight influence of the Koine (common Greek). The Attic style is broken only in passages in which Hebrews quotes the Septuagint. Plays on words and synonyms with similar beginnings for emphasis show the author's literary craftsmanship.

There are more Old Testament citations in Hebrews than in any other New Testament book. They are drawn mainly from the Pentateuch and some psalms.

The church is viewed as being in danger of discouragement in the face of persecution and possible apostasy. If faithless, church members risk total loss, for no second repentance is possible. Through his special Christology, the author seeks to help the readers by showing that Christ is the saviour superior to any other and that as Saviour, Son of God, High Priest, pioneer, guide, and forerunner, he who has already suffered and been glorified will lead the wandering people of God to their eternal Sabbath rest, an eschatological future state of peace and renewal.

Christological and eschatological motifs

This high type of Christology is combined with much stress on Jesus' humanity. He partook of man's nature and overcame death to destroy the power of the devil in order to deliver man. Thus, having been made like his brethren he has become a faithful High Priest to make expiation for the sins of the people. Because he himself suffered and was tested, he can help those who are tested and tempted. Through suffering, tears, and obedience Jesus was made perfect and thus the source of help and salvation, being designated by God a High Priest after the order of Melchizedek, king of Salem and priest of God Most High in Abraham's time.

Christ and his once for all (*ephapax*) sacrifice has superseded and made all Old Testament sacrifices and cultic practices obsolete. Christ is superior to the prophets because he is a son, superior to the angels because they worship him, and (in the light of his cosmic role as apostle and High Priest) superior to Moses, who brought God's Law to Israel, because Moses was a servant in God's house and Christ a son. Christ is also superior to Moses' successor Joshua, because Joshua did not bring the wandering people into a perfect rest; superior to the Old Testament priesthood of Aaron, because Christ, the true High Priest, has sacrificed himself once for all and is without sin; and superior to the patriarch Abraham, because Abraham paid tithes to the priest of Salem, Melchizedek, who as the prototype of Christ had no human antecedents. Christ, High Priest forever by obedient suffering and perfection in that he lives up to the demand, has become the source of salvation. He is High Priest in the heavenly tabernacle and mediator for the new covenant. On the basis of this Christology and ecclesiology, the rest of Hebrews is composed of injunctions to faithful life in all situations, spiritual or temporal. In chapter 11, verse 1, Hebrews gives a programmatic statement that should be translated: "Faith is the Reality [rather than "assurance," as in the usual translation] of what is hoped for and the Proof concerning what is invisible." In Hebrews, Jesus is that Reality and that Proof, and everything else is unreal or at best an earthly copy or a shadow. The heroes and martyrs of old were looking toward his coming (chapter 11) and those now under persecution look toward him and find strength (chapter 12) as they leave the ultimately unreal structures of this world, seeking the "coming city" and going out to him who was executed outside the walls of the city made with hands. Thus, the message of Hebrews is: Reality versus sham and shadow, Christ's sacrifice (priest and victim in one) versus the cult of temples, and the real heavenly rest and heavenly city versus the sabbath and Jerusalem.

THE CATHOLIC LETTERS

As the history of the New Testament canon shows, the seven so-called Catholic Letters (*i.e.,* James, I and II Peter, I, II, and III John, and Jude) were among the last of the literature to be settled on before the agreement of East and West in 367. During the 2nd and 3rd centuries, only I John and I Peter were universally recognized and, even after acceptance of all seven, their varying positions in Greek manuscripts and early versions revealed some conflict concerning their inclusion. The designation Catholic Letters was already known and used by the church historian Eusebius in the 4th century for a group of seven letters, among which he especially mentions James and Jude. The word catholic meant general—*i.e.,* addressed to the whole, universal church as distinguished, for example, from Pauline letters addressed to particular communities or individuals. The earliest known occurrence of the adjective "catholic" referring to a letter is in the account of an anti-Montanist, Apollonius (*c.* 197) in his rebuke of a Montanist writer who "dared, in imitation of the Apostle [probably John] to compose a catholic epistle" for general instruction. In the time of Origen (*c.* 230), the term catholic was also applied to the *Letter of Barnabas* as well as to I John, I Peter, and Jude.

The meaning of "Catholic Epistles"

In the West, however, "catholic" took on the meaning in Christian usage as implying a value judgment as to orthodoxy or general acceptance. Thus, the West used it for all the New Testament letters that were in the canon along with the four gospels and Acts. All letters considered authoritative and of equal standing with those of Paul were therefore termed canonical in the West. Not until the Middle Ages did both East and West designate the seven as "catholic epistles" in the sense of being addressed to the whole Christian Church, in order to distinguish them from letters with more particular addresses. Had not the main tradition placed Hebrews in the Pauline corpus, it would perhaps rather have been counted among the Catholic Letters. Hebrews, however, looked "Pauline" rather than "Catholic" in that it presented an extensive theological argument to which the parenesis (advice or counsel) was applied at the end.

These seven letters are grouped together despite their disparate authorship and dates because of a number of characteristics common to all of them. Though the three Johannine letters, and especially I John, are distinctly Johannine in character, the four other Catholic Letters are of special interest precisely because they lack strong personal or peculiar traits both in their theological and in their ethical statements. This characteristic makes them a good source for understanding the piety and life-style of the majority of early Christians. These letters differ from the Pauline letters in that they seem to have been written for

general circulation throughout the church, rather than for specific congregations. Though Paul wrote as a missionary responsible for his recent Gentile converts, these letters address established congregations in more general terms. It is interesting to note, for example, that in I Pet. 2:12 the word Gentiles refers to "non-Christians" without any awareness of its older and Pauline meaning of "non-Jews."

The purpose of the Catholic Letters

The purpose of the Catholic Letters is to meet ordinary problems encountered by the whole church: refuting false docrines, strengthening the ethical implications of the Gospel message, sharing in the common catechetical and moral materials, and giving encouragement in the face of the delay of the Parousia and strength in the face of possible martyrdom under Roman persecution. They guide the ordinary Christian in his day-to-day life in the church.

The Catholic Letters preserve a considerable common legacy of ethical themes and quotations. Such themes and quotations (from the Old Testament) were handed down traditionally, though the writers interpreted them independently for their situations. For example, Proverbs, chapter 3, verse 34, showing God's scorn to scorners and favour to the humble, is used in James, chapter 4, verse 6, as a warning against involvement in the world and an exhortation to submission and humility, but in I Peter, chapter 5, verse 5, it exhorts Christians to humility and submission in relation to one another in the church and brotherhood. Because the Catholic Letters represent a common pool of Christian teaching, there are overlapping points, but these come from shared tradition rather than literary dependency. The virtues extolled in the early church are not particularly Christian but often coincide with those cultivated in Hellenistic culture, sometimes with a Jewish Hellenistic emphasis. An act of mercy and virtue valued in both Jewish and Hellenistic tradition is epitomized in hospitality (*e.g.,* I Peter 4:9). Similarly, Hellenistic lists of virtues and vices occur as needed from the general body of early Gentile Hellenistic tradition applied to the Christian communities. In these epistles, theological and credal statements are woven in and used for immediate ethical application. Thus, they differ from the Pauline style of extensive theological sections coupled with ethical applications that follow at the end of the epistle.

In the Catholic Letters, to be a Christian was to be in opposition to the world, a member of a minority church and thus at any time liable to be called as witness to the faith and perhaps to suffer and die for it. Eschatological trials are coming (*e.g.,* I Pet. 1:6f., 4:12–19; II Pet. 3:2–10; I John 2:18 ff., 4:1–4; Jude 17 ff.), and the Christian views false prophecy and heresy as well as hostile encounter with the world as part of the trials. The theme of joy in persecution, suffering, and the final trial or ultimate "testing" is based on Christ's victory over these events and the sense of being a member of his community. Thus, the Christian should show submission, nonretaliation, humility and patience, good conduct, and obedience to authorities, because his witness must be blameless when his faith is tested in the world, in the courtroom, and in martyrdom.

The Letter of James. The Letter of James, though often criticized as having nothing specifically Christian in its content apart from its use of the phrase the "Lord Jesus Christ" and its salutation to a general audience depicted as the twelve tribes in the dispersion (the Diaspora), is actually a letter most representative of early Christian piety. It depicts the teachings of the early church not in a missionary vein but to a church living dispersed in the world knowing the essentials of the faith but needing instruction in everyday ethical and communal matters with traditional critiques on wealth and status. In matters of church discipline and the practice of healing, there is stress on prayer, anointing, and confession of sin in order that the healing of the sick may be effected. Steadfastness, even joy, in persecution is based on pure religion with strong ethical demands, as noted in chapter 1, verses 2–4 and 19–27.

A debate as to how James' statement that "faith apart from works is dead" compares with Paul's "justification by faith without works" in Romans has a long history. The debate, central to the history of Christianity, has usually overlooked the simple fact that Paul speaks about "works of the Law" and does so with reference to those "works" that divide Jews and Gentiles—*e.g.,* circumcision and food laws. James, on the other hand, refers to works of mercy. Thus, the two statements are not only reconcilable but address themselves to quite distinct and different issues. Even Paul referred to mutual support of the brethren by the glorious phrase "the law of Christ" (Gal. 6:2) and this is the same as James' "royal law" (James 2:8). The Pauline language presumably was not in James' mind. In James, chapter 2, the example of Abraham's faith is used to show justification by works. It is to be noted that Paul also used Abraham as the paradigm of righteousness to demonstrate justification by faith in Romans, chapter 4, again showing the difference in purpose and setting of the two epistles.

In view of the post-apostolic situation depicted, James, the son of Zebedee, who died as a martyr before AD 44, could not have been the author. From the content, neither could James, a brother of the Lord and the leader of the Jerusalem church; his martyrdom is reported as *c.* AD 62. Thus, James is pseudepigraphical, with the purpose of gaining apostolic authority for its needed message. The date of writing is probably at the turn of the 1st century, and its addressees are the whole church.

James' stress on ethical imperatives

Of James' 108 verses, 54 contain imperatives—an obvious proof that advice is stressed. Such admonitions are expressed in the form of general ethical wisdom sayings, Hellenistic Jewish lists of virtues and vices, and Christian as well as pagan aphorisms sometimes related to popular preaching of the Stoic Cynic style.

In chapter 5 the community is enjoined to patience, steadfastness, and good behaviour. The Old Testament prophets, who spoke in the name of the Lord, are used as examples of suffering and endurance as they awaited the Judge. Thus, reference to the Parousia of Christ may have been conflated by the Christian writer to the coming of the Lord in judgment, an interpretation with "the day of the Lord" in mind. "Behold, the Judge is standing at the doors" is accompanied by the admonition, "You also be patient. Establish your hearts, for the coming of the Lord is at hand," (chapter 5, verses 8 and 9).

The purpose of I Peter

The First Letter of Peter. The purpose of the First Letter of Peter is exhortation directed to "the exiles of the Dispersion" in Asia Minor in order that they "stand fast" in God's grace in the face of persecution. On the one hand, such persecution is viewed as part of the trials of the end-time that the community must undergo before the coming of the new age. On the other, persecution is viewed as a simple fact of Christian community life in the world. In imitation of Christ, tribulations and testing can be a basis for joy.

In the address, the author calls himself "Peter, an apostle of Jesus Christ," and in chapter 5, verse 1, a "fellow-elder and witness of the suffering of Christ." Any Christian, not just a fellow eyewitness, however, might be such a witness and hope to partake in the future "glory that is to be revealed." The writer or the redactor of I Peter used Pauline and gospel theology and terminology both in quotations and in allusions and, if literary dependency cannot always be demonstrated, there is dependence on the catechetical traditions known in the post-apostolic church.

The milieu of the letter seems to reflect the time and temper of the correspondence of the emperor Trajan with Pliny the Younger, governor of Bithynia (*c.* 117). Pliny requested clarification as to the punishment of Christians "for the name itself" or for crimes supposedly associated with being a Christian. I Peter, chapter 4, verse 15, appears to reflect this situation: that a Christian be blameless of all crime and, if punished, be persecuted only "as a Christian." Pliny continued that denounced Christians are executed if they persevere in their belief but that whatever their creed "contumacy and inflexible obstinacy deserved punishment"; Trajan's response was that those denounced as Christians be punished. The warning in I Peter, chapter 3, on a Christian's manner of defense and submissiveness to authorities points to a date in the first quarter of the 2nd century. Such a date does not preclude reflection on earlier persecutions, such as those under Domitian.

The Greek style is hardly in keeping with a Galilean Pe-

ter—described as illiterate or uneducated in Acts, chapter 4, verse 13. The Greek is fluid, and the Old Testament citations are from the Septuagint. The addressees appear to be Gentile Christians portrayed as the new Israel dispersed among the (heathen) Gentiles, based on the analogy of the old Israel, a diaspora among the nations.

The work is thus pseudonymous, attributed to Peter through Silvanus, whose name constitutes a part of the pseudepigraphic device that strengthens the authority of the epistle. I Peter is an excellent example of the testament form modelled on the traditions of an Apostle and the message of his martyrdom. Peter, whose death and traditions concerning him were known to the readers of the time of I Peter, gives weight and authority to the letter that is formed in many ways as a farewell and admonition to those who follow, in order that they may stand firm.

The letter in the form of a testament

Warnings are given from the Apostle's own example along with counter-virtues for vices. Such testament forms have a mixture of wisdom material, advice, exhortation, hymns for ethical admonition, and apocalyptic elements with accounts of trials to come. This mixture is found in strange arrangements, but is perhaps solved if read as a testament form. Peter had denied that Christ must suffer and in I Peter suffering is the way of discipleship and even of joy. In Luke, chapter 22, Peter's denial was prophesied, and Jesus interceded for him in order that he might repent and strengthen his brethren (*cf.* I Peter, chapter 5, verses 10 and 12). In Mark and Matthew the defection of the Apostles was foretold in terms of the scattering of the sheep when the shepherd was stricken, and Peter does deny his Lord. In John, chapter 21, the risen Lord paralleled Peter's threefold denial with a threefold question as to Peter's love. At each affirmation the Lord responds with the forgiving command to feed the sheep—to care for the community. This is a central motif in I Peter. Immediately following the charge to Peter in John is the prediction of his own martyr death, and in I Peter the church is urgently admonished to accept trials as nothing strange, because they are a sharing in the sufferings of Christ. In the Garden of Gethsemane, Peter in particular was rebuked because he did not watch, and in I Peter the church is admonished to watch and be vigilant against the Devil. Prayer against temptation is also stressed.

In the Matthean account, Peter is delegated to build the church, and in I Peter it is the chief Apostle (Peter) who points to Christ as Shepherd and Bishop, who through his suffering collected the wandering sheep to himself. In like manner—on the model of Christ or perhaps Peter—the elders are exhorted to feed their flocks humbly and faithfully. Thus, there is a typical testament form: Peter has failed and repented; and the church is warned, admonished, and strengthened as by the Apostle, who, on the analogy of Jesus' Passion and death in innocence, exhorts the church to share in the vocation of innocent suffering and to do good in innocence. Finally, I Peter, viewed as a "testament," is in itself an apocalyptic "witness," and with its admixture of advice, example, and general address to the faithful living in the Diaspora as sojourners, with the authority of its martyred "author," it constitutes authority and strength for the church that faces the persecution of the world. References in chapter 5 to Rome (called Babylon) and to Mark are then also part of the pseudepigraphic testament form, as they presuppose the common tradition of Peter's martyrdom in Rome and his connection with Mark.

Christological elements

There are three Christological hymnic fragments in I Peter: 1:18–21, ransom by Christ; 2:21–25, with reference to the Book of Isaiah, chapter 53, used as ethical admonition; and 3:18–20, Christ's descent into hell. The last is in the context of Christ's going and preaching to the spirits in prison (a reference to the apocryphal *First Book of Enoch* with Satan chained under the earth but his descendants at work in the world until the end-time) in order to show that Christ, through his descent, has overcome the powers that underlie and engender persecution of the Christians. This is reaffirmed in chapter 5 by encouraging Christians in their fight against the Devil, for, though suffering will be a part of this resistance, there will be victory at the end. Imitation of Christ is a basis for joy even in suffering. The end is viewed as near, and final salvation can thus be anticipated.

The Second Letter of Peter. The Second Letter of Peter was written as a letter to the whole church purporting to be similar in testament form to that of I Peter. It deals with the problems of the delay of the Parousia and accounts for it in terms of God's time being different from that of man and God's patience in waiting for all men to be better ethically. This letter, the latest of the New Testament, shows how Christendom dealt with the delay of the Parousia, discarded older Jewish apocalyptic ideas by substituting those with Hellenistic emphases, and is clearly in its content and exposition a methodically worked out artistic product, fictionalizing the older beliefs, in order to bring them into some agreement with traditional Christian terminology.

Problems dealt with in II Peter

II Peter names Simon Peter as its author and declares his position by setting down rules for true faith as he sees it. His work is different in meaning and interpretation from the earlier tradition and understanding of the church. He regards the Transfiguration of Jesus on the mountain as the first Parousia and urges patient waiting for the final coming of the Lord. Although he refers to his letter as a second letter of Peter, his Hellenistic concepts and rhetoric could hardly be attributed even to the author of I Peter. II Peter speaks of "partakers of divine nature," a term from the mystery religions, and mixes proverbs with familiar quotations from Hellenistic tradition. Thus, not only is this letter pseudepigraphic, but it is an even later fiction, probably nearer to AD 150 than the end of the 1st century.

Almost all of Jude is used in II Peter, but II Peter drops out a quotation from *I Enoch* in Jude 14 ff., possibly demonstrating some fear of using apocryphal writings. Heresies are attacked by criticism of their interpretation of scripture and misuse of set tradition, another evidence of the late date of II Peter. Reference is made to "all the epistles of Paul which contain things hard to understand" and to "other scriptures," evidence of a New Testament canon well on its way to being delineated over against the Old Testament. Though skillfully composed, II Peter cannot hide the Gnosticism included in its view and much misinterpretation of the traditional body of faith of the early church. Thus, II Peter is an example of the church at a relatively late period, de-eschatologized for the most part and brought near to early institutionalized religion with a ministry but depending on ideas and a theology so changed that it is almost unrecognizable.

Attacks on heresies and eschatological views

The eschatology of II Peter awaits a new heaven and a new earth after the dissolution by fire of the old, evil earth with its unrepentant people. The Parousia no longer is Christological in nature but anthropologically oriented, with a vindication of the good and a punishment of the wicked. II Peter presents a picture of the church at the latest point in the canon and illustrates the necessity to reevaluate and recall more normative Christian traditions.

The Johannine Letters: I, II, and III John. The three epistles gathered under the name of John were written to guide and strengthen the post-apostolic church as it faced both attacks from heresies and an ever increasing need for community solidarity—along with the concomitant love and ethics necessary to such unity.

The purpose of the Johannine Letters

I John, though lacking any formal epistolary salutation or ending, directs itself to a circle of readers with whom the writer is acquainted. Taking the form of an anonymous "homily" for admonition against heresy and instruction in faith and love, it was directed to a wide audience or was to be circulated beyond a particular congregation. II and III John are brief letters from an author described only as "the elder," implying a position of some authority. II John, chapter 1, is addressed to an "elect lady and her children," probably a designation of a church with difficulties similar to those found in I John. III John is the most personal, being addressed by the elder "to the beloved Gaius," who has been praised particularly for his hospitality (probably to missionaries) and his brotherly love. The presbyter (elder), probably the author of II and III John, apparently was a man who was authoritative enough to influence and direct mission activities. All three letters, despite their differences of address, appear to have

been accepted among the Catholic Letters as having been circulated for the church at large.

I, II, and III John share much common terminology, style, and general situation. They are all called Johannine because they are loosely related to the Gospel According to John in style and terminology and could be the outcome of its theology.

The early church attributed I, II, and III John to John, the Apostle, the son of Zebedee. Although II and III John may possibly have been written by the same presbyter, this "elder" is not necessarily the author of I John, although it is commonly accepted that the three Johannine letters came from a "Johannine" inner circle. The earliest reference to the Johannine letters is in the *Letter to the Philippians* by Polycarp of Smyrna (7:1). Papias, who was a 2nd-century bishop of Hierapolis, mentions I John and quotes it several times, but he distinguishes between John, the Apostle, and John, the presbyter. Polycarp, Papias, and internal evidence point to the region of Asia Minor as the probable sources of the Johannine literature. These references and the organization of the churches indicated in the letters, as well as the lack of signs of persecution, suggest a date for the letters at around the beginning of the 2nd century.

Apologetics against Docetic Gnosticism

The First Letter of John. I John assumes a knowledge of the Johannine Gospel (the author of I John may be the ecclesiastical redactor of the Gospel According to John) and adds ethical admonition and instruction regarding the well-being of the church as it confronts heresy and stresses the lack of moral concern that springs from it. There is strong defense against the threat of a type of Gnosticism called Docetism that denied the reality of Jesus' earthly life and thus the meaning of the cross. Possessing special spiritual knowledge, the Docetic Gnostics had no need of the earthly Jesus and the humanity of Christ. This Docetic heresy led them to reject the Lord's Supper, but not Baptism. Their special possession of the Spirit had led them erroneously to consider themselves sinless and to deny the fellowship that has the cleansing of sins. Because the heresy may have led to libertinism, the ethics of Christians must accord with their faith and find expression in the love of the brethren in the church. "He who hears my word and . . . believes has passed from death to life" (John 5:24) is continued in I John 3:14, "We have passed out of death into life, because we love the brethren." The Gnostics separated themselves from the church in schism and have thereby committed the "sin unto death." They are false prophets and deceivers described by the term Antichrist. The true Christians, the "children of God," hold the true faith evidenced by their loyalty to the church and their charity toward its members.

The emphasis on love in I John

A constant theme in I John is that of God's love, which makes Christians the children of God. As children of God they keep the new commandment of love, which is of light—that of brotherly love—and resist the world, evil, and false teaching. Because Christ gave his life for man, the Christian's response is also to be self-giving. Through obedience and faith, God forgives even when man's heart condemns him, "for God is greater than his heart." It is of interest to note that in I John 2:1–2, Jesus is referred to as paraclete (advocate), but in the Gospel According to John, such references are to the Spirit. John 14:16, however, refers to "another Counselor." This discrepancy can be resolved by interpreting Jesus with his disciples as their advocate with another to come (the Spirit), and, in I John 2:1–2, the risen Lord becomes the advocate for the expiation of all sin. Righteousness and faith are emphasized in chapters 4–5, and again these characteristics are those of the children of God, who will finally in the end-time be like him who gave the promise, the commandment, and the joy of love.

The Second Letter of John. II John warns a specific church (or perhaps churches), designated as "the elect lady and her children," against the influence of the Docetic heresy combatted in I John, whose proponents lured Christians from "following the truth, just as we have been commanded by the Father." In II John, as in the Gospel According to John and I John, the light–darkness images are similar to those of the Dead Sea Scrolls. To "walk in the truth" in II John is to reject heresy and follow the doctrine of Christ.

The Third Letter of John. III John, addressed to Gaius, shows that the writer is concerned about and has responsibility as presbyter for the missionaries of the church. It is somewhat of a short note concerned with church discipline, encouraging hospitality to true missionaries, and thus not unconnected with true doctrine and the command of love.

The purpose and authorship of Jude

The Letter of Jude. The Letter of Jude, after a salutation that attributes it to Jude, the brother of James, and addresses itself to the church as a whole, develops the theme of the short letter—a polemic against heretics who have abandoned the transmitted traditional faith and who will thus be judged by the Lord. They deny Christ, and punishment similar to that of Sodom and Gomorrah in the Old Testament for such a denial is threatened. Heretical beliefs led to various sins and libertinism, and the judgment that will come upon them is cited from *Enoch* 1:9, demonstrating that this short letter reflects the post-biblical Jewish apocalyptic train of thought in the early Christian era.

"Jude, a servant of Jesus Christ and brother of James" is probably meant pseudepigraphically to relate this Jude to James the brother of the Lord so that this Jude is also a brother of the Lord. This, however, is impossible because the letter reflects a later time. Verse 17 refers to "the predictions of the apostles of our Lord Jesus Christ" concerning mockers and sinners. Thus, the author is recalling a former time that was prophesied regarding the heresies and trials of the end-time. Such a bearer of apostolic tradition is violently attacking heresy in the interest of transmitted traditional faith. Again, it would appear that the letter is pseudepigraphic and may have originated in Syria or Asia Minor.

The author struggles forcefully against heretics who deny God and Christ and attempts to strengthen his readers in their fight against such heresy that leads to wickedness and disorder. Libertinism is a characteristic of such heresy and the punishment of the heretics will be similar to that which befell the unfaithful in the Old Testament patriarchal times. Only steadfastness in faith, true doctrine, and prayer can lead to mercy, forgiveness, restoration, and final salvation. An attempt to bring the erring to repentance may save them. The letter concludes with a typical doxology.

The form is less a catholic letter than a declared position that lays down general rules. The date is probably near the end of the 1st century and before II Peter, which draws upon it.

THE REVELATION TO JOHN

The Revelation (*i.e.,* Apocalypse) to John is an answer in apocalyptic terms to the needs of the church in time of persecution, as it awaits the end-time expected in the near future. The purpose of the book is to encourage and admonish the church to be steadfast and endure. The form of an apocalypse shows affinities with contemporary Jewish, Oriental, and Hellenistic writings in which problems of the end of the world and of history are linked both with prophecy of an eschatological nature and with "sealed" secret mysteries. Such revelations are traditionally received in trances, characterized by strange symbols, numbers, images, and parables or allegories that represent people and historical situations. Apocalypticism is essentially dualistic, presenting the present eon as evil and the future as good, with an ultimate battle between the divine and the demonic to be won only after one or more cosmic catastrophes. The aim of apocalyptic literature is to depict in the age of present tribulation a knowledge of a future glorious victory and vindication, thus giving hope and assurance.

Apocalyptic themes

In Revelation it is God who gives the revelation to Jesus Christ to be shown by Christ through an angel to his servant John, in exile on the island of Patmos, in order that John become his seer and prophet to the church. John is to write down what he has seen, what is, and what is to come. In contradistinction to most Jewish apocalyptic works, Revelation is not pseudonymous and John is to

give finally unsealed, clear prophecy related to the present and to the end-time.

As in the rest of the New Testament, the starting point of eschatological hope is the saving act of God in Jesus, a historical centre pointing toward historical developments that will bring about the establishment of God's kingdom and vindication of his people, ransomed by the blood of Christ, the Lamb who was slain. It provides certainty and encouragement with the example of the faithfulness of those who have already witnessed unto death (martyrs) and their reward—special inheritance in the eternal kingdom.

The content of Revelation

After the introduction, Revelation continues first as a series of seven letters to seven churches in the province of Asia, thence to the whole church with an epistolary introduction and, after the apocalypse proper, an epistolary blessing as the last verse. The letters sent from the heavenly Christ through John (chapters 2 and 3) exhort, comfort, or censure the churches according to their condition under persecution or danger of heresy. From chapters 4–22 there are series of visions in three main cycles, each recapitulating but expanding the former in greater and clearer detail with groups of seven symbols predominating in each (seals, chapters 6–7; trumpets, chapters 8–10; and bowls, chapters 15–16). This material is interspersed with visions of God in His heavenly council, various visions of catastrophe and of Satan, the destroyer, the appearance of two witnesses and other martyr examples to spur the church to endurance, the victory of the archangel Michael over the dragon (Satan) by the blood of the Lamb (Christ), and the representation of the powers of emperor cult and false prophecy as beasts who bring destruction to the unfaithful in God's judgment. A heavenly woman who bears a messianic son is threatened by a dragon. Her child is carried up to heaven by God, and she escapes by hiding in a place prepared for her by God. The beasts who appear persecute the Christians and the "number" signifying the second beast is that of a man, "666" (or, in a variant reading, "616") probably indicating the emperor Nero. God's triumph in history is depicted in his judgment on the harlot Babylon (Rome), and the final consummation portrays the victory of Christ over the Antichrist and his followers. In chapter 20 the thousand-year reign of Christ with those who witnessed unto death is depicted. Satan, again loosed, is vanquished by fire from heaven with the beasts (empire power and false prophet), and the last judgment leads to a new heaven and a new earth, the new Jerusalem. This writing is, thus, a prophetic-apocalyptic work.

In summary, the seer reminds the reader that the words, because they are of God, are trustworthy and true. The motif that the Lord is coming soon is again repeated. This reflection of the early Christian watchword suggests a sacred liturgical style. The last verse is the closing benediction—perhaps not only of the letters in the beginning of Revelation but of the whole of Revelation, which was to be read aloud in a worship setting.

Authorship and style

After AD 70 (the fall of Jerusalem), apocalypticism was introduced into Asia Minor and *c.* 80–90 a prophetic circle was formed near Ephesus. Its leader was John, a prophet, who might well have been the author of Revelation, which is deeply steeped in apocalyptic traditions. The "Johannine circle" bearing the tradition of John, the Apostle of the Lord, and from which emerged the Gospel and letters bearing his name, might have been a continuation of the prophetic conventicle of Ephesus in which John was prominent. The various writings do not have to be consistent except in their basic faith in Jesus Christ; and, as the situations to which they addressed themselves were different, different styles and content were required. The seer was probably involved in an actual historical situation in the late 80s under Domitian, a time when there was open conflict between the church and the Roman state. There is a tradition supported by Irenaeus, a 2nd-century bishop of Lyons, that in this persecution punishment was death or banishment. John's prominence might have led to banishment to Patmos, an isle off the coast of Asia Minor, from his homeland in or around Ephesus. From Patmos he wrote a circular letter to the churches in Asia.

Though the style of Revelation is certainly eclectic in form and content, containing elements of a heavenly epistle and with more than three-fourths of the rest made up of prophetic-apocalyptic forms from varied sources, it reflects a systematic and careful plan. Even the apocalyptic, however, is "anti-apocalyptic" in that the seer's message is open and the mysteries serve not to conceal but to heighten what is seen and to be expected. Apocalyptic schemata and motifs are, however, used toward this purpose, and allegorical incorporation of sources is more a demonstration of the true, ultimate message than a literary device. Blurred images (*e.g.*, God, Christ, and angels; chiliastic [1,000-year] eras and temporal duplications; as well as interpretations) are part of the apocalyptic style, but a current concrete historical situation is the foundation. Revelation is written in fantastic imagery, blending Jewish apocalyptic, Babylonian mythology, and astrological speculation. It is pictorial, dramatic, and poetic.

Revelation contains long sections characterized by Greek that is grammatically and stylistically crude, strangely Hebraized to give a unique, almost Oriental, colour. This may have been deliberate. Although Revelation is replete with Old Testament allusions, there are no direct quotations, and this may reflect the seer's conviction that the work is a direct revelation from God. In other sections the poetry of Revelation might stem from the seer's experience in the heavenly throne room of God, from hearing the hymns of the angelic host, or from his recollection on Patmos of the liturgical practice of the church. The image of the Bride and wedding feast together with the "Come, Lord Jesus!" have associations with the eucharistic liturgy of the early church.

The recapitulation of images and the use of numbers

The recapitulations of the seven seals, trumpets, and bowls may be deliberate schematization. The purpose of such repetition and increasing revelation can be a way of heightening enthusiasm to encourage the church.

Mysterious numbers and divisions (such as 7, 3, 12) recur and are part of the theme of assurance, because God has numbers in their order as a sign of his plan of salvation, turning chaos to orderly cosmos. The mysterious name of the second beast, 666, in 13:18, can be calculated by "gematria," assigning their numerical values to letters of the word and summing them up. The most adequate solution is Nero (the numerical value of the Hebrew letters for *Caesar Neron* equals 666), a demonic Nero *redivivus* (revived), who returns from the dead as Antichrist. Astronomy and astrology have also been applied to Revelation in terms of the signs of the zodiac or a calendar of feasts and seasons as keys to understanding its structure, because it is God who orders the times and seasons.

Martyrological themes

Two witnesses described in chapter 11 have been assumed to be Elijah and Moses, Peter and Paul, or simply two examples of martyrs through whom God shows His punishment of the wicked and vindication of the righteous to his glory. There are strong martyrological themes throughout Revelation, and it seems to stand on the border line of the point at which the word witness (*martys*) became a technical term for a witness unto death, or martyr. The cosmic battle in heaven is fought by those willing to give their lives, who mix their blood with the blood of the Lamb, whose blood "ransomed men for God." The writer of Revelation based his hope for the church on perseverance, on endurance even to death, and on what the future will bring when the church will live with the glorified Christ, slain as a lamb. The harlot of Babylon will be destroyed and the church will endure; Babylon falls and the new Jerusalem, the city of God that is to come, is depicted in all its glory. These are the hopes to strengthen the persecuted church, assurance that God will soon triumph. With trumpet call and heavenly voices there is the joyful promise that "The kingdom of the world has become the kingdom of our Lord and of his Christ, and he shall reign for ever and ever." (K.St./E.T.Sa.)

New Testament Apocrypha

NATURE AND SIGNIFICANCE

The title New Testament Apocrypha may suggest that the books thus classified have or had a status comparable to that of the Old Testament Apocrypha and have been recognized as canonical. In a few instances such has been the

case, but generally these books were accepted only by individual Christian writers or by minority heretical groups. The word apocryphal (secret) is applied to Gnostic traditions and writings both by Gnostics and by their critics; from the 2nd century, for example, comes the *Apocryphon* (secret book) *of John.* In the 4th century the word referred to books not publicly read in churches. It meant apocryphal in the modern sense (*i.e.,* fictitious) only by implication, as when the church historian Eusebius speaks of some of "the so-called secret books" as forgeries composed by heretics.

Pseudepigraphical gospels, acts, letters and apocalypses

Like the New Testament books themselves, the New Testament apocryphal books consist of gospels, acts, letters, and apocalypses. The apocryphal writings, however, are almost exclusively pseudepigraphical—*i.e.,* written in the name of the apostles or disciples or concerning individual apostles. In general, they were created after and in imitation of the New Testament books but before the time when a relatively restricted canon, or list, of approved books was being formulated. They arose chiefly during the 2nd century, when the lines between orthodoxy and heresy were not absolutely fixed and when popular piety seems to have been rather freely expressed. What these works tell about Jesus and his disciples resembles the imaginative Midrashic (didactic commentarial) retelling of Old Testament stories among Jewish teachers.

As the New Testament canon was gradually given definite shape, these apocryphal books came to be excluded, first from public reading in churches, then from private reading as well. With the development of creeds and of systematic theologies based on the nascent canon, the apocryphal books were neglected and suppressed. Most of them have survived only in fragments, although a few have been found in Greek and Coptic papyri from Egypt. They are valuable to the historian primarily because of the light they cast on popular semi-orthodox beliefs and on Gnostic revisions of Christianity; occasionally, they may contain fairly early traditions about Jesus and his disciples. In the 3rd century, Neoplatonists (followers of the philosopher Plotinus, who advocated a system of levels of reality) joined Christians in attacking such books as "spurious," "modern," and "forged."

The difficulties the New Testament apocryphal books caused at the end of the 2nd century are well illustrated in a letter by Serapion, bishop of Antioch. He stated that he accepts Peter and the other apostles "as Christ" but rejects what is falsely written in their name. When some Christians showed him the *Gospel of Peter,* he allowed them to read it, but after further investigation he discovered that its teaching about Christ was false, and he had to withdraw his permission.

In the early 4th century Eusebius himself found it difficult to create categories for the various books then in circulation or used by earlier authors. He seems to have concluded that the books could be called "acknowledged," "disputed," "spurious," and absolutely rejected. Thus, the *Acts of Paul,* the *Apocalypse of Peter,* and the *Gospel According to the Hebrews* were rather well attested, and he called them spurious but disputed. He definitely rejected books used by heretics but not by church writers: the gospels ascribed to Peter, Thomas, and Matthias, and the *Acts of Andrew,* John, and other apostles. About a century earlier, the North African theologian Tertullian had written about how a presbyter who wrote the *Acts of Paul* had been deposed.

Categories of authenticity and spuriousness

Without reference to the standards of canonicity and orthodoxy gradually being worked out by the churches of the 2nd through 4th centuries, it is evident that many of these books reflect the kinds of rather incoherent Christian thought that church leaders were trying to prune and shape from the 1st century onward. Often such works represented what was later viewed as inadequate orthodoxy because the views presented had become obsolete. All the apocrypha taken together show the variety of expression from which the canon was a critical selection.

THE NEW TESTAMENT APOCRYPHAL WRITINGS

This section will classify these documents in relation to their literary forms: gospels, acts, letters, and apocalypses.

Gospels. A few papyrus fragments come from gospels not known by name (*e.g.,* Egerton Papyrus 2, Oxyrhynchus Papyrus 840, Strasbourg Papyrus 5–6). There are also the *Gospel* produced in the 2nd century by Marcion (a "semi-Gnostic" heretic from Asia Minor), who removed what he regarded as interpolations from the Gospel According to Luke; the lost Gnostic *Gospel of Perfection;* and the *Gospel of Truth,* published in 1956 and perhaps identical with the book that Irenaeus (*c.* 185), bishop of Lyon, said was used by the followers of Valentinus, a mid-2nd-century Gnostic teacher. The *Gospel of Truth* is a mystical–homiletical treatise that is Jewish–Christian and, possibly, Gnostic in origin. In addition, there were gospels ascribed to the Twelve (Apostles) and to individual apostles, including the *Protevangelium of James,* with legends about the birth and infancy of Jesus; the lost Gnostic *Gospel of Judas* (Iscariot); the *Gospel of Peter,* with a legendary account of the resurrection; the *Gospel of Philip,* a Valentinian Gnostic treatise; the *Gospel of Thomas,* published in 1959 and containing "the secret sayings of Jesus" (Greek fragments in Oxyrhynchus papyri 1, 654, and 655); and an "infancy gospel" also ascribed to Thomas. Beyond these lie gospels ascribed to famous women, namely Eve and Mary (Magdalene), or named after the groups that used them: Ebionites (a Jewish Christian sect), Egyptians, Hebrews, and Nazarenes (an Ebionite sect).

Acts. The various acts, close in form and content to the contemporary Hellenistic romances, turned the apostolic drama into melodrama and satisfied the popular taste for stories of travel and adventure, as well as for a kind of asceticism that was generally rejected by Christian leaders: Andrew (including the *Acts of Andrew and Matthias Among the Cannibals*), Barnabas (a companion of St. Paul), Bartholomew, John (with semi-Gnostic traits), Paul (including the *Acts of Paul and Thecla,* with a Christian version of the story of Androcles and the lion), Peter—with the apostle's question to the risen Lord, "Lord, where are you going?" ("Domine, quo vadis?") and Peter's crucifixion upside down, Philip, Thaddaeus (his conversion of a king of Edessa), and Thomas (with the Gnostic "Hymn of the Pearl").

Popular stories of travel and adventure

Letters. Among the apocryphal letters are: a 2nd-century *Epistula Apostolorum* ("Epistle of the Apostles"; actually apocalyptic and antiheretical), the *Letter of Barnabas,* a lost *Letter of Paul to the Alexandrians* (said to have been forged by followers of Marcion), the late-2nd-century letter called "III Corinthians" (part of the *Acts of Paul* and composed largely out of the genuine letters of Paul), along with a letter from the Corinthians to Paul, and a Coptic version of a letter from Peter to Philip. There is also a famous forgery purporting to have been written by Jesus to Abgar, king of Edessa (noted in Eusebius, *Church History* I. 13).

Apocalypses. Other than the Revelation to John, which some early Christian writers rejected, there are apocalypses ascribed to two Jameses, the Virgin Mary, Paul, Peter, Philip, Stephen, and Thomas. Only the *Apocalypse of Peter* won any significant acceptance and is important for its vivid description of the punishment of the wicked.

In addition, it should be noted that there were apocryphal books with titles not so closely related to the New Testament. Among these are: the *Didachē,* or *Teaching of the Twelve Apostles* (and its later revisions, such as the *Didascalia Apostolorum,* or the "Teaching of the Apostles," and the *Apostolic Constitutions*), and the *Kerygma of Peter,* a favourite at Alexandria, as well as various Gnostic works, such as *The Dialogue of the Redeemer, Pistis Sophia* ("Faith-Wisdom"), and the *Sophia Jesu Christi* ("Wisdom of Jesus Christ"). From the 5th century there is even a *Testamentum Domini* ("Testament of the Lord"), an expansion of the 2nd–3rd-century Roman Church leader and theologian Hippolytus' *Apostolic Tradition.*

(R.M.G.)

Biblical literature in liturgy

BIBLICAL LITERATURE IN THE LITURGY OF JUDAISM

The liturgy of Judaism is that of the synagogue, which arose during and after the Babylonian Exile of 586–538

BCE and gradually replaced the Temple cult as the spiritual centre of Jewish life. The Hebrew biblical canon and the liturgy of the synagogue, to a great extent, grew up together.

The synagogue liturgy

Because the synagogue arose in a land separated from the Jerusalem Temple with its sacrificial emphasis and its priestly class, worship in the synagogue differed from what went before it in several respects. A local congregation worshipped together on a certain day of the week in a place set apart for that purpose, rather than primarily on special festival days and periods. The people worshipped without priest or cultic sacrifice, yet consciously as a community within a larger covenant fellowship and in response to a divine word that was written down in a holy scripture. Bible reading and interpretation, the singing of psalms, and prayers, both corporate and individual, were the staple content of the liturgy. The ancient synagogue liturgy has come down to the present in two books: the *Siddur,* or daily prayer book, and the *Mahzor,* or festival prayer book.

The biblically prescribed rhythm of days, weeks, months, and years gave order to the lives of the people. The Bible became familiar to old and young by being read aloud in the synagogue, and no part of worship was esteemed more highly than the reading of scripture. The Torah, the first five books of the Bible, is handwritten on a scroll. Viewed as the holiest object in the synagogue, it is kept in a sacred cabinet called the ark. Special prayers and ceremonies accompany its being taken out and replaced in the ark, and during the course of the year it is read in its entirety at the sabbath services. Torah portions are also read on the religious holidays.

A reading from the Prophets, called the Haftarah, follows each Torah reading. One of the five Megillot (Scrolls) is read on certain holidays: the Song of Solomon at Pesah (Passover), the Book of Ruth at Shavuot (Weeks), Lamentations of Jeremiah at Tisha be-Av (Av 9), Ecclesiastes at Sukkot (Tabernacles), and the Book of Esther at Purim (Lots). The Book of Jonah is read on the afternoon of Yom Kippur (Day of Atonement). Psalms are said or sung in every service. From the chanting of biblical texts, especially the Psalms, the music of the synagogue's cantor has developed into an incomparable art form (see also JUDAISM).

BIBLICAL LITERATURE IN THE LITURGY OF CHRISTIANITY

Eastern Orthodoxy. The first Christians were Jews, and they worshipped along with other Jews in the synagogue. The earliest Gentile converts also attended the synagogue. When Christians met outside the synagogue, they still used its liturgy, read its Bible, and preserved the main characteristics of synagogue worship. Every historic liturgy is divided into (1) a Christian revision of the sabbath service in the synagogue and (2) a celebration of Jesus' Last Supper with his disciples as a fulfillment of the Passover and a new covenant with a newly redeemed people of God. Thus, the church was never without traditional forms of worship.

The use of the Old and New Testament texts

For more than 100 years Christians had no authorized New Testament, the Old Testament being read, as had been done previously, in the worship service. By the middle of the 2nd century, however, Christian writings also were in the Sunday service. The Old Testament, the version used most generally in its Greek translation (the Septuagint), was the Bible from which the Gospel was preached. Its reading preceded that of the Christian writings, and the reading was far more extensive than it is in modern Christian churches.

As the liturgies grew longer and more elaborate, the biblical readings were reduced, and the New Testament gradually displaced the Old Testament. No Old Testament lesson remained in the Greek or Russian liturgy or in the Roman mass, though it has been reintroduced in the 20th century in most liturgies. All liturgies have at least two readings from the New Testament: one from a letter or other (non-Gospel) New Testament writing, and one from a Gospel, in that order. The Eastern liturgies all honour the Gospel with a procession called the Little Entrance. This action is accompanied by hymns and prayers that interpret the Gospel as the coming of Christ to redeem the world.

The Eastern liturgies, especially after the great theological controversies of the first four centuries, have favoured composed texts of prayers, hymns, and choral anthems that summarize the thought of many biblical passages, thus becoming short sermons or confessions of faith. The Nicene Creed (4th century) itself is one such text, in contrast with the Shema ("Hear, O Israel"—a type of creed) in Judaism, which consists of verbatim passages from Deuteronomy and Numbers.

The Divine Liturgy of the Eastern Orthodox churches contains many such composed texts, such as prayers that proclaim Orthodox theology (*e.g.,* the "Only begotten Son and Word of God" following the second antiphon). Isaiah, chapter 6, verse 3 ("Holy, holy, holy is the Lord of hosts; the whole earth is full of his glory"), used in the Jewish Kedusha (Glorification of God), generates two separate texts in the Eastern liturgy: the Trisagion (a solemn threefold acclamation to God) at the Little Entrance and the Greek original of the "Holy, holy, holy" in the eucharistic liturgy.

Psalms are sung extensively at the daily hours of prayer in the East as in the West. At the beginning of the Sunday service, entire psalms or more than one psalm are sometimes sung. More often, however, a psalm verse or two are combined with other material into a composite text of a hymn or anthem. A mosaic of selected psalm verses may be used either as a text for music or a spoken prayer. Most characteristic of all, especially in the Greek Church's tradition, however, is the freely composed and imaginative hymn text, based on a biblical incident or person, or an extended paraphrase of a passage of scripture. In addition to such biblically based psalms and other hymns, there are the famous Cherubic Hymn of the Greek and Russian liturgies and the original texts of hymns that have become well known in the Western churches—*e.g.,* "O gladsome light of the Father immortal," and "Let all mortal flesh keep silent."

Roman Catholicism. Liturgical worship in both Judaism and Christianity is an action that moves within the framework of biblical ideas and explains itself in biblical language. Preoccupied with really different views from opposite windows, Jews and Christians have often overlooked the common heritage that they share. This has likewise been true of the differences between Eastern and Western Christians.

Development of the Latin rite

At Rome, the liturgy was sung and said in Greek until the 4th century and was probably more like the liturgy of Syria at that time than that of Rome after the 16th century. The Latin rite developed many distinctive features, but what happened in Rome happened also to some extent in the East. The biblical readings at mass were reduced to two: the first reading, formally called the Epistle, was usually from an apostolic letter but sometimes from the Acts of the Apostles or even the Old Testament, and the second was a Gospel passage selected as appropriate for that particular day in the Church Year. The West, like the East, retained the Jewish week and developed a yearly cycle of Easter–Pentecost and Christmas–Epiphany celebrations with appropriate biblical selections. The development of the Church Year became so elaborate in the West, however, that the Roman calendar provided for every day in the year.

In the West as in the East, monastic and other religious communities observed the daily hours of prayer, in which there was little Bible reading as such but a great deal of corporate praying as well as the reading or singing of psalms. The Roman canonical hours were further enriched with homilies and legends from many sources, with Latin metrical hymns, and with biblical canticles, including a daily singing of the early Christian songs that are quoted in the Gospel According to Luke: the "Benedictus" ("Song of Zechariah") in chapter 1, verses 68–79, at Lauds (morning prayer), the "Magnificat" ("Song of Mary") in chapter 1, verses 46–55, at Vespers (evening prayer), and the "Nunc Dimittis" ("Song of Simeon") in chapter 2, verses 29–32, at Compline (prayer at the end of the day). The great anonymous canticle called the "Te Deum," a vast array of

biblical images ascribing praise and glory to God, is sung every day at Matins (an early morning prayer).

The mass is an abbreviation of a much longer liturgy. Many items are mere vestiges of more elaborate actions or texts. The psalms once sung at the entrance, for example, have been reduced to a traditional form of a sung text: an antiphon of one or two verses from a psalm, the first verse of the psalm, the "Glory be to the Father," and the antiphon repeated. The same has occurred in other parts of the mass. Psalms were once interspersed among the readings of scripture. The traditional gradual was a formalized text sung between the Epistle and Gospel, but in the reformed mass it becomes a responsorial psalm between the first and second readings. The short texts at the Offertory (offering of the bread and wine) and Communion are fragments in biblical language, but they are also masterpieces of the Latin genius for brevity, clarity, and order—as are the inimitable Latin collects (prayers), each basing its definite petition on an equally definite biblical revelation.

For centuries the mass was heard only in Latin and repeated the same readings on the same days every year, with the result that only a limited number of unconnected passages were heard in church. The second Vatican Council (1962–65) approved the plan of having a three-year cycle of biblical readings, providing an Old Testament lesson for every mass, a more nearly continuous reading from one of the Gospels each year, and a reading from one of the letters or other New Testament books over a period of weeks.

Protestantism. The term Protestant covers so wide a variety of theological views and religious and cultural groups and so many different ways of worshipping and using the Bible in worship that it is virtually impossible to say anything about the liturgy or the Bible's place in worship that would be true of all Protestants. Among Anglicans, what was said of the Bible in the Roman Catholic liturgy would generally apply. It would also apply to most Lutherans in the 20th century, but not to all Lutherans. On the other hand, there have been and are Protestants who claim or tacitly assume that nothing but the Bible should be used in worship. The use of the Bible in Protestant liturgy lies between these extremes.

Variations in Protestant liturgies

In the 16th century, the New Testament was appealed to as a guide for reforming the worship as well as the doctrine of the time. Because the worship reflected in the New Testament is synagogue worship, Protestant worship of the less liturgical kind became, in many respects, a return to synagogue worship. Protestants separated the two services (instructional and Eucharistic) that had been joined together in the historic liturgy of Christendom. The Protestant Sunday service is the Liturgy of the Learners, a new revision of the synagogue liturgy. It centres in the biblical word read and preached. The congregation worships in anticipation of and response to the scriptural word. Praise becomes corporate only in hymns sung by the congregation, and prayer voices human need and misery as revealed in the Bible and claims the promises heard there.

The absence of a developed liturgy generally limits the amount and variety of scripture read in the course of a year, as well as the forms of congregational participation. On the one hand, it limits worship to the resources and skill of local ministers, but, on the other hand, it also leaves a freedom to choose what is useful from any source—this has become an increasing practice in almost every Protestant church in the 20th century. Such freedom has been welcomed by many in the latter part of the 20th century—when all Protestant and Catholic liturgies seem likely to change without much advance notice (see also CHRISTIANITY). (H.G.D.)

The critical study of biblical literature: exegesis and hermeneutics

Exegesis, or critical interpretation, and hermeneutics, or the science of interpretive principles, of the Bible have been used by both Jews and Christians throughout their histories for various purposes. The most common purpose has been that of discovering the truths and values of the Old and New Testaments by means of various techniques and principles, though very often, due to the exigencies of certain historical conditions, polemical or apologetical situations anticipate the truth or value to be discovered and thus dictate the type of exegesis or hermeneutic to be used. The primary goal, however, is to arrive at biblical truths and values by an unbiassed use of exegesis and hermeneutics.

NATURE AND SIGNIFICANCE

Biblical exegesis is the actual interpretation of the sacred book, the bringing out of its meaning; hermeneutics is the study and establishment of the principles by which it is to be interpreted. Where the biblical writings are interpreted on a historical perspective, just as with philological and other ancient documents, there is little call for a special discipline of biblical hermeneutics. But it has been widely held that the factors of divine revelation and inspiration in the Bible, which, according to Jewish and Christian belief, set it apart from other literature, impose their appropriate hermeneutical principles, although there has been divergence of opinion on what these principles are. Again, because of the place that the biblical writings have occupied in synagogue and church, their exploitation for apologetical or polemical ends, their employment as a source for dogma or as a means of grace, fostering individual and community devotion, and the use of certain parts (especially the psalms) in the congregational liturgy, the science of hermeneutics has been studiously cultivated as a theological discipline. To treat the Bible like any other book (even in order to discover that it is not like any other book) has been condemned by believers as an unworthy, not to say impious, attitude.

Varying standards and presuppositions

At times the languages in which the biblical texts were originally composed have for that reason been treated as sacred languages. Hebrew may be to the philologist a Canaanite dialect, not substantially different from Phoenician, or Moabite, or other Semitic languages, but for some people even today this language is invested with an aura of sacredness. As for the language of the New Testament, in the days before its place within the general development of Hellenistic Greek was properly appreciated, it could be called a "language of the Holy Ghost," as it was by the German Lutheran theologian Richard Rothe (1799–1867). And even scholars who know very well the true character of the biblical languages are tempted at times to make the Old and New Testament vocabularies, down to the very prepositions, bear a greater weight of theological significance than sound linguistic practice permits. Where in other Greek literature the context would be allowed to determine the precise force of this or that synonym, there is a tendency to approach the New Testament with definitions ready made and to impose them on the text: to give one example, of two common Greek words meaning "new," it is sometimes laid down in advance that *kainos* denotes new in character and *neos* new in time ("young"). Often such distinctions are valid, but their validity must be established by the context; where the context discourages such precise differentiations, they must not be forced upon it.

Again, it is a truism in linguistic study that the meaning of a word depends on its usage, not on its derivation. It may be of interest to know that the Hebrew word for "burnt offering" (*'ola*) etymologically means "ascending" (*cf.* the verb *'ala,* "ascend"), and to trace the stages by which it attained its biblical meaning, but this knowledge is almost wholly irrelevant to the understanding of the word in the Old Testament ritual vocabulary, and any attempt to link it, say, with the ascension of Jesus in the New Testament, as has been done, can lead only to confusion.

Similarly there has been a tendency to place the history contained in the biblical writings on a different level from "ordinary" history. Here the increasing knowledge of the historical setting of the biblical narrative, especially in the Old Testament, has helped to remove the impression that the persons and peoples portrayed in this narrative are not quite "real"; it has integrated them with contemporary life and promoted a better understanding of what they had in

common with their neighbours and what their distinctive qualities were.

BIBLICAL CRITICISM

A prerequisite for the exegetical study of the biblical writings, and even for the establishment of hermeneutical principles, is their critical examination. Most forms of biblical criticism are relevant to many other bodies of literature.

Textual criticism. Textual criticism is concerned with the basic task of establishing, as far as possible, the original text of the documents on the basis of the available materials. For the Old Testament, until 1947, these materials consisted principally of: (1) Hebrew manuscripts dated from the 9th century AD onward, the Masoretic text, the traditional Jewish text with its vocalization and punctuation marks as recorded by the editors called Masoretes (Hebrew *masora,* "tradition") from the 6th century to the end of the 10th; (2) Hebrew manuscripts of medieval date preserving the Samaritan edition of the Pentateuch (first five books of the Bible); (3) Greek manuscripts, mainly from the 3rd and 4th centuries AD onward, preserving the text of the pre-Christian Greek version of the Hebrew Bible together with most of the apocryphal books (the Septuagint); (4) manuscripts of the Syriac (Peshitta) and Latin (Vulgate) versions, both of which were based directly on the Hebrew. Since 1947 the discovery of Hebrew biblical texts at Qumrān (then Jordan) and other places west of the Dead Sea has made it possible to trace the history of the Hebrew Bible back to the 2nd century BC and to recognize, among the manuscripts circulating in the closing generations of the Second Jewish Commonwealth (*c.* 450 BC–*c.* AD 135), at least three types of Hebrew text: (1) the ancestor of the Masoretic text, (2) the Hebrew basis of the Septuagint version, and (3) a popular text of the Pentateuch akin to the Samaritan edition. A comparative examination of these three indicates that the ancestor of the Masoretic text is in the main the most reliable; the translators of the Revised Standard Version (1952) and New English Bible (1970) have continued to use the Masoretic text as their Old Testament basis.

Biblical manuscript texts and editions

For the New Testament the chief text-critical materials are (1) manuscripts of the Greek text, from the 2nd to the 15th centuries, of which some 5,000 are known, exhibiting the New Testament text in whole or in part; (2) ancient versions in Syriac, Coptic, Latin, Armenian, Georgian, Ethiopic, and other languages; and (3) citations in early Christian writers. A comparative study of this material enables scholars to get behind the Byzantine type of text (the type that first diffused from Constantinople from the 4th century onward, gained currency throughout Greek-speaking Christendom, and formed the basis of the earliest printed editions of the Greek Testament) to a variety of types current in various localities in the generations immediately preceding; but the more recent discovery of manuscripts (mainly on papyrus) of the 3rd and even 2nd centuries, which cannot be neatly assigned to one or another of these types, makes the earlier history of the text more problematic, and the Revised Standard Version and New English Bible are both based on an eclectic text (in which, where the witnesses show variant readings, the reading preferred is that which best suits the context and the author's known style).

New Testament textual problems

Philological criticism. Philological criticism consists mainly in the study of the biblical languages in their widest scope, so that the vocabulary, grammar, and style of the biblical writings can be understood as accurately as possible with the aid not only of other biblical writings but of other writings in the same or cognate languages. New Testament Greek, for example, is a representative of Hellenistic Greek written in the 1st century AD, ranging from the literary Hellenistic of Hebrews, I Peter, and portions of Luke–Acts, to the colloquial or vernacular idiom of some other books (*e.g.,* the conversations in the Gospels). Some Aramaic influences have been discerned in parts of the New Testament that have a Palestinian setting, but not to a point where scholars are obliged to conclude that some books, or parts of books, were originally composed in Aramaic. Moreover, the Septuagint version exercised on some New Testament writers the kind of influence that the King James Version has exercised on many English writers, especially in the provision of a theological vocabulary in areas such as law, ethics, atonement, and sacrifice. The study of Old Testament Hebrew has been enriched by the study of other Semitic languages—Akkadian and Ugaritic among the ancient languages, and Arabic, which preserves many archaic features. Such comparative study has led to the suggestion of new meanings for a considerable number of biblical Hebrew words—a tendency that is amply illustrated by the New English Bible—but this department of philological criticism requires much more carefully defined guiding lines than have hitherto been laid down.

Literary criticism. Literary criticism endeavours to establish the literary genres (types or categories) of the various documents and to reach conclusions about their structure, date, and authorship. These conclusions are based as far as possible on internal evidence, but external evidence is also very helpful, especially where date is concerned. If the document under consideration is unmistakably quoted in another composition, for example, that quotation forms a *terminus ante quem* (later limiting point in time) for dating purposes. If, on the other hand, the document is clearly dependent on another document that can be dated on independent grounds, the date of the earlier document provides a *terminus post quem* (earlier limiting point in time).

Importance of literary genres

Proved dependence on such an earlier document may also throw light on the structure of the work being studied. But much of the evidence for the history of its structure is internal. The evaluation of such evidence is the province of what used to be called the higher criticism, a term first employed with a biblical reference by the German biblical scholar and orientalist Johann Gottfried Eichhorn (1752–1827):

> I have been obliged to bestow the greatest amount of labour on a hitherto entirely unworked field, the investigation of the inner constitution of the separate books of the Old Testament by the aid of the higher criticism (a new name to no humanist).

Eichhorn paid special attention to the Pentateuch; his work marks an important step forward in Pentateuchal criticism. The chronological arrangement of the successive law codes contained in the Pentateuch, or of the successive editions of one fundamental law code, has been related to the history of Israelite culture and religion recorded in the other Old Testament books—histories, prophecies, and psalms—with the mounting aid supplied by contemporary non-Israelite documents. The development of some Old Testament books is indicated expressly in their contents: one can note the composition of the first and second editions of the Book of Jeremiah in Jer. 36:4, 32; and scholars can reach some conclusions about later editions by a comparison of the longer edition in the Masoretic text with the shorter edition in the Septuagint (now also attested in a fragmentary Hebrew text from Qumrān). In the absence of such explicit evidence, conclusions about the structure of other prophetic books, such as Isaiah and Ezekiel, must be more tentative.

In the New Testament, literary criticism has centred principally on the Gospels. In the Synoptic Gospels (that is, those having a common source; *i.e.,* Matthew, Mark, and Luke) indicators as to source and composition are provided by the presence of so much material common to two or to all three of them. The majority opinion for well over a century has been that Mark served as a source for Matthew and Luke, and that the two latter had a further common source, generally labelled Q (for *Quelle,* the German term for "source"), comprising mainly sayings of Jesus. Aspects of the Gospel problem that literary criticism leaves unsolved are more likely to be illuminated by other critical approaches. The Fourth Gospel (John), having much less in common with the Synoptic Gospels than the latter three have among themselves, presents an independent line of transmission, and a comparative study of those areas where the Johannine and Synoptic traditions touch each other yields valuable conclusions for the beginnings of the gospel story.

Tradition criticism. Tradition criticism takes up where

Tracing the oral tradition

literary criticism leaves off; it goes behind the written sources to trace the development of oral tradition, where there is reason to believe that this preceded the earliest documentary stages, and attempts to trace the development of the tradition, phase by phase, from its primary life setting to its literary presentation. The development of the tradition might cover a lengthy period, as in the Old Testament narratives of the patriarchs—Abraham, Isaac, and Jacob—and the judges, such as Deborah and Samuel, many of which were originally attached to particular sanctuaries. The recognition of the life setting of each successive phase is necessary to the interpretation of the material received and delivered by one generation after another.

In the New Testament, too, special attention has been paid to the oral stage of the Gospel tradition, though here the preliterary period is measured in decades, not (as in the Old Testament) in generations and centuries. Not only the record of the ministry of Jesus but the development of Christian theology in the short preliterary stage has formed the subject matter of this study.

Form criticism. Form criticism has become one of the most valuable tools for the reconstruction of the preliterary tradition. This discipline classifies the literary material according to the principal "forms"—such as legal, poetic, and other forms—represented in its contents, and examines these in order to discover how they were handed down and what their successive life settings were until they assumed their present shape and position. In their various ways laws, narratives, psalms, and prophecies are amenable to this approach. By this means some scholars have undertaken to recover the *ipsissima verba* ("very own words") of Jesus by removing the accretions attached to them in the course of transmission. The exegetical task assumes a threefold shape as scholars work back from (1) interpretation of the present Gospels through (2) interpretation of the tradition lying behind them to (3) reconstruction of the proclamation of Jesus.

Scholars are not left completely to speculation as they attempt to reconstruct the stages by which the Gospel tradition attained its final form: here and there in the New Testament letters, and in some of the speeches included in Acts (which convey the general sense of what was said and should not be regarded as the author's free creations), there are fragments and outlines of the story of Jesus and of his teaching. Sometimes the characteristic terminology of tradition ("I received . . . I delivered") is used when such fragments are introduced, a decade or so before the composition of the earliest Gospel (*cf.* I Cor. 11:23; 15:3).

Other types of exegetical critical techniques. *Redaction criticism.* Redaction criticism concentrates on the end product, studying the way in which the final authors or editors used the traditional material that they received and the special purpose that each had in view in incorporating this material into his literary composition. It has led of late to important conclusions about the respective outlooks and aims of the four evangelists, Matthew, Mark, Luke, and John.

Biblical texts in relation to their historical milieu

Historical criticism. Historical criticism places the documents in their historical setting and promotes their interpretation in the light of their contemporary environment. This is necessary for their understanding, whether they are historical in character or belong to another literary genre. If they are historical in character it is important to establish how faithfully they reflect their dramatic date—the date of the events they record (as distinct from the date of final composition). This test has been applied with singularly positive results to Luke–Acts, especially in relation to Roman law and institutions; and in general the biblical outline of events from the middle Bronze Age (*c.* 21st–*c.* mid-16th centuries BC) to the 1st century AD fits remarkably well into its Near Eastern context as recovered by archaeological research.

"History of religions" criticism. "History of religions" criticism, to use an ungainly expression, relates Old and New Testament religion to the religious situation of the contemporary world of the writings and tries to explain biblical religion as far as possible in terms of current religious attitudes and practices. This is helpful to a point, insofar as it throws into relief those features of Hebrew and Christian faith that are distinctive; it is carried to excess when it attempts to deprive those features of their unique qualities and to account completely for them in religious–historical terms. When the cult of Israel was practically indistinguishable from that of the Canaanites, the protests of the 8th-century-BC Hebrew prophets Amos or Hosea stand out over against popular Yahweh worship (Hebrew) and Baal worship (Canaanite) alike. Another attempt has been made by historians of religion to recreate for the 1st century AD a pre-Christian Gnostic myth—referring to an esoteric dualism in which matter is viewed as evil and spirit good—of the primal or heavenly man who comes from the realm of light to liberate particles of a heavenly essence that are imprisoned on earth in material bodies and to impart the true knowledge. By men's acceptance of this secret salvatory knowledge (gnosis), the heavenly essence within man is released from its thraldom and reascends to its native abode. Fragments of this myth have been recognized in several books of the New Testament. But the attempt has not been successful: according to many recent (latter half of the 20th century) New Testament scholars and historians of the early church, it is probable that the concepts of primal man and redeemer-revealer were not brought together in Gnosticism *except* under the influence of the Christian apostolic teaching, in which Jesus fills the role of Son of man (or Second Adam) together with that of Saviour and Revealer.

On the other hand, the Iranian religious influence, primarily that of Zoroastrianism, on the angelology and eschatology (concepts of the last times) of Judaism in the last two centuries BC is unmistakable, especially among the Pharisees (a liberal Jewish sect emphasizing piety) and the Qumrān community (presumably the Essenes) near the Dead Sea. In the latter, indeed, Zoroastrian dualism finds clear expression, such as in the concept of a war between the sons of light and the sons of darkness, although it is subordinated to the sovereignty of the one God of Israel.

The value of these critical methods of Bible study lies in their enabling the reader to interpret the writings as accurately as possible. By their aid he can ascertain better what the writers meant by the language they used at the time they wrote and how their first readers would have understood their language. If the understanding of readers today is to have any validity, it must bear a close relation to what the original readers were intended to understand.

For additional information about the various forms of biblical criticism, see above: *Old Testament canon, texts, and versions*; and *New Testament canon, texts, and versions.*

TYPES OF BIBLICAL HERMENEUTICS

As has been said, the importance of biblical hermeneutics has lain in the Bible's status as a sacred book in Judaism and Christianity, recording a divine revelation or reproducing divine oracles. The "oracles" are primarily prophetic utterances, but often their narrative setting has also come to acquire oracular status. Quite different hermeneutical principles, however, have been inferred from this axiom of biblical inspiration: whereas some have argued that the interpretation must always be literal, or as literal as possible (since "God always means what he says"), others have treated it as self-evident that words of divine origin must always have some profounder "spiritual" meaning than that which lies on the surface, and this meaning will yield itself up only to those who apply the appropriate rules of figurative exegesis. Or again, it may be insisted that certain parts must be treated literally and others figuratively; thus some expositors who regard the allegorical (symbolic) interpretation of the Old Testament histories as the only interpretation that has any religious value maintain that in the apocalyptic writings that interpretation which is most literal is most reliable.

Interpretation of biblical inspiration

Literal interpretation. Literal interpretation is often, but not necessarily, associated with the belief in verbal or plenary inspiration, according to which not only the biblical message but also the individual words in which that message was delivered or written down were divinely chosen. In an extreme form this would imply that God dictated the

message to the speakers or writers word by word, but most proponents of verbal inspiration repudiate such a view on the reasonable ground that this would leave no room for the evident individuality of style and vocabulary found in the various authors. Verbal inspiration received classic expression by the 19th-century English biblical scholar John William Burgon:

> The Bible is none other than *the voice of Him that sitteth upon the Throne!* Every Book of it, every Chapter of it, every Verse of it, every word of it, every syllable of it, (*where* are we to *stop?*) every letter of it, is the direct utterance of the Most High! (From *Inspiration and Interpretation,* 1861).

This explains Burgon's severe judgment that the revisers of the English New Testament (1881), in excluding what they believed to be scribal or editorial additions to the original text, "stand convicted of having deliberately rejected the words of Inspiration in every page" (*The Revision Revised,* p. vii, London, 1883). Such a high view of inspiration has commonly been based on the statement in II Tim. 3:16 that "all [Old Testament] scripture is God-breathed" (Greek *theopneustos,* which means "inspired by God") or Paul's claim in I Cor. 2:13 to impart the gospel "in words not taught by human wisdom but taught by the Spirit, interpreting spiritual truths in spiritual language." On this latter passage the English bishop and biblical scholar Joseph Barber Lightfoot (1828–89) remarked:

> The notion of a verbal inspiration in a certain sense is involved in the very conception of an inspiration at all, because words are at once the instruments of carrying on and the means of expressing ideas, so that the words must both lead and follow the thought. But the passage gives no countenance to the popular doctrine of verbal inspiration, whether right or wrong (From *Notes on Epistles of St. Paul from Unpublished Commentaries,* 1895).

The detailed attention that Lightfoot and his Cambridge University colleagues, Brooke Foss Westcott (1825–1901), successor of Lightfoot as bishop of Durham, and Fenton John Anthony Hort (1828–92), paid in their exegesis to the vocabulary and grammatical construction of the biblical documents, together with their concern for the historical context, sprang from no dogmatic attachment to any theory of inspiration but represented the literal method of interpretation at its best. Such grammatico-historical exegesis can be practiced by anyone with the necessary linguistic tools and accuracy of mind, irrespective of confessional commitment, and is likely to have more permanent value than exegesis that reflects passing fashions of philosophical thought. Biblical theology itself is more securely based when it rests upon such exegesis than when it forms a hermeneutical presupposition.

The Bible as a guide to conduct

Moral interpretation. Moral interpretation is necessitated by the belief that the Bible is the rule not only of faith but also of conduct. The Jewish teachers of the late pre-Christian and early Christian Era, who found "in the law the embodiment of knowledge and truth" (Rom. 2:20), were faced with the necessity of adapting the requirements of the Pentateuchal codes to the changed social conditions of the Hellenistic Age (3rd century BC–3rd century AD). This they did by means of a body of oral interpretation, which enabled the conscientious Jew to know his duty in the manifold circumstances of daily life. If, for example, he wished to know whether this or that activity constituted "work" that was forbidden on the sabbath, the influential school of legal interpretation headed by the rabbi Hillel (late 1st century BC to early 1st century AD) supplied a list of 39 categories of activity that fell under the ban.

The Christian Church rejected the Jewish "tradition of the elders" but for the most part continued to regard the Ten Commandments as ethically binding and devised new codes of practice, largely forgetting Paul's appeal to the liberty of the Spirit, or viewing it as an invitation to indulge in allegory. In order to deduce moral lessons from the Bible, allegorization was resorted to, as when the *Letter of Barnabas* (*c.* AD 100) interprets the Levitical food laws prescribed in the book of Leviticus as forbidding not the flesh of certain animals but the vices imaginatively associated with the animals. To set up principles of exegesis by which ethical lessons may be drawn from all parts of the Bible is not easy, since many of the commandments enjoined upon the Israelites in the Pentateuch no longer have any obvious relevance, such as the ban on boiling a kid in its mother's milk (Ex. 23:19b, etc.), or on wearing a mixed woollen and linen garment (Deut. 22:11); and much of the teaching of Jesus in the Sermon on the Mount is widely regarded as a counsel of perfection, impracticable for the average man, even when he professes the Christian faith. Even summaries of the biblical ethic, such as the golden rule (Matt. 7:12; *cf.* Tob. 4:15) or the twofold law of love to God and love to one's neighbour (Deut. 6:5; Lev. 19:18), in which the Decalogue (Ten Commandments) is comprehended (Mark 12:29–31; *cf.* Rom. 13:8–10), involve casuistic interpretation (fitting general principles to particular cases) when they are applied to the complicated relations of present-day life. The difficulties of applying biblical ethics to modern situations do not mean that the task of application should be abandoned but that it should not be undertaken as though it provided an easy shortcut to moral solutions.

Allegorical interpretation. Allegorical interpretation places on biblical literature a meaning that, with rare exceptions, it was never intended to convey. Yet at times this interpretation seemed imperative. If the literal sense, on which heretics, such as the 2nd-century biblical critic Marcion, and anti-Christian polemicists, such as the 2nd-century philosopher Celsus, insisted, was unacceptable, then allegorization was the only procedure compatible with a belief in the Bible as a divine oracle. Law, history, prophecy, poetry, and even Jesus' parables yielded new meanings when allegorized. The surface sensuous meaning of the Canticles (the Song of Solomon) was gladly forgotten when its mutual endearments were understood to express the communion between God and the soul, or between Christ and the church. There are still readers who can reconcile themselves to the presence of a book such as Joshua in the canon only if its battles can be understood as pointing to the warfare of Christians "against the spiritual hosts of wickedness in the heavenly places" (Eph. 6:12). As for the Gospel parables, when in the story of the good Samaritan (Luke 10:30–37) an allegorical meaning is sought for the thieves, the Samaritan's beast, the inn, the innkeeper, and the two pence, the result too often is that the explicit point of the story, "Go and do likewise," is blunted.

Typological interpretation

Closely allied to allegorical interpretation, if not indeed a species of it, is typological interpretation, in which certain persons, objects, or events in the Old Testament are seen to set forth at a deeper level persons, objects, or events in the New. In such interpretations, Noah's ark (Gen. 6:14–22) is interpreted to typify the church, outside which there is no salvation; Isaac carrying the wood for the sacrifice (Gen. 22:6) typifies Jesus carrying the cross; Rahab's scarlet cord in the window (Jos. 2:18–21) prefigures the blood of Christ; and so on. These are not merely sermon illustrations but rather aspects of a hermeneutical theory that maintains that this further significance was designed (by God) from the beginning. Traces of typology appear in the New Testament, as when Paul in Rom. 5:14 calls Adam a "type" of the coming Christ (as the head of the old creation involved its members in the results of his disobedience, so the head of the new creation shares with its members the fruit of his obedience), or when in I Cor. 10:11 he says that the Israelites' experiences in the wilderness wanderings befell them "typically," so as to warn his own converts of the peril of rebelling against God. The fourth evangelist stresses the analogy between the sacrificial Passover lamb of the Hebrews and Christ in his death (John 19). The writer of the Hebrews treats the priest-king of Salem, Melchizedek, who was involved with Abraham as a type of Christ (Heb. 7)—without using the word "type"—and the Levitical ritual of the Day of Atonement as a model (though an imperfect one) of Christ's sacrificial ministry (Heb. 9).

Other hermeneutical principles. *Anagogical interpretation.* Anagogical (mystical or spiritual) interpretation seeks to explain biblical events or matters of this world so that they relate to the life to come. Jordan is thus interpreted as the river of death; by crossing it one enters into the heavenly Canaan, the better land, the "rest that remains

for the people of God." "The Jerusalem that now is" points to the new Jerusalem that is above. In Judaism of the closing centuries BC, the Eden of Genesis, the earthly paradise, lent its name to the heavenly paradise mentioned occasionally in the New Testament (Luke 23:43; II Cor. 12:3; Rev. 2:7).

Another form of mystical interpretation is the Mariological (referring to Mary, the mother of Jesus) application of scriptures that have another contextual sense. Thus Mary is the second Eve, whose offspring bruises the serpent's head (Gen. 3:15); Mary is the star-crowned woman of Rev. 12, whose son is caught up to the throne of God, and in more popular piety the dark-faced Madonna of the monastery at Montserrat, near Barcelona, Spain, can be identified with the "black but comely" bride of the Song of Solomon.

Parallelism. Parallelism, the interpretation of Scripture by means of Scripture, is a corollary of the belief in the unity of Scripture. But as a hermeneutical principle it must be employed sparingly, since the unity of Scripture should be based on comprehensive exegetical study, rather than itself provide a basis. Where one or two biblical documents (*e.g.,* the letters to the Romans and to the Galatians) are treated as the norm of biblical doctrine, there is a danger that other parts of the volume (*e.g.,* the Letter to the Hebrews) will be forced to yield the same sense as the "normative" documents; the distinctiveness of certain biblical authors will then be blurred. One naive form of parallelism is the "concordant" method, in which it is axiomatic that a Hebrew or Greek word will always (or nearly always) have the same force wherever it occurs in the Bible, no matter who uses it. There is, again, a harmonistic tradition that smooths out disparities in the biblical text (*e.g.,* as between the gospel narratives or the parallel records of Kings and Chronicles) in a manner that imposes a greater strain on faith than do the disparities themselves.

Concordant and harmonistic methods

One exegetical device of the Jewish rabbis (teachers, biblical commentators, and religious leaders) was that of *gezera shawa,* "equal category," according to which an obscure passage might be illuminated by reference to another containing the same key term. There are several examples in Paul's Old Testament exegesis, one of the best known being in Gal. 3:10–14, where the mystery of Christ's dying the death that incurred the divine curse (Deut. 21:23) is explained by his bearing vicariously the curse incurred by the lawbreaker (Deut. 27:26). One may compare the explanation in Heb. 4:3–9 of God's "rest" mentioned in Ps. 95:11 by reference to his resting on the seventh day after creation's work (Gen. 2:3)—an explanation dependent on the Septuagint, not the Hebrew.

Analogical interpretation. Analogical interpretation traditionally includes not only interpretation according to the analogy of Scripture (parallelism, in other words) but also interpretation according to the "analogy of faith"—an expression that misapplies the language of Rom. 12:6 in the King James Version of 1611. It has at times been pressed to mean that no biblical interpretation is valid unless it conforms to the established teaching of a religious community, to the verdict of tradition, or to the "unanimous consensus of the fathers." Where the established teaching is based, in intention, on Scripture, then an interpretation of Scripture that conflicts with it naturally calls for further scrutiny, but such conflict does not rule out the interpretation beforehand; if the conflict is confirmed, it is the established teaching that requires revision.

Other types. There is an unconscious tendency to conform hermeneutical principles to the climate of opinion in and around the community concerned, and to change the hermeneutic pattern as the climate of opinion changes. It is not surprising that in the circles where Pseudo-Dionysius (early-6th-century writings attributed to Dionysius, a convert of St. Paul) was revered as a teacher, Scripture was interpreted in Neoplatonic (idealistic and mystical) categories, and if in the latter half of the 20th century there is an influential and persuasive school of existential hermeneutics, this may be as much due to a widespread contemporary outlook on life as was the liberal hermeneutic of the preceding generations.

Existential interpretations

At a far different level contemporary movements continue to influence biblical interpretation. The interpretation of prophecy and apocalyptic in terms of events of the interpreter's day, which has ancient precedent, is still avidly pursued. Just as in the 16th century the apocalyptic beast of Revelation was interpreted to be the papacy or Martin Luther (in accordance with the interpreter's viewpoint), so also today in some nonacademic circles the ten kings denoted by the beast's horns in Revelation are identified with the European Economic Community in its ultimate development, or the threat to "destroy the tongue of the sea of Egypt" (Isa. 11:15) is believed to be fulfilled in the condition of the Suez Canal in the years following 1967. Whatever critical exegetes think of such aberrations, historians of exegesis will take note of them and recognize the doctrine of Scripture that underlies them.

THE DEVELOPMENT OF BIBLICAL EXEGESIS AND HERMENEUTICS IN JUDAISM

Early stages. The beginnings of biblical exegesis are found in the Old Testament itself, where earlier documents are interpreted in later documents, as in the recasting of earlier laws in later codes, or the Chronicler's reworking of material in Samuel and Kings. In addition, even before the Babylonian Exile (586 BC) there is evidence of the kind of midrashic exposition (nonliteral interpretations) familiar in the rabbinical period (*c.* 300 BC–*c.* AD 500) and after.

In Isa. 40 and following, the restoration of Israel after the return from exile is portrayed as a new creation: the characteristic verbs of the Genesis creation narrative—"create" (*bara*), "make" (*'asa*) and "form" (*yatzar*)—are used of this new act of God (*e.g.,* Isa. 43:7). Even more clearly are the same events portrayed as a new Exodus: on their journey back from Babylon, as earlier through the wilderness, the God of Israel makes a way for his people; he protects them before and behind; he champions them "with a mighty hand and an outstretched arm," he brings water from the rock for their sustenance (Isa. 43:2, 16, 19; 48:21; 52:12; Ezek. 20:33).

A pattern of divine action in mercy and judgment is discernible as one moves from the earlier prophets to the later prophets and apocalyptists (those concerned with the intervention of God in history). Yahweh's "strange work" in bringing the Assyrians against Israel in the 8th century BC (Isa. 28:21; 29:14) is repeated a century later when he raises up the Chaldaeans (Babylonians) to execute his judgment (Hab. 1:5 fol.). Ezekiel's visionary figure Gog is the invader whose aggression was foretold in earlier days by Yahweh through his "servants the prophets" (Ezek. 38:17), and one may recognize in him a revival not only of Isaiah's Assyrian (Isa. 10:4 fol.) but also of Jeremiah's destroyer from the north (Jer. 1:14 fol.; 4:6 fol.). The same figure reappears in the last "king of the north" in Dan. 11:40 fol.; he too is diverted from his path by "tidings from the east and the north" (*cf.* Isa. 37:7) and "shall come to his end, with none to help him" (*cf.* Isa. 31:8).

Later events as recapitulations of earlier events

In some degree these later predictions are interpretations, or reinterpretations, of the earlier ones, as when the non-Israelite prophet Balaam's "ships . . . from Kittim" (Num. 24:24) are interpreted in Dan. 11:30 as the Roman vessels off Alexandria in 168 BC that frustrated the Syrian king Antiochus IV Epiphanes (*c.* 215–164/163 BC) in his attempt to annex Egypt.

Ezra (*c.* 400 BC), whose role as the archetypal "scribe" is magnified by tradition, is said in the canonical literature to have brought the law of God from Babylonia to Jerusalem (Ezra 7:14), where it was read aloud to a large assembly by relays of readers "with interpretation"—and "they gave the sense, so that the people understood the reading" (Neh. 8:8). This may be the first recorded use of an Aramaic Targum—a paraphrase of the Hebrew that included interpretation as well as translation.

In the scribal and rabbinic tradition, two forms of exposition were early distinguished—*peshaṭ,* "plain meaning," and *derash,* "interpretation," by which religious or social morals were derived, often artificially, from the text. There was, however, no sense of conflict between the two.

The Hellenistic period. The translation of the Hebrew Bible into Greek by Alexandrian Jews in the 2nd and 3rd

centuries BC provided opportunities for recording interpretations that were probably current in Hellenistic Judaism. Literal translations might be misleading to Greek readers; metaphors natural in Hebrew were rendered into less figurative Greek. "Walking with God" or "walking before God" was rendered as "pleasing God." Such renderings are scarcely to be called anti-anthropomorphisms (that is, against depicting God in human terms or forms). In certain books there are some renderings that might be so described: in Ex. 24:10, for example, "they saw the God of Israel" becomes "they saw the place where the God of Israel stood"; but an examination of the Hebrew context suggests that this is precisely what was seen.

Metaphorical and philosophical interpretations

There was a tendency to universalize certain particularist statements of the Hebrew: in Amos 9:11 fol. the prophecy that David's dynasty will repossess the residue of Edom becomes a promise that the residue of men (the Gentiles) will seek the true God—a promise that is quoted in the New Testament as a "testimony" to the Christian Gentile mission.

The other main contribution to biblical exegesis in Alexandria was made by the Jewish philosopher Philo (*c.* 30/*c.* 20 BC–after AD 40), whose interpretation of the Pentateuch in terms of Platonic idealism and Stoic ethics had more influence on Christian than on Jewish hermeneutics.

In Palestinian Judaism the most distinctive exegetical work in the Hellenistic period was that of the Qumrān community (*c.* 130 BC–AD 70). The community, believing itself raised up to prepare for the new age of everlasting righteousness, interpreted Scripture so as to find there the divine purpose about on the point of fulfillment, together with its own duty in the impending crisis. Biblical prophecies in the Qumrān commentaries refer to persons and events of the recent past, the present, or the imminent future. The time of their fulfillment was concealed from the prophets; only when this was revealed to the Teacher of Righteousness, the organizer of the community, could their intent be grasped.

Rabbinic exegesis was present in all the varieties of rabbinic literature but is found especially in the Targumim and Midrashim (plural of Targum and Midrash). Among the former, special interest attaches to the early Palestinian Pentateuch Targum; it preserves, for example, messianic (referring to the expected anointed deliverer) exegesis of certain passages to which later rabbis gave a different interpretation because of the Christians' appeal to them. The earlier Midrashim—those whose contents are not later than AD 200—expound Exodus, Leviticus, Numbers, and Deuteronomy and are almost entirely Halakhic—*i.e.,* recording legal interpretations from various schools. The later Midrashim are more homiletic and include a considerable element of Haggada; *i.e.,* illustrative material drawn from all sources.

Rabbinic exegesis was not haphazard; it observed certain rules, which were variously formulated in the schools. The name of the famous interpreter Hillel is linked with seven *middot*, or norms; (1) inference from less important to more important and vice versa, (2) inference by analogy, (3) the grouping of related passages under an interpretative principle that primarily applies to one of them, (4) similar grouping where the principle primarily applies to two passages, (5) inference from particular to general and vice versa, (6) exposition by means of a similar passage, (7) inference from the context. By the time of Rabbi Ishmael (*c.* AD 100) these rules were expanded to 13, and Eliezer ben Yose the Galilaean (*c.* AD 150) formulated 32 rules, reflecting rational principles of exegesis, which remained normative into the Middle Ages.

Norms of interpretation

The medieval period. By the beginning of the Middle Ages the Masoretes of Babylonia and Palestine (6th–10th century) had fixed in writing, by points and annotation, the traditional pronunciation, punctuation, and (to some extent) interpretation of the biblical text. The rise of the Karaites, who rejected rabbinic tradition and appealed to Scripture alone (8th century onward) stimulated exegetical study in their own sect and in Judaism generally: in reaction against them Sa'adia ben Joseph (882–942), who was the *gaon*, or head, of the Sura academy in Babylonia, did some of his most important work. He adopted as one basic principle that biblical interpretation must not contradict reason. He translated most of the Bible into Arabic and composed an Arabic commentary on the text.

The French Jewish biblical and Talmudic scholar Rashi (Rabbi Shlomo Yitzhaqi of Troyes, 1040–1105), the most popular of all Jewish commentators, paid careful heed to the language and rejected those midrashic traditions that were inconsistent with the plain meaning of the text. Abraham ibn Ezra, of Spanish birth (1092/93–1167), in some respects anticipated the Pentateuchal literary criticism of later centuries. Other important names are Joseph Qimḥi of Narbonne and his sons Moses and David, the last of whom (*c.* 1160–1235) commented on the prophets and psalms; his psalms commentary took issue especially with Christian exegesis.

The great philosopher and codifier Maimonides (Moses ben Maimon, 1135–1204) composed, among many other works, his *Guide of the Perplexed* to help readers who were bewildered by apparent contradictions between the biblical text and the findings of reason. Like his younger contemporary David Qimḥi, he classified some biblical narratives as visionary accounts.

Far removed from the rational exegesis of these scholars was the mystical tradition, or Kabbala, which combined with an earlier mysticism—involving reflection on Ezekiel's inaugural chariot vision—the Neoplatonic doctrine of emanations. Adherents of this mystical exegesis found encouragement in the Pentateuch commentary of the Spanish Talmudist, Kabbalist, and biblical commentator Moses ben Naḥman (*c.* 1195–1270). The tracing of mystical significance in the numerical values of Hebrew letters and words (*gematria*) made a distinctive contribution to mystical exegesis. The chief monument of mystical exegesis is the 13th-century Spanish *Sefer ha-zohar* ("Book of Splendour"), in form a midrashic commentary on the Pentateuch. In the *Zohar* the *peshaṭ* (literal) and *derash* (nonliteral meanings) types of interpretation are accompanied by those called *remez* ("allusion"), including typology and allegory, and *sod* ("secret"), the mystical sense. The initials of the four were so arranged as to yield the word PaRDeS ("Paradise"), a designation for the fourfold meaning. The highest meaning led by knowledge through love to ecstasy and the beatific vision.

Mystical interpretation

The modern period. Following a line marked out earlier by the Spanish philosopher and poet Moses ibn Ezra (1060–1139), Benedict de Spinoza (1632–77) put forward a thoroughgoing reappraisal of the traditional account of the origin of the Pentateuch in his *Tractatus Theologico-Politicus* (1679). In the following century the Jewish Enlightenment (Haskala) brought a fresh appreciation of the Bible as literature. The pioneer of the Enlightenment, Moses Mendelssohn (1729–86), prepared a German translation of the Pentateuch, which he furnished (along with Solomon Dubno and others) with a commentary; he also translated the psalms and the Song of Solomon.

The tradition of orthodox Jewish exegesis has been maintained to the present day. In the 19th century the Russian rabbi Meir ben Yehiel Michael, "Malbin," (1809–79) wrote commentaries on the prophets and the writings, making a special point of explaining differences between synonyms. In the 20th century the traditional values of Judaism were popularly expounded in Joseph Herman Hertz's commentary on *The Pentateuch and Haftorahs* (1929–36) and in the Soncino *Books of the Bible* (1946–51). Martin Buber (1878–1965), the great modern Jewish philosopher, imparted to his many studies in biblical literature and religion—including his revolutionary German translation of the Bible (1926 and following), partly executed in association with the religious philosopher Franz Rosenzweig (1886–1926)—the qualities of his personal genius that was influenced by Ḥasidic (18th-century mystical) piety and an existential interpretation of life.

20th-century commentators

In recent decades the most valuable Jewish exegesis has been in association with the wider world of biblical scholarship. Journals such as the *Jewish Quarterly Review* and the *Hebrew Union College Annual* welcome contributions from non-Jewish scholars; in interconfessional projects such as the Anchor Bible, Jewish scholars cooperate in the Old and New Testament alike.

The whole field of biblical study, including exegesis, is cultivated most intensively in Israel. Yehezkel Kaufmann (1890–1963) produced the encyclopaedic *History of Israelite Religion from Its Beginnings to the End of the Second Temple* (8 vol., 1937–56) in Hebrew that pursues a path involving a radical revision of current biblical criticism and interpretation. Mosheh Zevi Hirsh Segal (died 1968) dealt with a wide area of biblical and related literature, maintaining the essential Mosaic authorship of the Pentateuch (supplemented by later editors who worked in Moses' spirit). The most ambitious enterprise in this field is the "Bible Project" of the Hebrew University of Jerusalem, which aims to produce a critical edition of the Hebrew Bible but also fosters a number of ancillary studies in biblical text and interpretation, mostly published in its annual report *Textus,* in which non-Jewish as well as Jewish scholars participate.

THE DEVELOPMENT OF BIBLICAL EXEGESIS AND HERMENEUTICS IN CHRISTIANITY

Early stages. The earliest Christian exegesis of the Old Testament is found in the New Testament, not in the written texts only but in the oral tradition lying behind them. Some lines of exegesis are present in so many separate strands of primitive Christian teaching that they are most reasonably assigned to Jesus, who began his Galilaean ministry with the announcement that the time appointed for the fulfillment of prophecy, and the Kingdom of God that was its main theme, had arrived. If the accomplishment of his ministry involved his death, that was accepted in the same spirit; he submitted to his captors with the words: ". . . Let the scriptures be fulfilled" (Mark 14:49). The church began with the conviction that Jesus, crucified and risen, was the one of whom the prophets spoke. He was the prophet like Moses, prince of the house of David, priest of the order of Melchizedek, servant of the Lord, Son of man, and exalted Lord. If the prophets themselves were uncertain about the person or time indicated by their oracles, the early Christians were certain: the person was Jesus, the time was now. The New Testament writers shared a creative and flexible principle of exegesis that has regard for the literary and historical context and traces a consistent pattern of divine action in judgment and mercy, reproduced repeatedly in the history of Israel and manifested definitively in Christ. This exegesis is elaborated at times by means of typology and allegory, as when Paul illustrates the relationship between law and gospel by the story of Hagar and Sarah, the concubine and wife of Abraham, respectively (Gal. 4:21–31), or when Israel's tabernacle in the wilderness becomes the material counterpart to the heavenly sanctuary in which believers of the new age offer spiritual worship to God (Heb. 8:2 fol.). The writer to the Hebrews, indeed, occasionally relates the old order to the new order platonically in terms of the earthly copy of an eternal archetype.

At an early date Christians developed a line of Old Testament exegesis designed to show that they, not the Jews, stand in the true succession of the original people of God. This line is seen in the *Letter of Barnabas,* the apologist Justin's (*c.* 100–*c.* 165) *Dialogue with Trypho,* and the 3rd-century *Against the Jews* ascribed to the North African bishop Cyprian (*c.* 200–258).

Alexandrian allegorization

The patristic period. Alexandria had long boasted a school of classical study that practiced the allegorical interpretation of the Homeric epics and the Greek myths. This method of exegesis was taken over by Philo and from him by Christian scholars of Alexandria in the 2nd and 3rd centuries. Clement of Alexandria (*c.* 150–*c.* 215) and Origen (*c.* 185–*c.* 254) did not completely rule out the literal sense of Scripture—Origen's *Hexapla,* a six column edition of various biblical versions, was a monument to his painstaking study of the text—but claimed that the most meaningful aspects of divine revelation could be extracted only by allegorization. Clement stated that the Fourth Gospel was a "spiritual gospel" because it unfolds the deeper truth concealed in the matter-of-fact narratives of the other three. Origen treated literal statements as "earthen vessels" preserving divine treasure; their literal sense is the body as compared with the moral sense (the soul) and the spiritual sense (the spirit). The true exegete, he claimed, pursues the threefold sense and recognizes the spiritual (allegorical) as the highest.

Later, the Antiochene fathers, represented especially by Theodore of Mopsuestia (*c.* 350–428/429) and John Chrysostom (*c.* 347–407), patriarch of Constantinople, developed an exegesis that took more account of literal meaning and historical context. But the allegorizers could claim that their method yielded lessons that (while arbitrary) were more relevant and interesting to ordinary Christians.

In the West, the Alexandrian methods were adopted by Ambrose (*c.* 339–397), bishop of Milan, and Augustine (354–430), bishop of Hippo, especially as formulated in the seven "rules" of Tyconius (*c.* 380), a Donatist heretic (one who denied the efficacy of sacraments administered by an allegedly unworthy priest), which classified allegorical interpretation in relation to: (1) the Lord and his church, (2) true and false believers, (3) promise and law, (4) genus and species, (5) numerical significance, (6) "recapitulation," and (7) the devil and his followers. There were other Latin exegetes, like Ambrosiaster (commentaries ascribed to Ambrose) and, supremely, Jerome (*c.* 347–419/420), the learned Latin Father, who paid close attention to the grammatical sense. In the Old Testament Jerome appealed from the Greek version to the "Hebraic verity" and in such a work as his commentary on Daniel provided some fine examples of historical exegesis. Augustine, though not primarily an exegete, composed both literal and allegorical commentaries and expository homilies on many parts of Scripture, and his grasp of divine love as the essential element in revelation supplied a unifying hermeneutical principle that compensates for technical deficiencies.

The medieval period. As the patristic age gave way to the scholastic age, the English monk Bede of Jarrow (died 735) wrote commentaries designed to perpetuate patristic exegesis, mainly allegorical: thus Elkanah with his two wives (1 Sam. 1:2) is interpreted as referring to Christ with the synagogue and the church.

The fourfold sense of Scripture

In the early Middle Ages the fourfold sense of Scripture—developed from Origen's threefold sense by subdividing the spiritual sense into the allegorical (setting forth the doctrine) and the anagogical (relating to the coming world)—was increasingly expounded and received its final authority from Thomas Aquinas (1225/26–74). For Thomas, the literal sense, expressing the author's intention, was a fit object of scientific study; the figurative senses unfolded the divine intention.

Medieval exegesis was greatly influenced by the *Glossa Ordinaria,* a digest of the views of the leading fathers and early medieval doctors (teachers) on biblical interpretation. This compilation owed much in its initial stages to Anselm of Laon (died 1117); it had reached its definitive form by the middle of the 12th century and provided the exegetical norm of the *Summa theologiae* ("Summation of Theology") of Thomas Aquinas and others.

For all the interest in allegory, literal interpretation was cultivated in many centres in the West, often with the aid of Hebrew, knowledge of which was obtainable from Jewish rabbis. One such centre was the Abbey of Saint-Victor at Paris, where Hugh (died 1141) compiled biblical commentaries that fill three volumes of J.-P. Migne's (1800–75) *Patrologiae Cursus Completus* (Series Latina) and indicate the commentator's dependence on Rashi as well as on his Christian predecessors. Of Hugh's disciples, Andrew, abbot of Wigmore (died 1175), carried on his master's tradition of literal scholarship, and Richard, the Scottish-born prior of Saint-Victor (died 1173) pursued a line more congenial to his mystical temperament. Herbert of Bosham (*c.* 1180) produced a commentary on Jerome's Hebrew Psalter. Robert Grosseteste, bishop of Lincoln (died 1253), wrote commentaries on the days of creation and the Psalter that both drew on the Greek fathers and profited by his direct study of the Hebrew text. Nicholas of Lyra (*c.* 1265–*c.* 1349), the greatest Christian Hebraist and expositor of the later Middle Ages, compiled *postillae,* or commentaries, both literal and figurative, on the whole Bible; he insisted that only the literal sense could establish proof. Luther ranked him among the best exegetes: "a fine soul, a good Hebraist and a true Christian."

The Reformation period. The English theologian John Colet (*c.* 1466–1519) broke with medieval scholasticism when he returned from the Continent to Oxford in 1496 and lectured on the Pauline letters, expounding the text in terms of its plain meaning as seen in its historical context. The humanist Erasmus (*c.* 1466–1536) owed to him much of his insight into biblical exegesis. By the successive printed editions of his Greek New Testament (1516 and following), Erasmus made his principal, but not his only, contribution to biblical studies.

Reformation principles

Martin Luther (1483–1546) was a voluminous expositor, insisting on the primacy of the literal sense and dismissing allegory as so much rubbish—although he indulged in it himself on occasion. The core of Scripture was to him its proclamation of Christ as the one in whom alone lay man's justification before God. John Calvin (1509–64), a more systematic expositor, served his apprenticeship by writing a youthful commentary on the Roman statesman and philosopher Seneca the Younger's (*c.* 4 BC–AD 65) *De clementia* ("Concerning Mercy"); systematic theologian though he was, he did not allow his theological system to distort the plain meaning of Scripture, and his philological–historical interpretation is consulted with profit even today.

Scientific exegesis was pursued on the Catholic side by scholars such as F. de Ribera (1591) and L. Alcasar (1614), who showed the way to a more satisfactory understanding of the Revelation. On the Reformed side, the *Annotationes in Libros Evangeliorum* (1641–50) by the jurist Hugo Grotius (1583–1645) were so objective that some criticized them for rationalism.

The modern period. The modern period is marked by advances in textual criticism and in the study of biblical languages and history, all of which contribute to the interpretation of the Bible. The German theologian J.A. Bengel's (1687–1752) edition of the Greek text of the New Testament with critical apparatus (1734), in which he framed the canon that "the more difficult reading is to be preferred," was followed by his exegetical *Gnomon Novi Testamenti* ("Introduction to the New Testament," 1742): "apply thyself wholly to the text," he directed; "apply the text wholly to thyself." The English bishop Robert Lowth's (1710–87) Oxford lectures on *The Sacred Poetry of the Hebrews,* published in Latin in 1753, greatly promoted the understanding of the poetry of the Old Testament by expounding the laws of its parallelistic structure. The German philologist Karl Lachmann (1793–1851) applied his expertise in classical criticism to editing the text of the New Testament; to him also belongs the credit of arguing that Mark was the earliest of the Gospels and a main source of Matthew and Luke (1835).

Source analysis and historical interpretations

The problem of the source analysis of the Pentateuch was given what for long appeared to be its final solution by Julius Wellhausen (1844–1918), who related the successive law codes to the development of the Israelite cultus. For the period preceding the 9th century BC, however, he operated in a historical vacuum that Near Eastern archaeology was in his day only beginning to fill; its subsequent findings have dictated radical modifications in his reconstruction of Israel's religious history. In the middle half of the 19th century, New Testament exegesis was overshadowed by the school of Ferdinand Christian Baur (1792–1860), which envisaged a sharply opposed Petrine (Peter) and Pauline (Paul) antithesis in the primitive church, followed in the 2nd century by a synthesis that is reflected in most of the New Testament writings. In France, Ernest Renan's (1823–92) works on early Christianity were helpful philological and historical studies; the most popular volume, his *Vie de Jésus* (1863), was the least valuable. In England, where the poet and educator Matthew Arnold (1822–88) endeavoured to find an impregnable moral foundation for biblical authority, New Testament exegesis received contributions of unsurpassed worth between 1865 and the end of the century from J.B. Lightfoot, B.F. Westcott, and F.J.A. Hort.

At the beginning of the 20th century a new direction was given to Gospel interpretation by the German scholar William Wrede (*Das Messiasgeheimnis in den Evangelien,* 1901) and the medical missionary theologian Albert Schweitzer (*The Quest of the Historical Jesus,* Eng. trans., 1910), who so emphasized the eschatological orientation of Jesus' mind and message that New Testament scholarship can never be the same again. The writings of the biblical scholar C.H. Dodd (*The Parables of the Kingdom,* 1935; *The Apostolic Preaching and Its Developments,* 1936) stressed realized eschatology—that the standards of the last times were realized by Jesus and his disciples—in the preaching of Jesus and of the primitive church; he has been a leading pioneer of the "biblical theology" movement. Karl Barth's (1886–1968) commentary on Romans (1919) launched an existential interpretation of the New Testament, which has been pursued more radically by Rudolf Bultmann (1884–1976), under the influence of Wilhelm Dilthey (1833–1911), according to whom the interpreter must project himself into the author's experience so as to relive it, and of Martin Heidegger (1889–1976), whose conception of the truly authentic man as capable of freedom because he has faced reality provides the "pre-understanding" for Bultmann's existential theology. Bultmann's disciple Ernst Fuchs considers the hermeneutical task to be the creation of a "language event" in which the authentic language of Scripture encounters one now, challenging decision, awakening faith, and accomplishing salvation. The chief rival to existential exegesis is the "salvation-history" hermeneutic espoused by Oscar Cullmann.

Existential and form-critical exegesis

Rudolf Bultmann and Martin Dibelius (1883–1947) pioneered the modern form-critical study of the Gospels. The form-critical method was fruitfully applied to the Old Testament by Hermann Gunkel (1862–1932) and Sigmund Mowinckel (1884–1965). Among Catholic scholars, exegetical studies are vigorously promoted by Jean Daniélou (with his researches into early Jewish Christianity), the Dominicans of the École Biblique et Archéologique (The School of the Bible and Archeology) in Jerusalem (to whom one must credit the Jerusalem Bible), and the Jesuits of the Pontifical Biblical Institute and others.

The encouragement given by the second Vatican Council (1962–65) of the Roman Catholic Church to biblical scholarship, to be cultivated in association with "separated brethren" and with consideration for the requirements of non-Christians, is one indication of a new direction in biblical exegesis, in which this study will no longer be pursued as a vindication of sectional traditions but rather as a cooperative enterprise aiming at making widely available the permanent value of the Bible. (F.F.B.)

BIBLIOGRAPHY

Biblical literature. *Nature and significance.* General articles and notes in *The Oxford Annotated Bible* (1962), *The Jerusalem Bible* (1966), and the *Genesis* volume of *The Anchor Bible,* by E.A. SPEISER (1964); E. H. GOMBRICH, *The Story of Art,* 12th rev. ed. (1972); ABRAHAM J. HESCHEL, *Man Is Not Alone* (1951), a classic statement of modern Judaism; ERICH AUERBACH, *Mimesis: Dargestellte Wirklichkeit in der abendländischen Literatur* (1946; Eng. trans., *Mimesis: The Representation of Reality in Western Literature,* 1953), a classic work; WILLIAM R. MUELLER, *The Prophetic Voice in Modern Fiction* (1959), religious themes interpreting Joyce, Camus, Kafka, Faulkner, Greene, and Silone; NATHAN A. SCOTT, JR. (ed.), *The Tragic Vision and the Christian Faith* (1957), essays by 12 writers on faith and the tragic dimension of existence; JAMES BARR, *The Scope and Authority of the Bible* (1981), questions the divine inspiration of biblical texts.

Old Testament canon, texts, and versions: OTTO EISSFELDT, *Einleitung in das Alte Testament,* 3rd ed. (1964; Eng. trans., *The Old Testament: An Introduction,* 1965); *The Cambridge History of the Bible* (CHB), 3 vol. (1963–70). (*The Canon*): FRANTS BUHL, *Kanon und Text des Alten Testaments* (1891; Eng. trans., *Canon and Text of the Old Testament,* 1892); MAX L. MARGOLIS, *The Hebrew Scriptures in the Making* (1922); HERBERT E. RYLE, *The Canon of the Old Testament,* 2nd ed. (1895); SOLOMON ZEITLIN, "An Historical Study of the Canonization of the Hebrew Scriptures," *Proceedings of the American Academy for Jewish Research,* pp. 121–158 (1932). (*Textual criticism, texts and manuscripts, and early versions*): FRANK MOORE CROSS, *The Ancient Library of Qumrân and Modern Biblical Studies,* 2nd ed. (1961); "The History of the Biblical Text in the Light of Discoveries in the Judaean Desert," *Harvard Theological Review,* 57:281–299 (1964); and "The Contribution of the Qumrân Discoveries to the Study of the Biblical Text," *Israel Exploration Journal,* 16:81–95 (1966);

CHRISTIAN D. GINSBURG, *Introduction to the Massoretico: Critical Edition of the Hebrew Bible* (1897, reprinted 1966); MOSHE H. GOSHEN-GOTTSTEIN, *Linguistic Structure and Tradition in the Qumran Documents* (1958); "Theory and Practice of Textual Criticism," *Textus,* 3:130–158 (1963); and *The Book of Isaiah: Sample Edition with Introduction* (1965); MOSHE GREENBERG, "The Stabilization of the Text of the Hebrew Bible," *Journal of the American Oriental Society,* 76:157–167 (1956). PAUL KAHLE, *The Cairo Genizah,* 2nd ed. (1959); FREDERICK G. KENYON, *The Bible and the Ancient Manuscripts,* 5th ed. rev. (1958); HARRY M. ORLINSKY, "The Textual Criticism of the Old Testament," in GEORGE E. WRIGHT (ed.), *The Bible and the Ancient Near East,* pp. 113–132 (1961); BLEDDYN J. ROBERTS, *The Old Testament Text and Versions* (1951); and "The Old Testament: Manuscripts, Text and Versions," *CHB,* vol. 2, pp. 1–26 (1969); P.W. SKEHAN, "Qumran and the Present State of Old Testament Text Studies," *Journal of Biblical Literature,* 78:21–25 (1959); S. TALMON, "Aspects of the Textual Transmission of the Bible in the Light of Qumran Manuscripts," *Textus,* 4:95–132 (1964); ERNST WURTHWEIN, *Der Text des Alten Testaments* (1952; Eng. trans., *The Text of the Old Testament,* 1957). (*Later and modern versions–English versions*): DAVID DAICHES, *The King James Version of the English Bible* (1941, reprinted 1968); MARGARET DEANESLY, *The Lollard Bible and Other Medieval Biblical Versions* (1920, reprinted 1966); HERMAN HAILPERIN, *Rashi and the Christian Scholars* (1963); WILLIAM F. MOULTON, *The History of the English Bible,* 5th ed. (1911); ALFRED W. POLLARD, *Records of the English Bible* (1911); and, with G.R. REDGRAVE, *A Short-Title Catalogue of Books Printed in England, Scotland, and Ireland and of English Books Printed Abroad 1475–1640* (1926, reprinted 1969); B.F. WESTCOTT, *A General View of the History of the English Bible,* 3rd ed. rev. by W.A. WRIGHT (1905). (*Continental versions and non-European versions*): THOMAS H. DARLOW and HORACE F. MOULE, *Historical Catalogue of the Printed Editions of the Holy Scripture in the Library of the British and Foreign Bible Society,* 2 vol. (1903–11); JOSEF SCHMID (ed.), "Moderne Bibelübersetzungen," *Zeitschrift für katholische Theologie,* 82:290–332 (1960).

Old Testament history: Two current histories of Israel exhibit the full range of historiographical problems and methods relating to the subject: JOHN BRIGHT, *A History of Israel* (1959); and MARTIN NOTH, *Geschichte Israels,* 3rd ed. (1956; Eng. trans., *The History of Israel,* 1958). They differ mainly in where they begin; Bright begins with Abraham, Noth with the federation of tribes that calls itself Israel in the land of Canaan. They disagree about the demonstrability of such a community in the pre-Canaanite times because of their respective assessment of the character of the Pentateuch. Bright assumes that it was intended as a history concerned to record the early past, while Noth assumes that its thematic traditions were intended to define and celebrate the identity of the later Israel and hence do not constitute a usable historical resource about its earliest beginnings. This whole methodological problem in Israelite historiography is lucidly discussed and illustrated in a little book by JOHN BRIGHT—*Early Israel in Recent History Writing: A Study in Method* (1956). For the use of archaeology, geography, and history of religion in the study of the history of Israel, see GEORGE ERNEST WRIGHT. *Biblical Archaeology,* rev. ed. (1962); LUC H. GROLLENBERG, *Atlas van de Bijbel,* 3rd ed. (1954; Eng. trans., *Atlas of the Bible,* 1956); YEHEZKEL KAUFMANN, *The Religion of Israel, from Its Beginnings to the Babylonian Exile* (1960); and HELMER RINGGREN, *Israelitische Religion* (1963; Eng. trans., 1966).

Old Testament literature: For various modern critical methods of studying the formation of the Old Testament, see the "Old Testament Series" of *Guides to Biblical Scholarship:* NORMAN C. HABEL, *Literary Criticism of the Old Testament,* GENE M. TUCKER, *Form Criticism of the Old Testament,* and WALTER E. RAST, *Tradition History and the Old Testament* (1971–72). Among general introductions, the most exhaustive is OTTO EISSFELDT (*op. cit.*), based mainly on literary criticism. The other methods are reflected to a somewhat greater extent in AAGE BENTZEN, *Introduction to the Old Testament,* 3rd ed. (1957); and in the briefer, less original but very readable work of ARTUR WEISER, *Einleitung in das Alte Testament,* 4th ed. (1957; Eng. trans., *The Old Testament: Its Formation and Development,* 1961). For pioneering research in tradition analysis of the Pentateuch and the Former Prophets, see MARTIN NOTH, *Überlieferungsgeschichte des Pentateuch,* 3rd ed. (1966; Eng. trans., *A History of Pentateuchal Traditions,* 1972), and *Überlieferungsgeschichtliche Studien* (1957); the latter deals with what its author calls "The Deuteronomic History," an envisioned work containing the books of Deuteronomy, Joshua, Judges, Samuel, and Kings. The contribution of form criticism to the understanding of the history of the Book of Psalms may best be approached through HERMANN GUNKEL, *The Psalms: A Form-Critical Introduction* (1967), a translation of his article in *Die Religion in Geschichte und Gegenwart* (2nd ed.) summarizing his seminal work in *Die Psalmen* (1926) and *Einleitung in die Psalmen* (1928). ELMER A. LESLIE, *The Psalms, Translated and Interpreted in the Light of Hebrew Life and Worship* (1949), is heavily dependent on Gunkel and illustrates his use of form criticism. The celebrated work of SIGMUND MOWINCKEL on the Psalter, culminating in his masterful *Offersang ob Sangoffer* (1951; Eng. trans., *The Psalms in Israel's Worship,* 2 vol., 1962), combines the methods of Gunkel with those of the comparative historian of religion and locates the setting for the production of most of the psalms in the cult of the Solomonic temple. The application of the newer methods to the study of the Latter Prophets is evident in the essays in HAROLD H. ROWLEY (ed.), *Studies in Old Testament Prophecy* (1950). The new approaches were deeply under the impact of HENRIK S. NYBERG, *Studien zum Hoseabuche* (1935). Other books that amplify the implications of his assumptions include: JOHANNES LINDBLOM, *Prophecy in Ancient Israel* (1962); CURT KUHL, *Israels Propheten* (1956; Eng. trans., *The Prophets of Israel,* 1960); and SIGMUND MOWINCKEL, *Prophecy and Tradition: The Prophetic Books in the Light of the Study of the Growth and History of the Tradition* (1946). ABRAHAM J. HESCHEL, *The Prophets* (1962), though of independent origin, belongs with those new interpretations of the prophetic materials. An old classic in a new edition, OLIVER S. RANKIN, *Israel's Wisdom Literature: Its Bearing on Theology and the History of Religion* (1936, reprinted 1969), presents Israel's wisdom literature in relation both to its extra-Israelite cultural connections and to the rest of Israel's heritage in the Old Testament. Two new approaches to the legacy of wisdom literature—through literary form and through theology—are presented, respectively, in R.B.Y. SCOTT, *The Way of Wisdom in the Old Testament* (1971); and GERHARD VON RAD, *Weisheit in Israel* (1970). See also NORTHROP FRYE, *The Great Code: The Bible and Literature* (1982), and ELSA TAMEZ, *Bible of the Oppressed* (1982), an interpretation from a Latin, female theologian's perspective.

Intertestamental literature: Standard translations of the Jewish intertestamental literature are ROBERT H. CHARLES (ed.), *The Apocrypha and Pseudepigrapha of the Old Testament in English* (1913); and EMIL KAUTZSCH (ed.), *Die Apocryphen und Pseudepigraphen des Alten Testaments* (1900). PAUL RIESSLER, *Altjüdisches Schrifttum ausserhalb der Bibel* (1928), is indispensable because it contains translations of the fullest number of writings. The best translations of the Dead Sea Scrolls are GEZA VERMES, *The Dead Sea Scrolls in English* (1962); JOHANN MAIER, *Die Texte vom Toten Meer* (1960); and ANDRE DUPONT-SOMMER, *Les Écrits esséniens découverts près de la Mer Morte,* 3rd ed. (1964). ALBERT-MARIE DENIS, *Introduction aux Pseudépigraphes grecs d'Ancien Testament* (1970), does not treat the Apocrypha and is important mainly for its bibliography. Basic books dealing with intertestamental literature are R.H. PFEIFFER, *History of New Testament Times, with an Introduction to the Apocrypha* (1949); EMIL SCHURER, *Geschichte des jüdischen Volkes im Zeitalter Jesu Christi,* 3rd–4th ed., 3 vol. (1898–1901; Eng. trans., *A History of the Jewish People in the Time of Jesus Christ,* 2nd and rev. ed., 5 vol., 1885–91); and ROBERT H. CHARLES, *Religious Development Between the Old and the New Testaments* (1914). Still interesting is ROBERT TRAVERS HERFORD, *Talmud and Apocrypha* (1933, reprinted 1971). Information about the library of the Dead Sea Scrolls is in two books: JOZEF T. MILIK, *Dix Ans de découvertes dans le désert de Juda* (1957; Eng. trans., *Ten Years of Discovery in the Wilderness of Judaea,* 1959); and FRANK MOORE CROSS, *The Ancient Library of Qumrân and Modern Biblical Studies,* 2nd ed. (1961). A fragment of Ben Sira from antiquity was published by YIGAEL YADIN, *The Ben Sira Scroll from Masada, with Introduction, Emendations and Commentary* (1965). The best book about Jewish eschatology is PAUL VOLZ, *Die Eschatologie der jüdischen Gemeinde im neutestamentlichen Zeitalter* (1934). On Apocalyptic and Messianism, see HAROLD H. ROWLEY, *The Relevance of Apocalyptic,* 3rd ed. (1963); DAVID S. RUSSELL, *The Method and Message of Jewish Apocalyptic, 200 BC–AD 100* (1964); SIGMUND MOWINCKEL, *Han som kommer* (1951; Eng. trans., *He That Cometh,* 1954); ERIK SJOBERG, *Der Menschensohn im äthiopischen Henochbuch* (1946); and A.S. VAN DER WOUDE, *Die messianischen Vorstellungen der Gemeinde von Qumrân* (1957).

New Testament canon, texts, and versions: (*Canon*): For the relevant primary texts on the history of the canon, see DANIEL J. THERON (ed.), *Evidence of Tradition* (1957), with selected source material in Greek or Latin with English translation; and JAMES STEVENSON (ed.), *A New Eusebius* (1957). For introductions, see ALEXANDER SOUTER and C.S.C. WILLIAMS, *The Text and Canon of the New Testament,* 2nd ed. rev. (1954); and ROBERT M. GRANT, *The Formation of the New Testament* (1965). For a phenomenological approach, see GERARDUS VAN DER LEEUW, *Phänomenologie der Religion,* 2nd ed., 2 vol. (1956; Eng. trans., *Religion in Essence and Manifestation,* 2nd

ed., 2 vol., 1963), ch. 64. (*Texts*): The major text for further study is BRUCE H. METZGER, *The Text of the New Testament* (1964). (*Translations*): On translation in general, see REUBEN A. BROWER (ed.), *On Translation* (1959). For translation of the Bible into English, see FREDERICK F. BRUCE, *The English Bible: A History of Translations from the Earliest English Versions to the New English Bible,* 2nd ed. (1970).

New Testament history: (*Jewish culture and history*): Standard works are R.H. PFEIFFER (*op. cit.*); and GEORGE F. MOORE, *Judaism in the First Centuries of the Christian Era,* 3 vol. (1927–30). (*Qumrān, Dead Sea Scrolls*): FRANK MOORE CROSS, JR. (*op. cit.*); on Qumrān and New Testament problems, see KRISTER STENDAHL (ed.), *The Scrolls and the New Testament* (1958). (*Greco-Roman culture and history*): WILLIAM W. TARN, *Hellenistic Civilisation,* 3rd ed. rev. (1952). For a broad cultural comparison, see ERIC R. DODDS, *Pagan and Christian in an Age of Anxiety* (1965). (*Pauline chronology*): The debate can be best assessed by comparing JOHN KNOX, *Chapters in a Life of Paul* (1950), with DIETER GEORGI, *Die Geschichte der Kollekte des Paulus für Jerusalem* (1965).

New Testament literature: The following works are useful for commentary, survey articles, and bibliographic material: GEORGE A. BUTTRICK (ed.), *The Interpreter's Bible,* especially vol. 1, 7, and 12 (1952–57); MATTHEW BLACK (ed.), *Peake's Commentary on the Bible,* 2nd ed. (1962); and RAYMOND E. BROWN, JOSEPH A. FITZMYER, and ROLAND E. MURPHY (eds.), *The Jerome Biblical Commentary* (1968). WERNER G. KUEMMEL, *The New Testament: The History of the Investigations of Its Problems* (1972), covers the whole history of New Testament studies with ample excerpts from the major scholars since the 18th century. For a rich introduction to the 27 books of the New Testament with full and balanced reporting on all major issues of contemporary discussion and extensive bibliographies, see PAUL FEINE, JOHANNES BEHM, and WERNER G. KUEMMEL, *Einleitung in das Neue Testament,* 14th rev. ed. (1965; Eng. trans., *Introduction to the New Testament,* 1966). For a general dictionary to the Bible, see GEORGE A. BUTTRICK (ed.), *The Interpreter's Dictionary of the Bible,* 4 vol. (1962). The most extensive tool for the study of New Testament theological terms is GERHARD KITTEL (ed.), *Theological Dictionary of the New Testament,* vol. 1–8 (1964–72, in progress). For New Testament theologies, see RUDOLF BULTMANN, *Theologie des Neuen Testaments,* 3rd ed. (1958; Eng. trans., *Theology of the New Testament,* 2 vol., 1951–55); HANS CONZELMANN, *Grundriss der Theologie des Neuen Testaments,* 2nd ed. (1967; Eng. trans., *An Outline of the Theology of the New Testament,* 1969). For general commentary, see *The International Critical Commentary on the Holy Scriptures of the Old and New Testaments,* 41 vol. (1895–1920); for the major German commentary, see *Kritisch exegetischer Kommentar über das Neuen Testament* ("Meyer Series," frequently updated); *Handbuch zum Neuen Testament* (Lietzmann-Bornkamm); and *Das Neue Testament Deutsch (Göttinger Bibelwerk).* For a major French Protestant commentary, see *Commentaire du Nouveau Testament* and for major French and German Roman Catholic commentaries, see *Etudes bibliques* and *Das Neue Testament übersetzt und erklärt* (the *Regensburger New Testament*). (*Gospels—texts*): KURT ALAND (ed.), *Synopsis Quattuor Evangeliorum* (1964), a Greek synopsis, includes the Gospel of John and translations (Eng. trans. 1972) of the Coptic *Gospel of Thomas.* For a synopsis, see BURTON H. THROCKMORTON, JR. (ed.), *Gospel Parallels: A Synopsis of the First Three Gospels,* 3rd ed. (1967). For a general study of the Gospels and the Synoptic problem, see FREDERICK C. GRANT, *The Gospels: Their Origin and Their Growth* (1957). For arguments against the priority of Mark, see WILLIAM R. FARMER, *The Synoptic Problem* (1964). Significant new approaches to gospel study are found in JAMES M. ROBINSON and HELMUT KOESTER, *Trajectories Through Early Christianity* (1971). For form criticism, see RUDOLF BULTMANN, *Die Geschichte der synoptischen Tradition,* 3rd ed. (1958; Eng. trans., *The History of the Synoptic Tradition,* 1963). AMOS N. WILDER, *Early Christian Rhetoric* (1971), goes beyond form criticism by fuller attention to modern literary criticism. (*Mark*): ROBERT H. LIGHTFOOT, *The Gospel Message of St. Mark* (1950); and WILLI MARXSEN, *Der Evangelist Markus* (1959; Eng. trans., *Mark the Evangelist,* 1969), are two outstanding works representing different periods and methods of scholarship. (*Matthew*): For discussion of the arrangement, Old Testament citations, and theology of Matthew, see GUENTHER BORNKAMM, GERHARD BARTH, and HEINZ J. HELD, *Auslegung im Mattäusevangelium* (1960; Eng. trans., *Tradition and Interpretation in Matthew,* 1963). DAVID HILL, *New Testament Prophesy* (1980), a discussion of prophesy both in the Bible and in the church today. KRISTER STENDAHL, "Prayer and Forgiveness," in *Svensk Exegetisk Årsbok,* 22–23:75–86 (1957–58), in English; and *The School of St. Matthew and Its Use of the Old Testament,* 2nd ed. (1968). (*Luke*): HENRY J. CADBURY, *The Making of Luke-Acts,* 2nd ed. (1958); and HANS CONZELMANN, *Die Mitte der Zeit: Studien zur Theologie des Lukas,* 3rd ed. (1960; Eng. trans., *The Theology of St. Luke,* 1960), represent a classic treatment of Luke-Acts. (*John*): Among the most important recent studies on John are CHARLES H. DODD, *Historical Tradition in the Fourth Gospel* (1963), and *The Interpretation of the Fourth Gospel* (1953); ERNST KAESEMANN, *Jesu letzter Wille nach Johannes 17.* (1966; Eng. trans., *The Testament of Jesus: A Study of the Gospel of John in the Light of Chapter 17,* 1968); and JAMES L. MARTYN, *History and Theology in the Fourth Gospel* (1968). (*Acts*): See also *Luke* above. For Acts viewed in its own time, see HENRY J. CADBURY, *The Book of Acts in History* (1955). Literary style and methods of composition are discussed in MARTIN DIBELIUS, *Aufsätze zur Apostelgeschichte* (1951; Eng. trans., *Studies in the Acts of the Apostles,* 1956). The scope and purpose of Acts are treated in P.M. MENOUD, "Le Plan des Actes des Apôtres," *New Testament Studies,* 1:44–50 (1954–55); and W.C. VAN UNNIK, "The 'Book of Acts' the Confirmation of the Gospel," *Novum Testamentum,* 4:26–59 (1960). (*Paul*): For general works on Paul and the epistles, see GUENTHER BORNKAMM, *Early Christian Experience* (1970), and *Paulus* (1969; Eng. trans., 1971); WILLIAM D. DAVIES, *Paul and Rabbinic Judaism,* 2nd ed. (1955); MARTIN DIBELIUS and WERNER G. KUEMMEL, *Paulus* (1951; Eng. trans., 1953); JOHANNES MUNCK, *Paulus und die Heilsgeschichte* (1954; Eng. trans., *Paul and the Salvation of Mankind,* 1959); ARTHUR D. NOCK, *St. Paul* (1938); and HANS J. SCHOEPS, *Paulus: Die Theologie des Apostels...* (1959; Eng. trans., *Paul: The Theology of the Apostle...,* 1961). See also KRISTER STENDAHL, "The Apostle Paul and the Introspective Conscience of the West," *Harvard Theological Review,* 51: 199–215 (1963). For a survey of Pauline studies, see EDWARD E. ELLIS, *Paul and His Recent Interpreters* (1961); WAYNE A. MEEKS (ed.), *The Writings of St. Paul* (1972); and ERNST KAESEMANN, *Paulinische Perspektiven* (1969; Eng. trans., *Perspectives on Paul,* 1971). (*Romans*): JOHN KNOX, "A Note on the Text of Romans," *New Testament Studies,* 2: 191–192 (1955–56); KRISTER STENDAHL, "Hate, Non-Retaliation, and Love: 1QS x, 17–20 and Romans 12:19–21," *Harvard Theological Review,* 50:343–355 (1962). (*I Corinthians*): For a discussion of the heresies met in I Corinthians, see WALTER SCHMITHALS, *Die Gnosis in Korinth,* 3rd ed. (1969; Eng. trans., *Gnosticism in Corinth,* 1971). (*II Corinthians*): For the arrangement of the fragments of II Corinthians and their redaction, see GUENTHER BORNKAMM, "The History of the Origin of the So-Called 2nd Letter to the Corinthians," *New Testament Studies,* 8:258–264 (1961–62). For a discussion of Paul's opponents in II Corinthians, see DIETER GEORGI, *Die Gegner des Paulus im 2. Korintherbrief: Studien zur religiösen Propaganda in der Spätantike* (1964); and his shorter article on this subject, "Forms of Religious Propaganda," in HANS J. SCHULTZ (ed.), *Die Zeit Jesu* (1966; Eng. trans., *Jesus in His Time,* 1971). (*Galatians*): For a discussion of the heretics in Galatia, see WALTER SCHMITHALS, "Die Häretiker in Galatien," *Zeitschrift für die Neutestamentliche Wissenschaft und die Kunde der Aelteren Kirche* (ZNW), pp. 25–67 (1956). (*Ephesians*): For the meaning and goal of Ephesians, see EDGAR J. GOODSPEED, *The Meaning of Ephesians* (1933), and *The Key to Ephesians* (1956). See also C. LESLIE MITTON, *The Epistle to the Ephesians* (1951). (*Philippians*): For the place of Philippians in the Pauline collection and the meaning of its various sections, see HELMUT KOESTER, "The Purpose of the Polemic of a Pauline Fragment (Philippians III)," *New Testament Studies,* 8:317–332 (1961–62). For the concept of Philippians as a testament, see DIETER GEORGI, "Ein Testament des Paulus (Phil. 3, 2ff.)," *ZNW* (1972). (*Philemon*): JOHN KNOX, *Philemon Among the Letters of Paul,* 2nd ed. (1959). (*Pastoral Epistles*): For evidence against Pauline authorship, see PERCY N. HARRISON, *The Problem of the Pastoral Epistles* (1921). See also EDUARD SCHWEIZER, *Church Order in the New Testament* (1961). (*Hebrews*): Concerning the Christology of Hebrews and the idea of the "wandering people of God," see ERNST KAESEMANN, *Das wandernde Gottesvolk,* 3rd ed. (1959). An approach to the eschatology of Hebrews and an origin connected with followers of Stephen is found in WILLIAM MANSON, *The Epistle to the Hebrews* (1951). (*Catholic Epistles*): For the typical admixture of parenesis, apocalyptic, and the general address of the Catholic Epistles, see CARL ANDRESEN, "Zum Formular frühchristlicher Gemeindebriefe," *ZNW,* 56:233–259 (1965). The similarity of style of the Catholic Epistles to later Christian Greek literature is treated in A. WIFSTRAND, "Stylistic Problems in the Epistles of James and Peter," *Studia Theologica,* 11:35–60 (1948). (*James*): For a solution to the apparent contradiction of Paul and James concerning "works," see JOACHIM JEREMIAS, "Paul and James," *Expository Times,* 66:368–371 (1954–55); for clarification of special passages with a modern technique similar to rabbinic methodology, see ROY B. WARD, "The Works of Abraham: James 2:14–26," *Harvard Theological Review,* 61:283–290 (1968), and "Partiality in the Assembly: James 2:2–4," *ibid.,* 62:87–97 (1969). (*I Peter*): For a date in Trajan's time, see JOHN KNOX,

"Pliny and I Peter: A Note on I Pet. 4:14–16 and 3:15," *Journal of Biblical Literature,* 72:187–189 (1953); an interpretation of the Descent into Hell is found in B. REICKE, *The Disobedient Spirits and Christian Baptism: A Study of I Peter iii, 19 and Its Context* (1946). (*II Peter and Jude*): For motivation for the writing of II Peter, see ERNST KAESEMANN, "An Apologia for Primitive Christian Eschatology," in *Essays on New Testament Themes* (1964). (*Johannine Epistles*): For speculations as to authorship, date, and nature of the situation of the Johannine Epistles, see W.F. HOWARD, "The Common Authorship of the Johannine Gospel and Epistles," *Journal of Theological Studies* (1947). (*Revelation*): Concerning liturgical style and content in Revelations, see GUENTHER BORNKAMM, "On the Understanding of Worship; B," in *Early Christian Experience* (1969). For a study of Revelation as a creative revelatory poem with unity throughout, drawing upon apocalyptic imagery of its time, see AUSTIN M. FARRER, *A Rebirth of Images* (1949, reprinted 1963). A general survey of apocalypticism and apocalypses from 200 BC into the early Christian era is found in DAVID S. RUSSELL (*op. cit.*).

Apocrypha: EDGAR HENNECKE, *Neutestamentliche Apokryphen in deutscher Übersetzung* (1959; Eng. trans., *New Testament Apocrypha,* 2 vol., 1963–65), a standard work; MONTAGUE R. JAMES, *The Apocryphal New Testament* (1924, reprinted 1955), convenient but obsolete; R.M. GRANT, D.N. FREEDMAN, and W.R. SCHOEDEL, *The Secret Sayings of Jesus* (1960); B. PICK, *The Apocryphal Acts of Paul, Peter, John, Andrew and Thomas* (1909); for Greek texts, see R.A. LIPSIUS and M. BONNET, *Acta Apostolorum Apocryphe.* A Greek papyrus (late 3rd century) of the Acts of Paul was edited by C. SCHMIDT in *Praxeis Paulou* (1936); he notes other papyrus fragments. The Seneca letters were edited by C.W. BARLOW, *Epistolae Seneca ad Paulum et Pauli ad Senecam* (*quae vocantur*) (1938).

Biblical literature in liturgy: ABRAHAM Z. IDELSOHN, *The Jewish Liturgy and Its Development* (1932, reprinted 1967); JOSEPH H. HERTZ, *The Authorized Daily Prayer Book,* rev. ed. (1948), Hebrew and English with historical notes and commentary; FAN S. NOLI, *Three Liturgies of the Eastern Orthodox Church* (1955); DONALD ATTWATER, *Eastern Catholic Worship* (1945), eight Uniate liturgies in English; JOSEF A. JUNGMANN, *Missarum Sollemnia: Eine Genetische Erklärung der römischen Messe,* 2 vol. (1958; Eng. trans., *The Mass of The Roman Rite: Its Origins and Development,* abridged ed., 1959); CLEMENT J. MCNASPY, *Our Changing Liturgy* (1966), reforms following Vatican II; GREGORY DIX, *The Shape of the Liturgy* (1945); BARD THOMPSON, *Liturgies of the Western Church* (1961), includes the main Protestant traditions.

Biblical exegesis and hermeneutics. *The Cambridge History of the Bible,* 3 vol. (1963–70), includes contributions by specialists on biblical interpretation from pre-Christian times to the present day. J. BARR, *Old and New in Interpretation* (1966), discusses the relation between the Old and New Testaments and examines critically some of the interpretative principles favoured by exegetes and theologians; another work on this subject is E.C. BLACKMAN, *Biblical Interpretation* (1957). C.E. BRAATEN, *History and Hermeneutics* (1966), discusses the relevance of the historical-critical method to theological study and the idea of revelation through history; F.F. BRUCE, *Biblical Exegesis in the Qumran Texts* (1959), examines the interpretative principles followed by biblical commentaries and other documents among the Dead Sea Scrolls. The major work on the theme of salvation-history in the Bible is O. CULLMANN, *Salvation in History* (1967). C.H. DODD, *According to the Scriptures* (1952), shows the various ways in which the Christian interpretation of important areas of the Old Testament provided the substructure of New Testament theology. F.W. FARRAR, *History of Interpretation* (1886, reprinted 1961), provides a classical survey of biblical exegesis from the early rabbinical period to the 19th century; R.M. GRANT, *A Short History of the Interpretation of the Bible,* rev. ed. (1963), is probably the best work of its kind. B. LINDARS, *New Testament Apologetic* (1962), studies the Old Testament quotations in the New Testament as evidence, in their text and interpretation, for the developing life and thought of the primitive church. J.M. ROBINSON and J.B. COBB (eds.), *The New Hermeneutic* (1964), expounds modern hermeneutical concerns. B. SMALLEY, *The Study of the Bible in the Middle Ages,* 2nd rev. ed. (1952), remains the standard work on early medieval exegesis. G. VERMES, *Scripture and Tradition in Judaism* (1961), gives an account of the interaction of the written text and oral tradition in Jewish exegesis of the pre-Christian and early rabbinical age. An outline of the history of biblical interpretation and of the main exegetical trends of the mid-20th century is presented in J.D. WOOD, *The Interpretation of the Bible* (1958); A. RICHARDSON and W. SCHWEITZER (eds.), *Biblical Authority for Today* (1951), discusses the difficulties of applying biblical ethics to some of the most urgent concerns of the modern world.

(J.C.Ry./L.F./R.F./N.M.Sa./
H.G.D./D.Fl./K.St./E.T.Sa./R.M.G.)

Biochemical Components of Organisms

Every living system contains, besides water and minerals, a large number of organic compounds. In general, the majority of these compounds may be classified as proteins, carbohydrates, and lipids, or fats. Nucleic acids and certain other organic derivatives are also important constituents. Various substances that cannot be classified in the above categories occur as well, though usually not in appreciable amounts. Such substances include vitamins and hormones of the nonprotein variety, which are biologically effective in relatively small quantities.

Proteins are fundamental to life. Some serve as structural material, comprising a major constituent of cellular membranes and the principal component of skin, hair, horn, and feathers. Other proteins, such as antibodies, provide a defense against invading destructive forces. Still others, the enzymes, are essential biocatalysts that accelerate thousands of complex chemical reactions necessary for sustaining life. Carbohydrates, which include simple sugars, starch, and cellulose, provide a major energy source for organisms as well as vital structural components. Most of the lipids produced by cells serve the same functions. Moreover, the function of some enzymes is dependent on their attachment to lecithin and certain other lipids. The two nucleic acids, deoxyribonucleic acid (DNA) and ribonucleic acid (RNA), play an integral part in the synthesis of proteins and in the transmission of hereditary information from one generation to the next.

In most organisms all water-soluble vitamins, with the exception of vitamin C (ascorbic acid), are converted to coenzymes of enzymes that function in energy transfer or in the metabolism of proteins, carbohydrates, and lipids. Certain kinds of fat-soluble vitamins appear to contribute substantially to the structure of membranes or at least aid in maintaining their integrity. Some of them may also control the synthesis of various enzymes. Hormones, produced and secreted by endrocrine glands, help to regulate the activities of other tissues and maintain homeostasis—*i.e.*, a relatively stable internal environment.

This article provides a summary of the structure, properties, and significance of each of the major classes of biochemical substances. It also discusses in brief methods of laboratory synthesis and analysis. Additional information pertaining to the nature and characteristics of many such organic compounds can be found in the article CHEMICAL COMPOUNDS. For details on specialized subjects in which these substances play an important role, see CELLS; TISSUES AND FLUIDS; METABOLISM; GENETICS AND HEREDITY: *The gene.* (Ed.)

This article is divided into the following sections:

PROTEINS

Proteins are highly complex substances that are present in all living organisms. They are of great nutritional value and are directly involved in the chemical processes essential for life. The importance of proteins was recognized by the chemists in the early 19th century who coined the name for these substances from the Greek *proteios*, meaning "holding first place." Proteins are both species-specific and organ-specific; for instance, muscle proteins differ from those of the brain and liver.

A protein molecule is very large compared to molecules of sugar or salt and consists of many amino acids joined together to form long chains, much as beads are arranged on a string. There are about 20 different amino acids that occur naturally in proteins. Proteins of similar function have similar amino acid composition and sequence. Although it is not yet possible to explain all of the functions of a protein from its amino acid sequence, established correlations between structure and function can be attributed to the properties of the amino acids that compose proteins.

Amino acid production and requirements by plants and animals

Plants can synthesize all of the amino acids; animals cannot, even though all of them are essential for life. Plants can grow in a medium containing inorganic nutrients that provide nitrogen, potassium, and other substances essential for growth. They utilize the carbon dioxide in the air during the process of photosynthesis to form organic compounds such as carbohydrates. Animals, however, must obtain organic nutrients from outside sources. Because the protein content of most plants is low, very large amounts of plant material are required by animals, such as ruminants (*e.g.*, cows), that eat only plant material to meet their amino acid requirements. Nonruminant animals, including man, obtain proteins principally from animals and their products—*e.g.*, meat, milk, and eggs. The seeds of legumes are increasingly being used to prepare inexpensive protein-rich food (see NUTRITION: *Human nutrition and diet*).

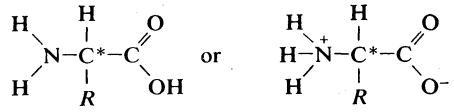
The protein content of animal organs is usually much higher than that of the blood plasma. Muscles, for example, contain about 30 percent protein, the liver 20 to 30 percent, and red blood cells 30 percent. Higher percentages of protein are found in hair, bones, and other organs and tissues with a low water content. The quantity of free amino acids and peptides in animals is much smaller than the amount of protein. Evidently, protein molecules are produced in cells by the stepwise alignment of amino acids and are released into the body fluids only after synthesis is complete.

The high protein content of some organs does not mean that the importance of proteins is related to their amount in an organism or tissue; some of the most important proteins, such as enzymes and hormones, occur in extremely small amounts. The importance of proteins is related principally to their function. All enzymes identified thus far are proteins. Enzymes, which are the catalysts of all metabolic reactions, enable an organism to build up the chemical substances necessary for life—proteins, nucleic acids, carbohydrates, and lipids—to convert them into other substances, and to degrade them. Life without enzymes is not possible. There are several protein hormones with important regulatory functions. In all vertebrates, the respiratory protein hemoglobin acts as oxygen carrier in the blood, transporting oxygen from the lung to body organs and tissues. A large group of structural proteins maintains and protects the structure of the animal body.

General structure and properties of proteins

THE AMINO ACID COMPOSITION OF PROTEINS

The common property of all proteins is that they consist of long chains of α-amino (alpha amino) acids. The general structure of α-amino acids is shown in Formula 1.



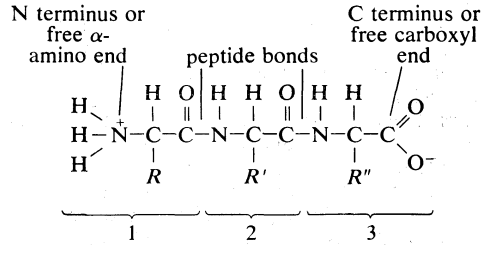
Formula 1: Generalized structure of all α-amino acids. C stands for a carbon atom; C* stands for the α-carbon. H is hydrogen, O is oxygen, and N is nitrogen. *R* is a general term for any of several different chemical structures that range from one hydrogen atom to large and complex molecular units containing many different atoms. The + and − signs represent electrical charges that exist when the molecule takes the configuration shown at the right.

The α-amino acids are so called because the α-carbon atom in the molecule (shown by an asterisk [*] in Formula 1) carries an amino group ($-NH_2$); the α-carbon atom also carries a carboxyl group (−COOH). In acidic solutions, when the pH is less than 4, the −COO groups combine with hydrogen ions (H^+) and are thus converted into the uncharged form (−COOH). In alkaline solutions, at pH above 9, the ammonium groups ($-NH^+_3$) lose a hydrogen ion and are converted into amino groups ($-NH_2$). In the pH range between 4 and 8, the amino acids exist almost exclusively in the structure shown at the right side of Formula 1. Because in this form they carry both a positive and a negative charge, they do not migrate in an electrical field. Such structures have been designated as dipolar ions, or zwitterions (*i.e.*, hybrid ions).

Although more than 100 amino acids occur in nature, particularly in plants, only 20 types are commonly found in most proteins (see Figure 1). In protein molecules the α-amino acids are linked to each other by peptide bonds between the amino group of one amino acid and the carboxyl group of its neighbour; the structure of the peptide bond is given in Formula 2. The condensation

Formula 2: The peptide bond.

(joining) of three amino acids yields the tripeptide shown in Formula 3.



Formula 3: A tripeptide. *R'* and *R''* represent the possibility that the three *R* groups (side chains) could be different.

Peptide structure

It is customary to write the structure of peptides in such a way that the free α-amino group (also called the N terminus of the peptide) is at the left side and the free carboxyl group (the C terminus) at the right side. Proteins are macromolecular polypeptides—*i.e.*, very large molecules composed of many peptide-bonded amino acids. Most of the common ones contain more than 100 amino acids linked to each other in a long peptide chain. The average molecular weight (based on the weight of a hydrogen atom as 1) of each amino acid is approximately 100 to 125; thus, the molecular weights of proteins are usually in the range of 10,000 to 100,000 daltons (one dalton is the weight of one hydrogen atom). The species-specificity and organ-specificity of proteins result from differences in the number and sequences of amino acids. Twenty different amino acids in a chain 100 amino acids long can be arranged in far more than 10^{100} ways (10^{100} is the number one followed by 100 zeroes).

Structures of common amino acids. The amino acids present in proteins differ from each other in the structure of their side (*R*) chains. The simplest amino acid is glycine, in which *R* is a hydrogen atom (see Figure 1). In a number of amino acids, *R* represents straight or branched carbon chains. One of these amino acids is alanine, in which *R* is the methyl group ($-CH_3$). Valine, leucine, and isoleucine, with longer *R* groups, complete the alkyl side-chain series. The alkyl side chains (*R* groups) of these amino acids are nonpolar; this means that they have no affinity for water but some affinity for each other. Although plants can form all of the alkyl amino acids, animals can synthesize only alanine and glycine; thus valine, leucine, and isoleucine must be supplied in the diet.

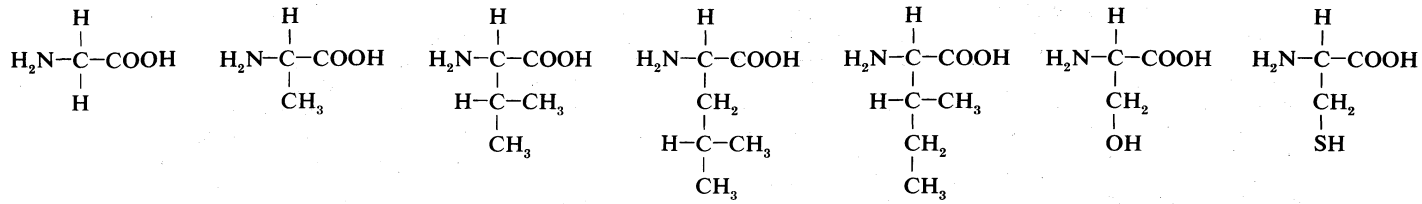
Two amino acids, each containing three carbon atoms, are derived from alanine; they are serine and cysteine. Serine contains an alcohol group ($-CH_2OH$) instead of the methyl group of alanine, and cysteine contains a mercapto group ($-CH_2SH$). Animals can synthesize serine but not cysteine or cystine. Cysteine occurs in proteins predominantly in its oxidized form (oxidation in this sense meaning the removal of hydrogen atoms), called cystine. Cystine consists of two cysteine molecules linked by the disulfide bond (−S−S−) that results when a hydrogen atom is removed from the mercapto group of each of the cysteines (see Figure 1). Disulfide bonds are important in protein structure because they allow the linkage of two different parts of a protein molecule to—and thus the formation of loops in—the otherwise straight chains. Some proteins contain small amounts of cysteine with free sulfhydryl (−SH) groups.

Four amino acids, each consisting of four carbon atoms, occur in proteins; they are aspartic acid, asparagine, threonine, and methionine. Aspartic acid and asparagine, which occur in large amounts, can be synthesized by animals. Threonine and methionine cannot be synthesized and thus are essential amino acids—*i.e.*, they must be supplied in the diet. Most proteins contain only small amounts of methionine.

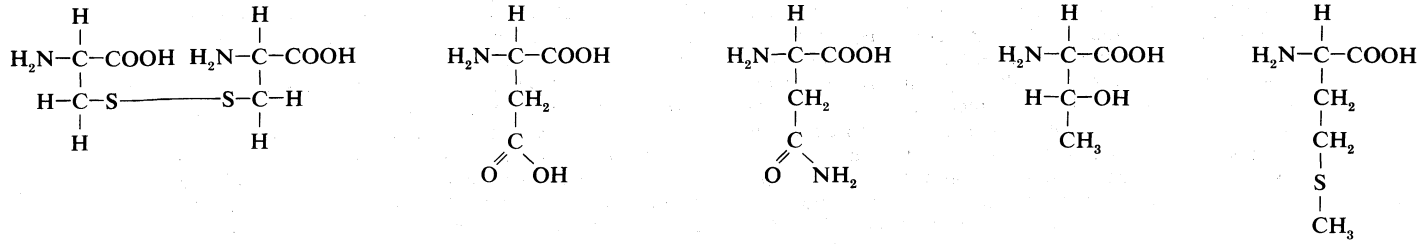
Proteins also contain an amino acid with five carbon atoms (glutamic acid) and an imino acid (proline), which is a structure with the amino group ($-NH_2$) bonded to the alkyl side chain, forming a ring. Glutamic acid and aspartic acid are dicarboxylic acids—that is, they have two carboxyl groups (−COOH). Glutamine is similar to asparagine in that both are the amides of their corresponding dicarboxylic acid forms; *i.e.*, they have an amide group ($-CONH_2$) in place of the carboxyl (−COOH) of the side chain (see Figure 1). Glutamic acid and glutamine are abundant in most proteins; *e.g.*, in plant proteins they sometimes comprise more than one third of the amino acids present. Both glutamic acid and glutamine can be synthesized by animals. The imino acids proline and hydroxyproline occur in large amounts in collagen, the protein of the connective tissue of animals (see Table 1). Proline and hydroxyproline lack free amino ($-NH_2$) groups because the amino group is enclosed in a ring structure with the side chain; they thus cannot exist in a zwitterion form. Although the imino group (>NH) of these amino acids can form a peptide bond with the carboxyl group of another amino acid, the bond so formed gives rise to a kink in the peptide chain—*i.e.*, the imino ring structure alters the regular bond angle of normal peptide bonds.

Most abundant amino acids

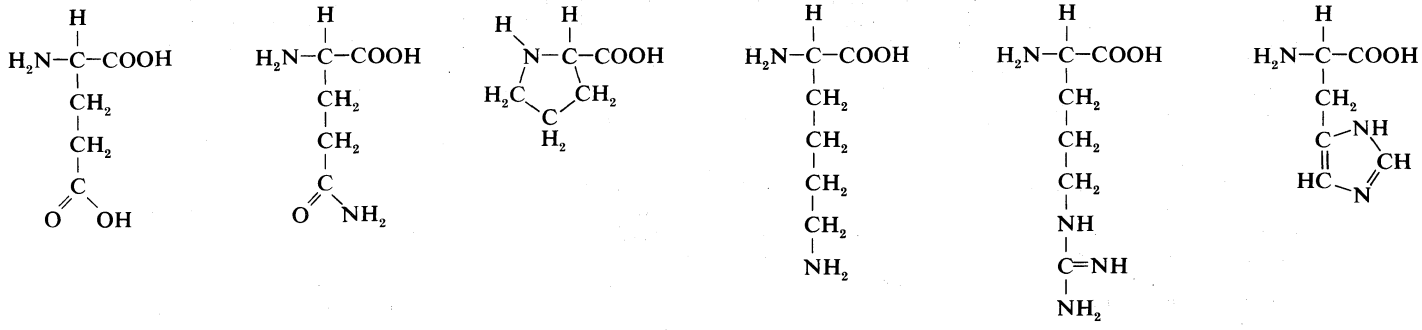
Proteins usually are almost neutral molecules; that is, they have neither acidic nor basic properties. This means that the acidic carboxyl ($-COO^-$) groups of aspartic and glutamic acid are about equal in number to the amino acids with basic side chains. Three such basic amino acids, each containing six carbon atoms, occur in proteins. The one with the simplest structure, lysine, is synthesized by plants but not by animals. Even some plants have a low lysine content. Arginine is found in all proteins; it occurs in particularly high amounts in the strongly basic protamines (simple proteins composed of relatively few amino acids) of fish sperm. The third basic amino acid is



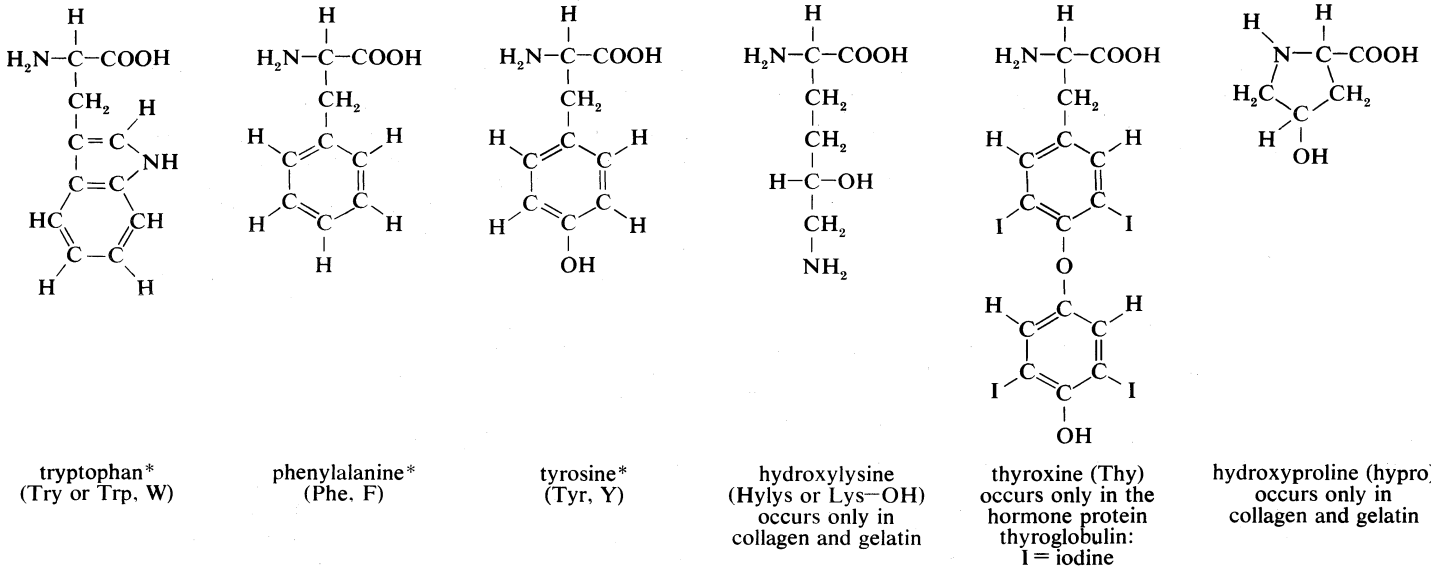
glycine (Gly, G) | alanine (Ala, A) | valine* (Val, V) | leucine* (Leu, L) | isoleucine* (Ile, I) | serine (Ser, S) | cysteine (CysH, C)



cystine (Cys-S-S-Cys, C-C) | aspartic acid (Asp, D; Asx or B) | asparagine ($AspNH_2$ or Asn, N; Asx or B) | threonine* (Thr, T) | methionine* (Met, M)



glutamic acid (Glu, E; Glx or Z) | glutamine ($GluNH_2$, GluN, or Gln, Q; Glx or Z) | proline (Pro, P) an imino acid | lysine* (Lys, K) | arginine (Arg, R) | histidine (His, H)



tryptophan* (Try or Trp, W) | phenylalanine* (Phe, F) | tyrosine* (Tyr, Y) | hydroxylysine (Hylys or Lys—OH) occurs only in collagen and gelatin | thyroxine (Thy) occurs only in the hormone protein thyroglobulin: I = iodine | hydroxyproline (hypro) occurs only in collagen and gelatin

Figure 1: Structures of amino acids found in proteins. Those amino acids marked with an asterisk (*) must be supplied in the diet of animals, which cannot synthesize them. The abbreviations in parentheses represent the shorthand notations (in three-letter codes and one-letter codes) used when indicating protein structures. The one-letter symbol for an unknown amino acid is *X*.

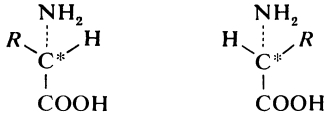
histidine. Both arginine and histidine can be synthesized by animals. Histidine is a weaker base than either lysine or arginine. The imidazole ring, a five-membered ring structure containing two nitrogen atoms in the side chain of histidine (see Figure 1), acts as a buffer (*i.e.*, a stabilizer of hydrogen ion concentration) by binding hydrogen ions (H^+) to the nitrogen atoms of the imidazole ring.

The remaining amino acids—phenylalanine, tyrosine, and tryptophan—have in common an aromatic structure; *i.e.*, a benzene ring is present (see Figure 1). Animals cannot synthesize the benzene ring, and these three amino acids are essential ones; but animals can convert phenylalanine to tyrosine. Because these amino acids contain benzene rings, they can absorb ultraviolet light at wavelengths between 270 and 290 nanometres (nm; one nanometre = 10^{-9} metre = 10 angstrom units). Phenylalanine absorbs very little ultraviolet light; tyrosine and tryptophan, however, absorb it strongly and are responsible for the absorption band most proteins exhibit at 280–290 nanometres. This absorption is often used to determine the quantity of protein present in protein samples.

Most proteins contain only the amino acids described above; however, other amino acids occur in proteins in small amounts. Thyroglobulin, the hormone of the thyroid gland, for example, contains thyroxine, which is an iodine-containing compound derived from tyrosine. The collagen found in connective tissue contains, in addition to hydroxyproline, small amounts of hydroxylysine. Other proteins contain some monomethyl-, dimethyl-, or trimethyllysine—*i.e.*, lysine derivatives containing one, two, or three methyl groups ($-CH_3$). The amount of these unusual amino acids in proteins, however, rarely exceeds 1 or 2 percent of the total amino acids.

Physicochemical properties of the amino acids. The physicochemical properties of a protein are determined by the analogous properties of the amino acids in it.

The α-carbon atom of all amino acids, with the exception of glycine, is asymmetric; this means that four different chemical entities (atoms or groups of atoms) are attached to it. As a result, each of the amino acids, except glycine, can exist in two different spatial, or geometric, arrangements (*i.e.*, isomers), which are mirror images akin to right and left hands (see Formula 4). These isomers exhibit the property of optical rotation.



L-amino acid D-amino acid

Formula 4: The tetrahedral (four-faced) arrangement of the bonds around the α-carbon (C*). The solid lines represent bonds that slant upward from the plane of the drawing (*i.e.*, toward the reader). The broken lines represent bonds that recede from the plane of the drawing (*i.e.*, away from the reader).

Optical rotation is the rotation of the plane of polarized light, which is composed of light waves that vibrate in one plane, or direction, only. Solutions of substances that rotate the plane of polarization are said to be optically active, and the degree of rotation is called the optical rotation of the solution. The direction in which the light is rotated is generally designed as plus, or *d*, for dextrorotatory (to the right), or as minus, or *l*, for levorotatory (to the left). Some amino acids are dextrorotatory; others are levorotatory.

Configuration of amino acids

With the exception of a few small proteins (peptides) that occur in bacteria, the amino acids that occur in proteins have the configuration shown on the left of Formula 4. For this reason all the amino acids found in proteins are designed as L-amino acids.

In bacteria, D-alanine and some other D-amino acids have been found as components of gramicidin and bacitracin. These peptides are toxic to other bacteria and are used in medicine as antibiotics. The D-alanine has also been found in some peptides of bacterial membranes.

In contrast to most organic acids and amines, the amino acids are insoluble in organic solvents. In aqueous solutions they are dipolar ions (zwitterions, or hybrid ions) that react with strong acids or bases in a way that leads to the neutralization of the negatively or positively charged ends, respectively. Because of their reactions with strong acids and strong bases, the amino acids act as buffers—stabilizers of hydrogen ion (H^+) or hydroxide ion (OH^-) concentrations. In fact, glycine is frequently used as a buffer in the pH range from 1 to 3 (acid solutions) and from 9 to 12 (basic solutions). In acid solutions, glycine has a positive charge and therefore migrates to the cathode (negative electrode of a direct-current electrical circuit with terminals in the solution). Its charge, however, is negative in alkaline solutions, in which it migrates to the anode (positive electrode). At pH 6.1 glycine does not migrate, because each molecule has one positive and one negative charge. The pH at which an amino acid does not migrate in an electrical field is called the isoelectric point. Most of the monoamino acids (*i.e.*, those with only one amino group) have isoelectric points similar to that of glycine. The isoelectric points of aspartic and glutamic acids, however, are close to pH 3; and those of histidine, lysine, and arginine are at pH 7.6, 9.7, and 10.8, respectively.

Amino acid sequence in protein molecules. Since each protein molecule consists of a long chain of amino acid residues, linked to each other by peptide bonds, the hydrolytic cleavage of all peptide bonds is a prerequisite for the quantitative determination of the amino acid residues. Hydrolysis is most frequently accomplished by boiling the protein with concentrated hydrochloric acid. The quantitative determination of the amino acids is based on the discovery that amino acids can be separated from each other by chromatography on filter paper and made visible by spraying the paper with ninhydrin. The amino acids of the protein hydrolysate are separated from each other by passing the hydrolysate through a column of adsorbents which adsorb the amino acids with different affinities and, on washing the column with buffer solutions, release them in a definite order. The amount of each of the amino acids can be determined by the intensity of the colour reaction with ninhydrin.

To obtain information about the sequence of the amino acid residues in the protein, the protein is degraded stepwise, one amino acid being split off in each step. This is accomplished by coupling the free α-amino group ($-NH_2$) of the N-terminal amino acid with phenyl isothiocyanate; subsequent mild hydrolysis does not affect the peptide bonds; the procedure, called the Edman degradation, can be applied repeatedly; it thus reveals the sequence of the amino acids in the peptide chain.

Unavoidable small losses that occur during each step make it impossible to determine the sequence of more than about 30 to 50 amino acids by this procedure. For this reason the protein is usually first hydrolyzed by exposure to the enzyme trypsin (see below *Enzymes*), which cleaves only peptide bonds formed by the carboxyl groups of lysine and arginine. The Edman degradation is then applied to each of the few resulting peptides produced by the action of trypsin. Further information can be gained by hydrolyzing another portion of the protein with another enzyme, for instance with chymotrypsin, which splits predominantly peptide bonds formed by the amino acids tyrosine, phenylalanine, and tryptophan. The combination of results obtained with two or more different proteolytic (protein degrading) enzymes was first applied by the English biochemist Frederick Sanger, and it enabled him to elucidate the amino acid sequence of insulin. The amino acid sequences shown in formulas 7 to 11 and those of many other proteins have been determined in this manner.

LEVELS OF STRUCTURAL ORGANIZATION IN PROTEINS

Primary structure. Analytical and synthetic procedures reveal only the primary structure of the proteins—that is, the amino acid sequence of the peptide chains. They do not reveal information about the conformation (arrangement in space) of the peptide chain—that is, whether the peptide chain is present as a long straight thread or is irregularly coiled and folded into a globule. The configuration, or conformation, of a protein is determined by mutual attraction or repulsion of polar or nonpolar groups in the side chains (*R* groups) of the amino acids. The former have positive or negative charges in their side chains; the

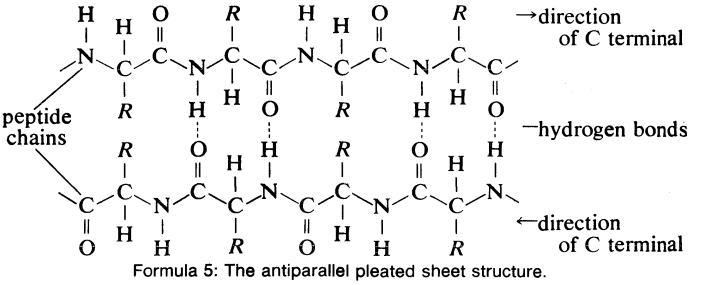
latter repel water but attract each other. Some parts of a peptide chain containing 100 to 200 amino acids may form a loop, or helix; others may be straight or form irregular coils.

Definitions of protein structural terms

The terms secondary, tertiary, and quaternary structure are frequently applied to the configuration of the peptide chain of a protein. A nomenclature committee of the International Union of Biochemistry (IUB) has defined these terms as follows: The primary structure of a protein is determined by its amino acid sequence without any regard for the arrangement of the peptide chain in space. The secondary structure is determined by the spatial arrangement of the main peptide chain without any regard for the conformation of side chains or other segments of the main chain. The tertiary structure is determined by both the side chains and other adjacent segments of the main chain, without regard for neighbouring peptide chains. Finally, the term quaternary structure is used for the arrangement of identical or different subunits of a large protein in which each subunit is a separate peptide chain.

Secondary structure. The nitrogen and carbon atoms of a peptide chain cannot lie on a straight line because of the magnitude of the bond angles between adjacent atoms of the chain; the bond angle is about 110°. Each of the nitrogen and carbon atoms can rotate to a certain extent, however, so that the chain has a limited flexibility. Because all of the amino acids, except glycine, are asymmetric L-amino acids, the peptide chain tends to assume an asymmetric helical shape; some of the fibrous proteins consist of elongated helices around a straight screw axis. Such structural features result from properties common to all peptide chains. The product of their effects is the secondary structure of the protein.

Tertiary structure. The tertiary structure is the product of the interaction between the side chains (*R*) of the amino acids composing the protein. Some of them contain positively or negatively charged groups, others are polar, and still others are nonpolar. The number of carbon atoms in the side chain varies from zero in glycine to nine in tryptophan (see Figure 1). Positively and negatively charged side chains have the tendency to attract each other; side chains with identical charges repel each other. The bonds formed by the forces between the negatively charged side chains of aspartic or glutamic acid on the one hand, and the positively charged side chains of lysine or arginine on the other hand, are called salt bridges. Mutual attraction of adjacent peptide chains also results from the formation of numerous hydrogen bonds. They are shown by dotted lines in the diagram of an antiparallel pleated sheet protein structure (see Formula 5). Hydrogen bonds form

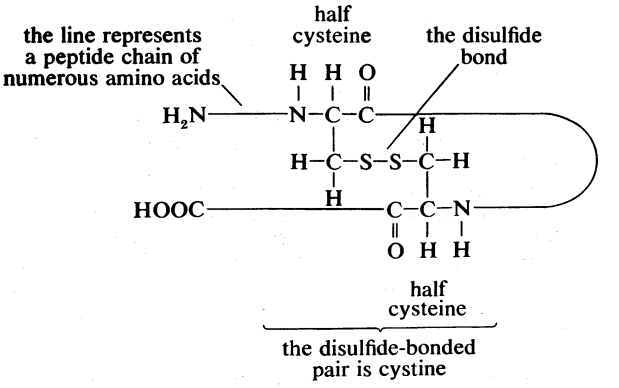


Formula 5: The antiparallel pleated sheet structure.

as a result of the attraction between the nitrogen-bound hydrogen atom (the imide hydrogen) and the unshared pair of electrons of the oxygen atom in the double bonded carbon–oxygen group (the carbonyl group) (>C=O). The result is a slight displacement of the imide hydrogen toward the oxygen atom of the carbonyl group. Although the hydrogen bond is much weaker than a covalent bond (*i.e.*, the type of bond between two carbon atoms, which equally share the pair of bonding electrons between them), the large number of imide and carbonyl groups in peptide chains results in the formation of numerous hydrogen bonds. Another type of attraction is that between nonpolar side chains of valine, leucine, isoleucine, and phenylalanine; the attraction results in the displacement of water molecules and is called hydrophobic interaction.

Hydrophobic interaction defined

In proteins rich in cystine, the conformation of the peptide chain is determined to a considerable extent by the disulfide bonds (—S—S—) of cystine. The halves of cystine may be located in different parts of the peptide chain and thus may form a loop closed by the disulfide bond, as shown in Formula 6. If the disulfide bond is reduced (*i.e.*, hydrogen is added) to two sulfhydryl (—SH) groups, the tertiary structure of the protein undergoes a drastic change—closed loops are broken and adjacent disulfide-bonded peptide chains separate.



Formula 6: The disulfide bridge between two cystine halves in an amino acid chain showing how loops in the chain are formed by this amino acid.

Quaternary structure. The nature of the quaternary structure is demonstrated by the structure of hemoglobin. Each molecule of human hemoglobin consists of four peptide chains, two α-chains and two β-chains; *i.e.*, it is a tetramer. The four subunits are linked to each other by hydrogen bonds and hydrophobic interaction. Because the four subunits are so closely linked, the hemoglobin tetramer is called a molecule, even though no covalent bonds occur between the peptide chains of the four subunits. In other proteins, the subunits are bound to each other by covalent bonds (disulfide bridges; see below the structure of insulin in Formula 8).

THE ISOLATION AND DETERMINATION OF PROTEINS

Animal material usually contains large amounts of protein and lipids and small amounts of carbohydrate; in plants, the bulk of the dry matter is usually carbohydrate. No general method exists for the isolation of proteins from organs or tissues. If it is necessary to determine the amount of protein in a mixture of animal foodstuffs, a sample is converted to ammonium salts by boiling with sulfuric acid and a suitable inorganic catalyst, such as copper sulfate (Kjeldahl method). The method is based on the assumption that proteins contain 16 percent nitrogen, and that nonprotein nitrogen is present in very small amounts. The assumption is justified for most tissues from higher animals but not for insects and crustaceans, in which a considerable portion of the body nitrogen is present in the form of chitin, a carbohydrate. Large amounts of nonprotein nitrogen are also found in the sap of many plants. In such cases, the precise quantitative analyses are made after the proteins have been separated from other biological compounds.

Concentration of protein molecules by dialysis

Proteins are sensitive to heat, acids, bases, organic solvents, and radiation exposure; for this reason, the chemical methods employed to purify organic compounds cannot be applied to proteins. Salts and molecules of small size are removed from protein solutions by dialysis; *i.e.*, by placing the solution into a sac of semipermeable material, such as cellulose or acetylcellulose, which will allow small molecules to pass through but not large protein molecules, and immersing the sac in water or a salt solution. Small molecules can also be removed either by passing the protein solution through a column of resin that adsorbs only the protein or by gel filtration. In gel filtration, the large protein molecules pass through the column, and the small molecules are adsorbed to the gel.

Groups of proteins are separated from each other by salting out—*i.e.*, the stepwise addition of sodium sulfate or ammonium sulfate to a protein solution. Some proteins,

called globulins, become insoluble and precipitate when the solution is half-saturated with ammonium sulfate or when its sodium sulfate content exceeds about 12 percent. Other proteins, the albumins, can be precipitated from the supernatant solution (*i.e.,* the solution remaining after a precipitation has taken place) by saturation with ammonium sulfate. Water-soluble proteins can be obtained in a dry state by freeze-drying (lyophilization), in which the protein solution is deep-frozen by lowering the temperature below −15° C (5° F) and removing the water; the protein is obtained as a dry powder.

Most proteins are insoluble in boiling water and are denatured by it—*i.e.,* irreversibly converted into an insoluble material. Heat denaturation cannot be used with connective tissue because the principal structural protein, collagen, is converted by boiling water into water-soluble gelatin.

Fractionation (separation into components) of a mixture of proteins of different molecular weight can be accomplished by gel filtration. The size of the proteins retained by the gel depends upon the properties of the gel. The proteins retained in the gel are removed from the column by solutions of a suitable concentration of salts and hydrogen ions.

Criteria of purity

Many proteins were originally obtained in crystalline form, but crystallinity is not proof of purity; many crystalline protein preparations contain other substances. Various tests are used to determine whether a protein preparation contains only one protein. The purity of a protein solution can be determined by such techniques as chromatography and gel filtration. In addition, a solution of pure protein will yield one peak when spun in a centrifuge at very high speeds (ultracentrifugation) and will migrate as a single band in electrophoresis (migration of the protein in an electrical field). After these methods and others (such as amino acid analysis) indicate that the protein solution is pure, it can be considered so. Because chromatography, ultracentrifugation, and electrophoresis cannot be applied to insoluble proteins, little is known about them; they may be mixtures of many similar proteins.

Very small (microheterogeneous) differences in some of the apparently pure proteins are known to occur; they are differences in the amino acid composition of otherwise identical proteins and are transmitted from generation to generation; *i.e.,* they are genetically determined; for example, some humans have two hemoglobins, hemoglobin A and hemoglobin S, which differ in one amino acid at a specific site in the molecule. In hemoglobin A the site is occupied by glutamic acid, and in hemoglobin S by valine. Refinement of the techniques of protein analysis has resulted in the discovery of other instances of "microheterogeneity."

The quantity of a pure protein can be determined by weighing or by measuring the ultraviolet absorbancy at 280 nanometres. The absorbency at 280 nanometres depends on the content of tyrosine and tryptophan in the protein (see above *The amino acid composition of proteins*). Sometimes the slightly less sensitive biuret reaction, a purple colour given by alkaline protein solutions upon the addition of copper sulfate, is used; its intensity depends only on the number of peptide bonds per gram, which is similar in all proteins.

PHYSICOCHEMICAL PROPERTIES OF PROTEINS

The molecular weight of proteins. The molecular weight of proteins cannot be determined by the methods of classical chemistry (*e.g.,* freezing-point depression) because they require solutions of a higher concentration of protein than can be prepared.

If a protein contains only one molecule of one of the amino acids or one atom of iron, copper, or another element, the minimum molecular weight of the protein or a subunit can be calculated; for example, the protein myoglobin contains 0.34 gram of iron in 100 grams of protein. The atomic weight of iron is 56; thus the minimum molecular weight of myoglobin is $(56 \times 100)/0.34 =$ about 16,500. Direct measurements of the molecular weight of myoglobin yield the same value. The molecular weight of hemoglobin, however, which also contains 0.34 percent iron, has been found to be 66,000 or $4 \times 16{,}500$; thus hemoglobin contains four atoms of iron.

Determination of molecular weight by ultracentrifugation

The method most frequently used to determine the molecular weight of proteins is ultracentrifugation; *i.e.,* spinning in a centrifuge at velocities up to about 60,000 revolutions per minute. Centrifugal forces of more than 200,000 times the gravitational force on the surface of the Earth are achieved at such velocities. The first ultracentrifuges, built in 1920, were used to determine the molecular weight of proteins. The molecular weights of a large number of proteins have been determined. Most consist of several subunits, the molecular weight of which is usually less than 100,000 and frequently ranges from 20,000 to 30,000. Proteins of very high molecular weights are found among hemocyanins, the copper-containing respiratory proteins of invertebrates; some range as high as several million. Although there is no definite lower limit for the molecular weight of proteins, short amino acid sequences are usually called peptides.

The shape of protein molecules. In the technique of X-ray diffraction the X-rays are allowed to strike a protein crystal; the X-rays, diffracted (bent) by the crystal, impinge on a photographic plate, forming a pattern of spots. This method reveals that peptide chains can assume very complicated, apparently irregular shapes. Two extremes in shape include the closely folded structure of the globular proteins and the elongated, unidimensional structure of the threadlike fibrous proteins; both were recognized many years before the technique of X-ray diffraction was developed. Solutions of fibrous proteins are extremely viscous (*i.e.,* sticky); those of the globular proteins have low viscosity (*i.e.,* they flow easily). A 5 percent solution of a globular protein—ovalbumin, for example—easily flows through a narrow glass tube; a 5 percent solution of gelatin, a fibrous protein, however, does not flow through the tube because it is liquid only at high temperatures and solidifies at room temperature. Even solutions containing only 1 or 2 percent of gelatin are highly viscous and flow through a narrow tube either very slowly or only under pressure. The elongated peptide chains of the fibrous proteins can be imagined to become entangled not only mechanically but also by mutual attraction of their side chains; in this way they incorporate large amounts of water. Most of the hydrophilic (water-attracting) groups of the globular proteins, however, lie on the surface of the molecules; as a result, globular proteins incorporate only a few water molecules. If a solution of a fibrous protein flows through a narrow tube, the elongated molecules become oriented parallel to the direction of the flow (see Figure 2), and the solution thus becomes birefringent like a crystal; *i.e.,* it splits a light ray into two components that travel at different velocities and are polarized at right angles to each other. Globular proteins do not show this phenomenon, which is called flow birefringence. Solutions of myosin, the contractile protein of muscles, show very high flow birefringence; other proteins with very high flow birefringence include solutions of fibrinogen, the clotting material of blood plasma, and solutions of tobacco mosaic virus. The gamma-globulins of the blood plasma show low flow birefringence; and none can be observed in solutions of serum albumin and ovalbumin.

Globular and fibrous proteins

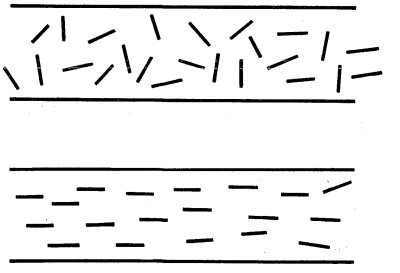


Figure 2: *Flow birefringence.*
The upper diagram shows a solution containing elongated, rodlike macromolecules that, in the resting solution, are randomly oriented. The lower diagram shows the same solution during flow through a horizontal tube.

Hydration of proteins. When dry proteins are exposed to air of high water content, they rapidly bind water up to a maximum quantity, which differs for different proteins; usually it is 10 to 20 percent of the weight of the protein. The hydrophilic groups of a protein are chiefly the positively charged groups in the side chains of lysine and arginine and the negatively charged groups of aspartic and glutamic acid. Hydration (*i.e.*, the binding of water) may also occur at the hydroxyl (−OH) groups of serine and threonine or at the amide ($-CONH_2$) groups of asparagine and glutamine.

The binding of water molecules to either charged or polar (partly charged) groups is explained by the dipolar structure of the water molecule; that is, the two positively charged hydrogen atoms form an angle of about 105°, with the negatively charged oxygen atom at the apex. The centre of the positive charges is located between the two hydrogen atoms; the centre of the negative charge of the oxygen atom is at the apex of the angle. The negative pole of the dipolar water molecule binds to positively charged groups; the positive pole binds negatively charged ones. The negative pole of the water molecule also binds to the hydroxyl and amino groups of the protein.

The water of hydration is essential to the structure of protein crystals; when they are completely dehydrated, the crystalline structure disintegrates. In some proteins this process is accompanied by denaturation and loss of the biological function.

In aqueous solutions, proteins bind some of the water molecules very firmly; others are either very loosely bound or form islands of water molecules between loops of folded peptide chains. Because the water molecules in such an island are thought to be oriented as in ice, which is crystalline water, the islands of water in proteins are called icebergs. Water molecules may also form bridges between the carbonyl (>C=O) and imino (>NH) groups of adjacent peptide chains, resulting in structures similar to those of the pleated sheet (see Formula 5) but with a water molecule in the position of the hydrogen bonds of that configuration. The extent of hydration of protein molecules in aqueous solutions is important, because some of the methods used to determine the molecular weight of proteins yield the molecular weight of the hydrated protein. The amount of water bound to one gram of a globular protein in solution varies from 0.2 to 0.5 gram. Much larger amounts of water are mechanically immobilized between the elongated peptide chains of fibrous proteins; for example, one gram of gelatin can immobilize at room temperature 25 to 30 grams of water.

The salting-out process

Hydration of proteins is necessary for their solubility in water. If the water of hydration of a protein dissolved in water is reduced by the addition of a salt such as ammonium sulfate, the protein is no longer soluble and is salted out, or precipitated. The salting-out process is reversible because the protein is not denatured (*i.e.*, irreversibly converted to an insoluble material) by the addition of such salts as sodium chloride, sodium sulfate, or ammonium sulfate. Some globulins, called euglobulins, are insoluble in water in the absence of salts; their insolubility is attributed to the mutual interaction of polar groups on the surface of adjacent molecules, a process that results in the formation of large aggregates of molecules. Addition of small amounts of salt causes the euglobulins to become soluble. This process, called salting in, results from a combination between anions (negatively charged ions) and cations (positively charged ions) of the salt and positively and negatively charged side chains of the euglobulins. The combination prevents the aggregation of euglobulin molecules by preventing the formation of salt bridges between them. The addition of more sodium or ammonium sulfate causes the euglobulins to salt out again and to precipitate.

Electrochemistry of proteins. Because the α-amino group and α-carboxyl group of amino acids are converted into peptide bonds (see Formula 2) in the protein molecule, there is only one α-amino group (at the N terminus) and one α-carboxyl group (at the C terminus). The electrochemical character of a protein is affected very little by these two groups. Of importance, however, are the numerous positively charged ammonium groups ($-NH_3^+$) of lysine and arginine and the negatively charged carboxyl groups ($-COO^-$) of aspartic acid and glutamic acid. In most proteins, the number of positively and negatively charged groups varies from 10 to 20 per 100 amino acids.

Electrometric titration. When measured volumes of hydrochloric acid are added to a solution of protein in salt-free water, the pH decreases in proportion to the amount of hydrogen ions added until it is about 4. Further addition of acid causes much less decrease in pH because the protein acts as a buffer at pH values of 3 to 4. The reaction that takes place in this pH range is the protonation of the carboxyl group—*i.e.*, the conversion of $-COO^-$ into −COOH. Electrometric titration of an isoelectric protein with potassium hydroxide causes a very slow increase in pH and a weak buffering action of the protein at pH 7; a very strong buffering action occurs in the pH range from 9 to 10 (see Figure 3). The buffering action at pH 7, which is caused by loss of protons (positively charged hydrogen) from the imidazolium groups (*i.e.*, the five-member ring structure in the side chain; see Figure 1) of histidine, is weak because the histidine content of proteins is usually low. The much stronger buffering action at pH values from 9 to 10 is caused by the loss of protons from the hydroxyl group of tyrosine and from the ammonium groups of lysine. Finally, protons are lost from the guanidinium groups (*i.e.*, the nitrogen-containing terminal portion of the arginine side chains; see Figure 1) of arginine at pH 12. A curve of the electrometric titration of glycine is shown in Figure 3. Electrometric titrations of proteins

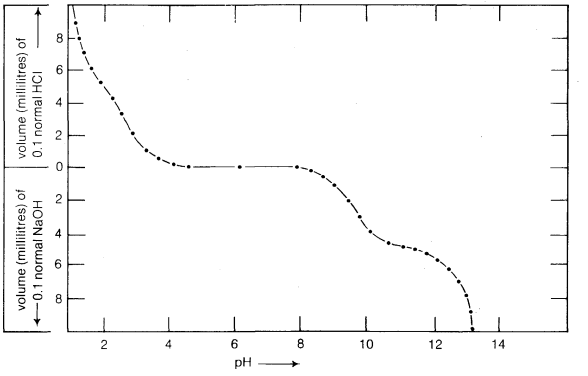


Figure 3: Electrometric titration of glycine. Addition of measured quantities of known-concentration hydrochloric acid (HCl) to glycine is shown in the upper half of the diagram; sodium hydroxide (NaOH) in the lower half. The dots indicate the experimental results. The addition of a trace of acid or base to pure glycine, the isoelectric point of which is close to pH 6.1, causes a strong change in pH. Glycine acts as a buffer, however, in the acidic pH range below 3 and in the alkaline pH range above 8.

yield similar curves. Electrometric titration makes possible the determination of the approximate number of carboxyl groups, ammonium groups, histidines, and tyrosines per molecule of protein.

Electrophoresis. The positively and negatively charged side chains of proteins cause them to behave like amino acids in an electrical field; that is, they migrate during electrophoresis at low pH values to the cathode (negative terminal) and at high pH values to the anode (positive terminal). The isoelectric point, the pH value at which the protein molecule does not migrate, is in the range of pH 5 to 7 for many proteins. Proteins such as lysozyme, cytochrome *c*, histone, and others rich in lysine and arginine (see Table 2), however, have isoelectric points in the pH range between 8 and 10. The isoelectric point of pepsin, which contains very few basic amino acids, is close to 1.

Zone electrophoresis

Free-boundary electrophoresis, the original method of determining electrophoretic migration, has been replaced in many instances by zone electrophoresis, in which the protein is placed in either a gel of starch, agar, or polyacrylamide or in a porous medium such as paper or cellulose acetate. The migration of hemoglobin and other coloured proteins can be followed visually. Colourless proteins are made visible after the completion of electrophoresis by staining them with a suitable dye.

CONFORMATION OF GLOBULAR PROTEINS

Results of X-ray diffraction studies. Most knowledge concerning secondary and tertiary structure of globular proteins has been obtained by the examination of their crystals using X-ray diffraction. In this technique X-rays are allowed to strike the crystal; the X-rays are diffracted by the crystal and impinge on a photographic plate, forming a pattern of spots. The measured intensity of the diffraction pattern, as recorded on a photographic film, depends particularly on the electron density of the atoms in the protein crystal. This density is lowest in hydrogen atoms, and they do not give a visible diffraction pattern. Although carbon, oxygen, and nitrogen atoms yield visible diffraction patterns, they are present in such great number—about 700 or 800 per 100 amino acids—that the resolution of the structure of a protein containing more than 100 amino acids is almost impossible. Resolution is considerably improved by substituting into the side chains of certain amino acids very heavy atoms, particularly those of heavy metals. Mercury ions, for example, bind to the sulfhydryl (−SH) groups of cysteine. Platinum chloride has been used in other proteins. In the iron-containing proteins, the iron atom already in the molecule is adequate.

Although the X-ray diffraction technique cannot resolve the complete three-dimensional conformation (that is, the secondary and tertiary structure of the peptide chain), complete resolution has been obtained by combination of the results of X-ray diffraction with those of amino acid sequence analysis. In this way the complete conformation of such proteins as myoglobin, chymotrypsinogen, lysozyme, and ribonuclease has been resolved.

The X-ray diffraction method has revealed regular structural arrangements in proteins; one is an extended form of antiparallel peptide chains that are linked to each other by hydrogen bonds between the carbonyl (≥C=O) and imino (≥NH) groups (shown in Formula 5). This conformation, called the pleated sheet, or β-structure, is found in some fibrous proteins. Short strands of the β-structure have also been detected in some globular proteins.

The α-helix

A second important structural arrangement is the α-helix (see Figure 4); it is formed by a sequence of amino acids wound around a straight axis in either a right-handed or a left-handed spiral. Each turn of the helix corresponds to a distance of 5.4 angstroms (= 0.54 nanometre) in the direction of the screw axis and contains 3.7 amino acids. Hence, the length of the α-helix per amino acid residue is 5.4 divided by 3.7, or 1.5 angstroms (one angstrom = 0.1 nanometre). The stability of the α-helix is maintained by hydrogen bonds between the carbonyl and imino groups of neighbouring turns of the helix. It was once thought, based on data from analyses of the myoglobin molecule, more than half of which consists of α-helices, that the α-helix is the predominant structural element of the globular proteins; it is now known that myoglobin is exceptional in this respect. The other globular proteins for which the structures have been resolved by X-ray diffraction contain only small regions of α-helix. In most of them the peptide chains are folded in an apparently random fashion (see Figure 5), for which the term random coil has been used. The term is misleading, however, because the folding is not random; rather, it is dictated by the primary structure and modified by the secondary and tertiary structures.

The first proteins for which the internal structures were completely resolved are the iron-containing proteins myoglobin and hemoglobin. The investigation of the hydrated crystals of these proteins at Cambridge by Max Perutz and J.C. Kendrew, who won a Nobel Prize for their work, revealed that the folding of the peptide chains is so tight that most of the water is displaced from the centre of the globular molecules. The amino acids that carry the ammonium ($-NH_3^+$) and carboxyl ($-COO^-$) groups were found to be shifted to the surface of the globular molecules, and the nonpolar amino acids were found to be concentrated in the interior.

Other approaches to the determination of protein structure. None of the several other physical methods that have been used to obtain information on the secondary and tertiary structure of proteins provides as much direct information as the X-ray diffraction technique. Most of the techniques, however, are much more simple than X-ray diffraction, which requires, for the resolution of the structure of one protein, many years of work and equipment such as electronic computers. Some of the simpler techniques are based on the optical properties of proteins—refractivity, absorption of light of different wavelengths, rotation of the plane polarized light at different wavelengths, and luminescence.

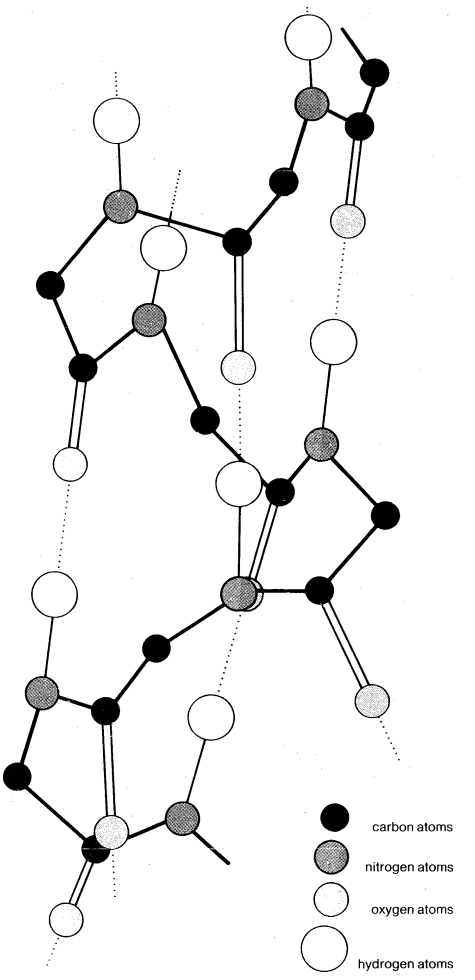


Figure 4: The α-helix (see text).

Spectrophotometric behaviour. Spectrophotometry of protein solutions (the measurement of the degree of absorbance of light by a protein within a specified wavelength) is useful within the range of visible light only with proteins that contain coloured prosthetic groups (the nonprotein components). Examples of such proteins include the red heme proteins of the blood, the purple pigments of the retina of the eye, green and yellow proteins that contain bile pigments, blue copper-containing proteins, and dark brown proteins called melanins. Peptide bonds, because of their carbonyl groups, absorb light energy at very short wavelengths (185–200 nanometres). The aromatic rings of phenylalanine, tyrosine, and tryptophan (see Figure 1), however, absorb ultraviolet light between wavelengths of 280 and 290 nanometres. The absorbance of ultraviolet light by tryptophan is greatest, that of tyrosine is less, and that of phenylalanine is least. If the tyrosine or tryptophan content of the protein is known, therefore, the concentration of the protein solution can be determined by measuring its absorbance between 280 and 290 nanometres.

Polarized light rotation ranges among proteins

Optical activity. It will be recalled that the amino acids, with the exception of glycine, exhibit optical activity (rotation of the plane of polarized light; see above, *Physicochemical properties of the amino acids*). It is not surprising, therefore, that proteins also are optically active. They are usually levorotatory (*i.e.*, they rotate the

plane of polarization to the left) when polarized light of wavelengths in the visible range is used. Although the specific rotation (a function of the concentration of a protein solution and the distance the light travels in it) of most L-amino acids varies from −30° to +30°, the amino acid cystine has a specific rotation of approximately −300°. Although the optical rotation of a protein depends on all of the amino acids, the most important ones are cystine and the aromatic amino acids phenylalanine, tyrosine, and tryptophan. The contribution of the other amino acids to the optical activity of a protein is negligibly small.

Chemical reactivity of proteins. Information on the internal structure of proteins can be obtained with chemical methods that reveal whether certain groups are present on the surface of the protein molecule and thus able to react or whether they are buried inside the closely folded peptide chains and thus are unable to react. The chemical reagents used in such investigations must be mild ones that do not affect the structure of the protein.

The reactivity of tyrosine is of special interest. It has been found, for example, that only three of the six tyrosines found in the naturally occurring enzyme ribonuclease can be iodinated (*i.e.,* reacted to accept an iodine atom). Enzyme-catalyzed breakdown of iodinated ribonuclease is used to identify the peptides in which the iodinated tyrosines are present. The three tyrosines that can be iodinated lie on the surface of ribonuclease; the others, assumed to be inaccessible, are said to be buried in the molecule. Tyrosine can also be identified by using other techniques; *e.g.,* treatment with diazonium compounds or tetranitromethane. Because the compounds formed are coloured, they can easily be detected when the protein is broken down with enzymes.

Cysteine can be detected by coupling with compounds such as iodoacetic acid or iodoacetamide; the reaction results in the formation of carboxymethylcysteine or carbamidomethylcysteine, which can be detected by amino acid determination of the peptides containing them. The imidazole groups of certain histidines can also be located by coupling with the same reagents under different conditions. Unfortunately, few other amino acids can be labelled without changes in the secondary and tertiary structure of the protein.

Association of protein subunits. Many proteins with molecular weights of more than 50,000 occur in aqueous solutions as complexes: dimers, tetramers, and higher polymers—*i.e.,* as chains of two, four, or more repeating basic structural units. The subunits, which are called monomers or protomers, usually are present as an even number. Less than 10 percent of the polymers have been found to have an odd number of monomers. The arrangement of the subunits is thought to be regular and may be cyclic, cubic, or tetrahedral. Some of the small proteins also contain subunits. Insulin, for example, with a molecular weight of about 6,000, consists of two peptide chains linked to each other by disulfide bridges (—S—S—). Similar interchain disulfide bonds have been found in the immunoglobulins. In other proteins, hydrogen bonds and hydrophobic bonds (resulting from the interaction between the amino acid side chains of valine, leucine, isoleucine, and phenylalanine) cause the formation of aggregates of the subunits. The subunits of some proteins are identical; those of others differ. Hemoglobin is a tetramer consisting of two α-chains and two β-chains.

Protein denaturation. When a solution of a protein is boiled, the protein frequently becomes insoluble—*i.e.,* it is denatured—and remains insoluble even when the solution is cooled. The denaturation of the proteins of egg white by heat—as when boiling an egg—is an example of irreversible denaturation. The denatured protein has the same primary structure as the original, or native, protein. The weak forces between charged groups and the weaker forces of mutual attraction of nonpolar groups are disrupted at elevated temperatures, however; as a result, the tertiary structure of the protein is lost. In some instances the original structure of the protein can be regenerated; the process is called renaturation.

Methods of denaturation

Denaturation can be brought about in various ways. Proteins are denatured by treatment with alkaline or acid, oxidizing or reducing agents, and certain organic solvents. Interesting among denaturing agents are those that affect the secondary and tertiary structure without affecting the primary structure. The agents most frequently used for this purpose are urea and guanidinium chloride. These molecules, because of their high affinity for peptide bonds, break the hydrogen bonds and the salt bridges between positive and negative side chains, thereby abolishing the tertiary structure of the peptide chain. When denaturing agents are removed from a protein solution, the native protein re-forms in many cases. Denaturation can also be accomplished by reduction of the disulfide bonds of cystine—*i.e.,* conversion of the disulfide bond (—S—S—) to two sulfhydryl groups (—SH). This, of course, results in the formation of two cysteines. Reoxidation of the cysteines by exposure to air sometimes regenerates the native protein. In other cases, however, the wrong cysteines become bound to each other, resulting in a different protein. Finally, denaturation can also be accomplished by exposing proteins to organic solvents such as ethanol or acetone. It is believed that the organic solvents interfere with the mutual attraction of nonpolar groups.

Some of the smaller proteins, however, are extremely stable, even against heat; for example, solutions of ribonuclease can be exposed for short periods of time to temperatures of 90° C (194° F) without undergoing significant denaturation. Denaturation does not involve identical changes in protein molecules; a common property of denatured proteins, however, is the loss of biological activity—*e.g.,* the ability to act as enzymes or hormones.

Although denaturation had long been considered an all-or-none reaction, it is now thought that many intermediary states exist between native and denatured protein. In some instances, however, the breaking of a key bond could be followed by the complete breakdown of the conformation of the native protein.

Although many native proteins are resistant to the action of the enzyme trypsin, which breaks down proteins during digestion, they are hydrolyzed by the same enzyme after denaturation. Evidently, the peptide bonds that can be split by trypsin are inaccessible in the native proteins but become accessible during denaturation. Similarly, denatured proteins give more intense colour reactions for tyrosine, histidine, and arginine than do the same proteins in the native state. The increased accessibility of reactive groups of denatured proteins is attributed to an unfolding of the peptide chains.

Unfolded peptide chains in denatured proteins

If denaturation can be brought about easily and if renaturation is difficult, how is the native conformation of globular proteins maintained in living organisms, in which they are produced stepwise, by incorporation of one amino acid at a time? Experiments on the biosynthesis of proteins from amino acids containing radioactive carbon or heavy hydrogen reveal that the protein molecule grows stepwise from the N terminus to the C terminus; in each step a single amino acid residue is incorporated. As soon as the growing peptide chain contains six or seven amino acid residues, the side chains interact with each other and thus cause deviations from the straight or β-chain configuration shown in Formula 3. Depending on the nature of the side chains, this may result in the formation of an α-helix (Figure 4) or of loops closed by hydrogen bonds (Formula 5) or disulfide bridges (Formula 6). The final conformation is probably frozen when the peptide chain attains a length of 50 or more amino acid residues.

Conformation of proteins in interfaces. Like many other substances with both hydrophilic and hydrophobic groups, soluble proteins tend to migrate into the interface between air and water or oil and water; the term oil here means a hydrophobic liquid such as benzene or xylene. Within the interface, proteins spread, forming thin films. Measurements of the surface tension, or interfacial tension, of such films indicate that tension is reduced by the protein film. Proteins, when forming an interfacial film, are present as a monomolecular layer; *i.e.,* a layer one molecule in height. Although it was once thought that globular protein molecules unfold completely in the interface, it has now been established that many proteins can be recovered from films in the native state. The application of

lateral pressure on a protein film causes it to increase in thickness and finally to form a layer with a height corresponding to the diameter of the native protein molecule. Protein molecules in an interface, because of Brownian motions (molecular vibrations), occupy much more space than do those in the film after the application of pressure. The Brownian motion of compressed molecules is limited to the two dimensions of the interface, since the protein molecules cannot move upward or downward.

The motion of protein molecules at the air–water interface has been used to determine the molecular weight of proteins. The technique involves measuring the force exerted by the protein layer on a barrier.

When a protein solution is vigorously shaken in air, it forms a foam, because the soluble proteins migrate into the air–water interface and persist there, preventing or slowing the reconversion of the foam into a homogeneous solution. Some of the unstable, easily modified proteins are denatured when spread in the air–water interface. The formation of a permanent foam when egg white is vigorously stirred is an example of irreversible denaturation by spreading in a surface.

Classification of proteins

CLASSIFICATION BY SOLUBILITY

Methods of classification

After two German chemists, Emil Fischer and Franz Hofmeister, independently stated in 1902 that proteins are essentially polypeptides consisting of many amino acids, an attempt was made to classify proteins according to their chemical and physical properties, because the biological function of proteins had not yet been established. (The protein character of enzymes was not proved until the 1920s.) Proteins were classified primarily according to their solubility in a number of solvents. This classification is no longer satisfactory, however, because proteins of quite different structure and function sometimes have similar solubilities; conversely, proteins of the same function and similar structure sometimes have different solubilities. The terms associated with the old classification, however, are still widely used. They are defined below.

Albumins are proteins that are soluble in water and in water half-saturated with ammonium sulfate. On the other hand, globulins are salted out (*i.e.*, precipitated) by half-saturation with ammonium sulfate. Globulins that are soluble in salt-free water are called pseudoglobulins; those insoluble in salt-free water are euglobulins. Both prolamins and glutelins, which are plant proteins, are insoluble in water; the prolamins dissolve in 50 to 80 percent ethanol, the glutelins in acidified or alkaline solution. The term protamine is used for a number of proteins in fish sperm that consist of approximately 80 percent arginine and therefore are strongly alkaline. Histones, which are less alkaline, apparently occur only in cell nuclei, where they are bound to nucleic acids. The term scleroproteins has been used for the insoluble proteins of animal organs. They include keratin, the insoluble protein of certain epithelial tissues such as the skin or hair, and collagen, the protein of the connective tissue. A large group of proteins has been called conjugated proteins, because they are complex molecules of protein consisting of protein and nonprotein moieties. The nonprotein portion is called the prosthetic group. Conjugated proteins can be subdivided into mucoproteins, which, in addition to protein, contain carbohydrate; lipoproteins, which contain lipids; phosphoproteins, which are rich in phosphate; chromoproteins, which contain pigments such as iron-porphyrins, carotenoids, bile pigments, and melanin; and finally, nucleoproteins, which contain nucleic acid.

The weakness of the above classification lies in the fact that many, if not all, globulins contain small amounts of carbohydrate; thus there is no sharp borderline between globulins and mucoproteins. Moreover, the phosphoproteins do not have a prosthetic group that can be isolated; they are merely proteins in which some of the hydroxyl groups of serine are phosphorylated (*i.e.*, contain phosphate). Finally, the globulins include proteins with quite different roles—enzymes, antibodies, fibrous proteins, and contractile proteins.

CLASSIFICATION BY BIOLOGICAL FUNCTIONS

In view of the unsatisfactory state of the old classification, it is preferable to classify the proteins according to their biological function. Such a classification is far from ideal, however, because one protein can have more than one function. The contractile protein myosin, for example, also acts as an ATPase (adenosine triphosphatase), an enzyme that hydrolyzes adenosine triphosphate (removes a phosphate group from ATP by introducing a water molecule). In addition, the definite function of a protein frequently is not known. A protein cannot be called an enzyme as long as its substrate (the specific compound upon which it acts) is not known. It cannot even be tested for its enzymatic action when its substrate is not known.

Special structure and function of proteins

Despite its weaknesses, a functional classification is used here in order to demonstrate, whenever possible, the correlation between the structure and function of a protein. The structural, fibrous proteins are presented first, because their structure is simpler than that of the globular proteins and more clearly related to their function, which is the maintenance of either a rigid or a flexible structure.

STRUCTURAL PROTEINS

Scleroproteins. *Collagen.* Collagen is the structural protein of bones, tendons, ligaments, and skin. For many years collagen was considered to be insoluble in water. Part of the collagen of calf skin, however, can be extracted with citrate buffer at pH 3.7. A precursor of collagen called procollagen is converted in the body into collagen. Procollagen has a molecular weight of 120,000. Cleavage of one or a few peptide bonds of procollagen yields collagen, which has three subunits, each with a molecular weight of 95,000; therefore, the molecular weight of collagen is 285,000 ($3 \times 95{,}000$). The three subunits are wound as spirals around an elongated straight axis. The length of each subunit is 2,900 angstroms, and its diameter is approximately 15 angstroms. The three chains are staggered, so that the trimer has no definite terminal limits.

Structure of collagen

The amino acid composition of collagen is shown in Table 1. It differs from all other proteins in its high content of proline and hydroxyproline. Hydroxyproline does not occur in significant amounts in any other protein except elastin. Most of the proline in collagen is present in the sequence glycine–proline-*X*, in which *X* is frequently alanine or hydroxyproline. Collagen does not contain cystine or tryptophan and therefore cannot substitute for other proteins in the diet. The presence of proline causes kinks

Table 1: Amino Acid Content of Some Proteins

amino acid*	protein					
	α-casein	gliadin	edestin	collagen (ox hide)	keratin (wool)	myosin
Lysine	60.9	4.45	19.9	27.4	6.2	85
Histidine	18.7	11.7	18.6	4.5	19.7	15
Arginine	24.7	15.7	99.2	47.1	56.9	41
Aspartic acid†	63.1	10.1	99.4	51.9	51.5	85
Threonine	41.2	17.6	31.2	19.3	55.9	41
Serine	63.1	46.7	55.7	41.0	79.5	41
Glutamic acid†	153.1	311.0	144.9	76.2	99.0	155
Proline	71.3	117.8	32.9	125.2	58.3	22
Glycine	37.3	—	68.0	354.6	78.0	39
Alanine	41.5	23.9	57.7	115.7	43.8	78
Half cystine	3.6	21.3	10.9	0.0	105.0	86
Valine	53.8	22.7	54.6	21.4	46.6	42
Methionine	16.8	11.3	16.4	6.5	4.0	22
Isoleucine	48.8	90.8‡	41.9	14.5	29.0	42
Leucine	60.3		60.0	28.2	59.9	79
Tyrosine	44.7	17.7	26.9	5.5	28.7	18
Phenylalanine	27.9	39.0	38.4	13.9	22.4	27
Tryptophan	7.8	3.2	6.6	0.0	9.6	—
Hydroxyproline	0.0	0.0	0.0	97.5	12.2	—
Hydroxylysine	—	—	—	8.0	1.2	—
Total	839	765	883	1,058	863	832
Average residual weight	119	131	113	95	117	120

*Number of amino acids is given per 100,000 daltons of protein—*i.e.*, the number of gram molecules of amino acid per 100,000 grams of protein.
†The values for aspartic and glutamic acid include asparagine and glutamine, respectively. ‡Isoleucine plus leucine.

in the peptide chain and thus reduces the length of the amino acid unit from 3.7 angstroms in the extended chain of the β-structure to 2.86 angstroms in the collagen chain. In the intertwined triple helix, the lycines are inside, close to the axis; the prolines are outside.

Native collagen resists the action of trypsin but is hydrolyzed by the bacterial enzyme collagenase. When collagen is boiled with water, the triple helix is destroyed, and the subunits are partially hydrolyzed; the product is gelatin. The unfolded peptide chains of gelatin trap large amounts of water, resulting in a hydrated molecule.

When collagen is treated with tannic acid or with chromium salts, cross links form between the collagen fibres, and it becomes insoluble; the conversion of hide into leather is based on this tanning process. The tanned material is insoluble in hot water and cannot be converted to gelatin. On exposure to water at 62° to 63° C (144° to 145° F), however, the cross links formed by the tanning agents collapse, and the leather contracts irreversibly to about one-third its original volume.

Collagen seems to undergo an aging process in living organisms that may be caused by the formation of cross links between collagen fibres. They are formed by the conversion of some lysine side chains to aldehydes (compounds with the general structure *R*CHO), and the combination of the aldehydes with the ε-amino groups of intact lysine side chains. The protein elastin, which occurs in the elastic fibres of connective tissue, contains similar cross links and may result from the combination of collagen fibres with other proteins. When cross-linked collagen or elastin is degraded, products of the cross-linked lysine fragments, called desmosins and isodesmosins, are formed.

Keratin. Keratin, the structural protein of epithelial cells in the outermost layers of the skin, has been isolated from hair, nails, hoofs, and feathers. Keratin is completely insoluble in cold or hot water; it is not attacked by proteolytic enzymes (*i.e.,* enzymes that break apart, or lyse, protein molecules), and therefore cannot replace proteins in the diet. The great stability of keratin results from the numerous disulfide bonds of cystine. The amino acid composition of keratin differs from that of collagen (see Table 1). Cystine may account for 24 percent of the total amino acids. The peptide chains of keratin are arranged in approximately equal amounts of antiparallel and parallel pleated sheets, in which the peptide chains are linked to each other by hydrogen bonds between the carbonyl (>C=O) and imino (>NH) groups.

Pleated-sheet structure of keratin

Reduction of the disulfide bonds to sulfhydryl groups results in dissociation of the peptide chains, the molecular weight of which is 25,000 to 28,000 each. The formation of permanent waves in the beauty treatment of hair is based on partial reduction of the disulfide bonds of hair keratin by thioglycol, or some other mild reducing agent, and subsequent oxidation of the sulfhydryl groups (−SH) in the reoriented hair to disulfide bonds (−S−S−) by exposure to the oxygen of the air.

The length of keratin fibres depends on their water content. They can bind approximately 16 percent of water; this hydration is accompanied by an increase in the length of the fibres of 10 to 12 percent.

The most thoroughly investigated keratin is hair keratin, particularly that of wool. It consists of a mixture of peptides with high and low cystine content. When wool is heated in water to about 90° C (190° F), it shrinks irreversibly. This is attributed to the breakage of hydrogen bonds and other noncovalent bonds; disulfide bonds do not seem to be affected.

Others. The most thoroughly investigated scleroprotein has been fibroin, the insoluble material of silk. The raw silk comprising the cocoon of the silkworm consists of two proteins. One, sericin, is soluble in hot water; the other, fibroin, is not. The amino acid composition of the latter differs from that of all other proteins. It contains large amounts of glycine, alanine, tyrosine, and serine; small amounts of the other amino acids; and no sulfur-containing ones. The peptide chains are arranged in antiparallel β-structures. Fibroin is partly soluble in concentrated solutions of lithium thiocyanate or in mixtures of cupric salts and ethylene diamine. Such solutions contain a protein of molecular weight 170,000, which is a dimer of two subunits.

Little is known about either the scleroproteins of the marine sponges or the insoluble proteins of the cellular membranes of animal cells. Some of the membranes are soluble in detergents; the membrane of the red blood cells contains an insoluble membrane protein that consists of a single peptide chain of molecular weight 200,000.

The muscle proteins. The total amount of muscle proteins in mammals, including man, exceeds that of any other protein. About 40 percent of the body weight of a healthy human adult weighing about 70 kilograms (150 pounds) is muscle, which is composed of about 20 percent muscle protein. Thus, the human body contains about five to six kilograms (11 to 13 pounds) of muscle protein. An albumin-like fraction of these proteins, originally called myogen, contains various enzymes—phosphorylase, aldolase, glyceraldehyde phosphate dehydrogenase, and others; it does not seem to be involved in contraction. The globulin fraction contains myosin, the contractile protein, which also occurs in blood platelets, small bodies found in blood. Similar contractile substances occur in other contractile structures; for example, in the cilia or flagella (whiplike organs of locomotion) of bacteria and protozoans. In contrast to the scleroproteins, the contractile proteins are soluble in salt solutions and susceptible to enzymatic digestion.

Occurrence of contractile proteins

The energy required for muscle contraction is provided by the oxidation of carbohydrates or lipids. The term mechano-chemical reaction has been used for this conversion of chemical into mechanical energy. Although the molecular process underlying the reaction is not yet completely understood, it is known to involve the fibrous muscle proteins, the peptide chains of which undergo a change in conformation during contraction.

Myosin, which can be removed from fresh muscle by adding it to a chilled solution of dilute potassium chloride and sodium bicarbonate, is insoluble in water. Myosin, solutions of which are highly viscous, consists of an elongated—probably double-stranded—peptide chain, which is coiled at both ends in such a way that a terminal globule is formed. The length of the molecule is approximately 160 nanometres and its average diameter 2.6 nanometres. The equivalent weight of each of the two terminal globules is approximately 30,000; the molecular weight of myosin is close to 500,000. Trypsin splits myosin into large fragments called meromyosin. Myosin contains many amino acids with positively and negatively charged side chains (see Table 1); they form 18 and 16 percent, respectively, of the total number of amino acids. Myosin catalyzes the hydrolytic cleavage of ATP (adenosine triphosphate). A smaller protein with properties similar to those of myosin is tropomyosin. It has a molecular weight of 70,000 and dimensions of 45 by 2 nanometres. More than 90 percent of its peptide chains are present in the α-helix form.

Myosin combines easily with another muscle protein called actin, the molecular weight of which is about 50,000; it forms 12 to 15 percent of the muscle proteins. Actin can exist in two forms—one, G-actin, is globular; the other, F-actin, is fibrous. Actomyosin is a complex molecule formed by one molecule of myosin and one or two molecules of actin. In muscle, actin and myosin filaments are oriented parallel to each other and to the long axis of the muscle. The actin filaments are linked to each other lengthwise by fine threads called S filaments. During contraction the S filaments shorten, so that the actin filaments slide toward each other, past the myosin filaments, thus causing a shortening of the muscle (for a detailed description of the process see MUSCLES AND MUSCLE SYSTEMS: *Muscle contraction*).

Fibrinogen and fibrin. Fibrinogen, the protein of the blood plasma, is converted into the insoluble protein fibrin during the clotting process. The fibrinogen-free fluid obtained after removal of the clot, called blood serum, is blood plasma minus fibrinogen. The fibrinogen content of the blood plasma is 0.2 to 0.4 percent.

Fibrinogen can be precipitated from the blood plasma by half-saturation with sodium chloride. Fibrinogen solutions are highly viscous and show strong flow birefringence. In

electron micrographs the molecules appear as rods with a length of 47.5 nanometres and a diameter of 1.5 nanometres; in addition, two terminal and a central nodule are visible. The molecular weight is 340,000. An unusually high percentage, about 36 percent, of the amino acid side chains are positively or negatively charged.

The clotting process

The clotting process is initiated by the enzyme thrombin, which catalyzes the breakage of a few peptide bonds of fibrinogen; as a result, two small fibrinopeptides with molecular weights of 1,900 and 2,400 are released. The remainder of the fibrinogen molecule, a monomer, is soluble and stable at pH values less than 6 (*i.e.,* in acid solutions). In neutral solution (pH 7) the monomer is converted into a larger molecule, insoluble fibrin; this results from the formation of new peptide bonds. The newly formed peptide bonds form intermolecular and intramolecular cross links, thus giving rise to a large clot, in which all molecules are linked to each other. Clotting, which takes place only in the presence of calcium ions, can be prevented by compounds such as oxalate or citrate, which have a high affinity for calcium ions.

ALBUMINS, GLOBULINS, AND OTHER SOLUBLE PROTEINS

The blood plasma, the lymph, and other animal fluids usually contain one to seven grams of protein per 100 millilitres of fluid, which includes small amounts of hundreds of enzymes and a large number of protein hormones. The discussion below is limited largely to the proteins that occur in large amounts and can be easily isolated from the body fluids. For further information on enzymes and hormones, see below *Enzymes* and *Hormones.*

Proteins of the blood serum. Human blood serum contains about 7 percent protein, two-thirds of which is in the albumin fraction; the other third is in the globulin fraction. Electrophoresis of serum reveals a large albumin peak and three smaller globulin peaks, the alpha-, beta-, and gamma-globulins. The amounts of alpha-, beta-, and gamma-globulin in normal human serum are approximately 1.5, 1.9, and 1.1 percent, respectively. Each globulin fraction is a mixture of many different proteins, as has been demonstrated by immuno-electrophoresis. In this method, the serum of a rabbit injected with human serum is allowed to diffuse into the four protein bands—albumin, alpha-, beta-, and gamma-globulin—obtained from the electrophoresis of human serum. Because the rabbit has previously been injected with human serum, its blood contains antibodies (substances formed in response to a foreign substance introduced into the body) against each of the human serum proteins; each antibody combines with the serum protein (antigen) that caused its formation in the rabbit. The result is the formation of about 20 regions of insoluble antigen-antibody precipitate, which appear as white arcs in the transparent gel of the electrophoresis medium. Each region corresponds to a different human serum protein.

Globulin fractions of serum

Serum albumin is much less heterogeneous (*i.e.,* contains fewer distinct proteins) than are the globulins; in fact, it is one of the few serum proteins that can be obtained in a crystalline form. Serum albumin combines easily with many acidic dyes (*e.g.,* Congo red and methyl orange); with bilirubin, the yellow bile pigment; and with fatty acids. It seems to act, in living organisms, as a carrier for certain biological substances. Present in blood serum in relatively high concentration, serum albumin also acts as a protective colloid, a protein that stabilizes other proteins. Albumin (molecular weight of 68,000) has a single free sulfhydryl (−SH) group, which on oxidation forms a disulfide bond with the sulfhydryl group of another serum albumin molecule, thus forming a dimer. The isoelectric point of serum albumin is pH 4.7.

The alpha-globulin fraction of blood serum is a mixture of several conjugated proteins. The best known are an α-lipoprotein (combination of lipid and protein) and two mucoproteins (combinations of carbohydrate and protein). One mucoprotein is called orosomucoid, or α_1-acid glycoprotein; the other is called haptoglobin because it combines specifically with globin, the protein component of hemoglobin. Haptoglobin contains about 20 percent carbohydrate.

The beta-globulin fraction of serum contains, in addition to lipoproteins and mucoproteins, two metal-binding proteins, transferrin and ceruloplasmin, which bind iron and copper, respectively. They are the principal iron and copper carriers of the blood.

The gamma-globulins are the most heterogeneous globulins. Although most have a molecular weight of approximately 150,000, that of some, called macroglobulins, is as high as 800,000. Because typical antibodies are of the same size and exhibit the same electrophoretic behaviour as γ-globulins, they are called immunoglobulins. The designation IgM or gamma M (γM) is used for the macroglobulins; the designation IgG or gamma G (γG) is used for γ-globulins of molecular weight 150,000.

Milk proteins. Milk contains the following: an albumin, α-lactalbumin; a globulin, beta-lactoglobulin; and a phosphoprotein, casein. If acid is added to milk, casein precipitates. The remaining watery liquid (the supernatant solution), or whey, contains lactalbumin and lactoglobulin. Both have been obtained in crystalline form; their molecular weights are 16,000 and 18,500, respectively. Lactoglobulin also occurs as a dimer of molecular weight 37,000. Small variations known to occur in the amino acid composition of lactoglobulin result from genetic variations. The amino acid composition and the tertiary structure of lactalbumin resemble that of lysozyme, an egg protein (see below).

Casein is precipitated not only by the addition of acid but also by the action of the enzyme rennin, which is found in gastric juice. Rennin from calf stomachs is used to precipitate casein, from which cheese is made. Milk fat precipitates with casein; milk sugar, however, remains in the supernatant (whey). Casein is a mixture of several similar phosphoproteins, called α-, β-, γ-, and κ-casein, all of which contain some serine side chains combined with phosphoric acid. Approximately 75 percent of casein is α-casein (see Table 1). Cystine has been found only in κ-casein. In milk, casein seems to form polymeric globules (micelles) with radially arranged monomers, each with a molecular weight of 24,000; the acidic side chains occur predominantly on the surface of the micelle, rather than inside.

Composition of casein

Egg proteins. About 50 percent of the proteins of egg white are composed of ovalbumin, which is easily obtained in crystals. Its molecular weight is 46,000 and its amino acid composition differs from that of serum albumin. Other proteins of egg white are conalbumin, lysozyme, ovoglobulin, ovomucoid, and avidin. Lysozyme is an enzyme that hydrolyzes the carbohydrates found in the capsules certain bacteria secrete around themselves; it causes lysis (disintegration) of the bacteria. The molecular weight of lysozyme is 14,100; its amino acid composition is shown in Table 2. Its three-dimensional structure, shown in Figure 5, is similar to that of α-lactalbumin, which stimulates the formation of lactose by the enzyme lactose synthetase. Lysozyme has also been found in the urine of patients suffering from leukemia.

Avidin is a glycoprotein that combines specifically with biotin, a vitamin. In animals fed large amounts of raw egg white, the action of avidin results in "egg-white injury." The molecular weight of avidin, which forms a tetramer, is 16,200. Its amino acid sequence is known.

Egg-yolk proteins contain a mixture of lipoproteins and livetins. The latter are similar to serum albumin, α-globulin, and β-globulin. The yolk also contains a phosphoprotein, phosvitin. Phosvitin, which has also been found in fish sperm, has a molecular weight of 40,000 and an unusual amino acid composition; one third of its amino acids are phosphoserine.

Protamines and histones. Protamines are found in the sperm cells of fish. The most thoroughly investigated protamines are salmine from salmon sperm and clupeine from herring sperm. The protamines are bound to deoxyribonucleic acid (DNA), forming nucleoprotamines. The amino acid composition of the protamines is simple; they contain, in addition to large amounts of arginine, small amounts of five or six other amino acids. The composition of the salmine molecule, for example, is: Arg_{51}, Ala_4, Val_4, Ile_1, Pro_7, and Ser_6, in which the subscript numbers indicate

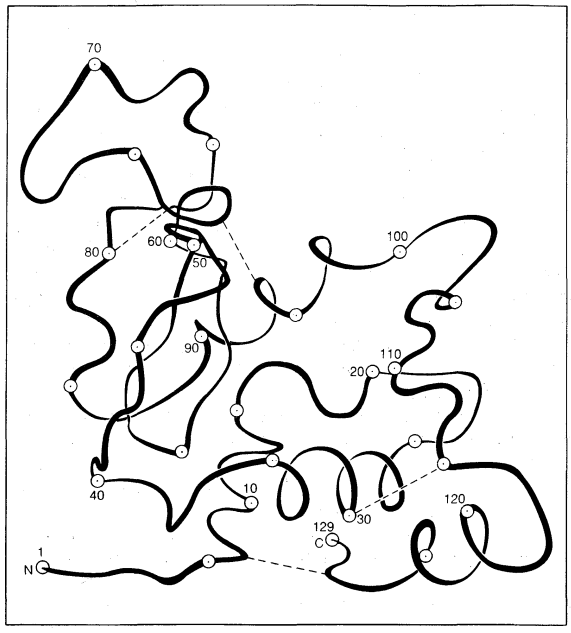


Figure 5: Conformation of lysozyme. Lysozyme from hen's egg white has a single peptide chain of 129 amino acids. The diagram shows the structure in simplified form. The amino acid residues are numbered from the terminal α amino group (N) to the terminal carboxyl group (C). Every fifth residue is shown by a circle, and every tenth residue is numbered. The four disulfide bridges are shown by broken lines. Alpha-helices are visible in the ranges 25 to 35, 90 to 100, and 120 to 125.

Protamine composition

the number of each amino acid in the molecule. Because of the high arginine content, the isoelectric points of the protamines are at pH values of 11 to 12; *i.e.*, the protamines are alkaline. The molecular weights of salmine and clupeine are close to 6,000. All of the protamines investigated thus far are mixtures of several similar proteins.

The histones are less basic than the protamines. They contain high amounts of either lysine or arginine and small amounts of aspartic acid and glutamic acid. Histones occur in combination with DNA as nucleohistones in the nuclei of the body cells of animals and plants, but not in animal sperm. The molecular weights of histones vary from 10,000 to 22,000. In contrast to the protamines, the histones contain most of the 20 amino acids, with the exception of tryptophan and the sulfur-containing ones. Like the protamines, histone preparations are heterogeneous mixtures. The amino acid sequence of some of the histones has been determined.

Plant proteins. Plant proteins, mostly globulins, have been obtained chiefly from the protein-rich seeds of cereals and legumes. Small amounts of albumins are found in seeds. The best known globulins, insoluble in water, can be extracted from seeds by treatment with 2 to 10 percent solutions of sodium chloride. Many plant globulins have been obtained in crystalline form; they include edestin from hemp, molecular weight 310,000; amandin from almonds, 330,000; concanavalin A (42,000) and B (96,000); and canavalin (113,000) from jack beans. They are polymers of smaller subunits; edestin, for example, is a hexamer of a subunit with a molecular weight of 50,000, and concanavalin B a trimer of a subunit with a molecular weight of 30,000. After extraction of lipids from cereal seeds by ether and alcohol, further extraction with water containing 50 to 80 percent of alcohol yields proteins that are insoluble in water but soluble in water–ethanol mixtures and have been called prolamins. Their solubility in aqueous ethanol may result from their high proline and glutamine content (see Table 1). Gliadin, the prolamin from wheat, contains 14 grams of proline and 46 grams of glutamic acid in 100 grams of protein; most of the glutamic acid is in the form of glutamine. The total amounts of the basic amino acids (arginine, lysine, and histidine) in gliadin are only 5 percent of the weight of gliadin. None of the prolamins has yet been obtained in a pure crystalline state. Because the lysine content is either low or nonexistent, human populations dependent on grain as a sole protein source suffer from lysine deficiency.

CONJUGATED PROTEINS

Chemical bonding in conjugated proteins

Combination of proteins with prosthetic groups. The link between a protein molecule and its prosthetic group is a covalent (electron-sharing) bond in the glycoproteins, the biliproteins, and some of the heme proteins. In lipoproteins, nucleoproteins, and some heme proteins, the two components are linked by noncovalent bonds; the bonding results from the same forces that are responsible for the tertiary structure of proteins: hydrogen bonds, salt bridges between positively and negatively charged groups, disulfide bonds, and mutual interaction of hydrophobic groups. In the metalloproteins (proteins with a metal element as a prosthetic group), the metal ion usually forms a centre to which various groups are bound.

Some of the conjugated proteins have been mentioned in preceding sections because they occur in the blood serum, in milk, and in eggs; others are discussed below in sections dealing with respiratory proteins and enzymes.

Mucoproteins and glycoproteins. The prosthetic groups in mucoproteins and glycoproteins are oligosaccharides (carbohydrates consisting of a small number of simple sugar molecules) usually containing from four to 12 sugar molecules; the most common sugars are galactose, mannose, glucosamine, and galactosamine. Xylose, fucose, glucuronic acid, sialic acid, and other simple sugars sometimes also occur. Some mucoproteins contain 20 percent or more of carbohydrate, usually in several oligosaccharides attached to different parts of the peptide chain. The designation mucoprotein is used for proteins with more than 3 to 4 percent carbohydrate; if the carbohydrate content is less than 3 percent, the protein is sometimes called a glycoprotein or simply a protein.

Mucoproteins, highly viscous proteins originally called mucins, are found in saliva, in gastric juice, and in other animal secretions. Mucoproteins occur in large amounts in cartilage, synovial fluid (the lubricating fluid of joints and tendons), and egg white. The mucoprotein of cartilage is formed by the combination of collagen with chondroitinsulfuric acid, which is a polymer of either glucuronic or iduronic acid and acetylhexosamine or acetylgalactosamine. It is not yet clear whether or not chondroitinsulfate is bound to collagen by covalent bonds.

Lipoproteins and proteolipids. The bond between the protein and the lipid portion of lipoproteins and proteolipids is a noncovalent one. It is believed that some of the lipid is enclosed in a meshlike arrangement of peptide chains and becomes accessible for reaction only after the unfolding of the chains by denaturing agents. Although

Table 2: Number of Amino Acids per Protein Molecule

amino acid	protein*						
	Cyto	Hb α	Hb β	RNase	Lys	Chgen	Fdox
Lysine	18	11	11	10	6	14	4
Histidine	3	10	9	4	1	2	1
Arginine	2	3	3	4	11	4	1
Aspartic acid†	8	12	13	15	21	23	13
Threonine	7	9	7	10	7	23	8
Serine	2	11	5	15	10	28	7
Glutamic acid†	10	5	11	12	5	15	13
Proline	4	7	7	4	2	9	4
Glycine	13	7	13	3	12	23	6
Alanine	6	21	15	12	12	22	9
Half cystine	2	1	2	8	8	10	5
Valine	3	13	18	9	6	23	7
Methionine	3	2	1	4	2	2	0
Isoleucine	8	0	0	3	6	10	4
Leucine	6	18	18	2	8	19	8
Tyrosine	5	3	3	6	3	4	4
Phenylalanine	3	7	8	3	3	6	2
Tryptophan	1	1	2	0	6	8	1
Total	104	141	146	124	129	245	97

*Cyto = human cytochrome *c*; Hb α = human hemoglobin A, α-chain; Hb β = human hemoglobin A, β-chain; RNase = bovine ribonuclease; Lys = chicken lysozyme; Chgen = bovine chymotrypsinogen; Fdox = spinach ferredoxin. †The values recorded for aspartic and glutamic acid include asparagine and glutamine.

lipoproteins in the α- and β-globulin fraction of blood serum are soluble in water (but insoluble in organic solvents), some of the brain lipoproteins, because they have a high lipid content, are soluble in organic solvents; they are called proteolipids. The β-lipoprotein of human blood serum is a macroglobulin with a molecular weight of about 1,300,000, 70 percent of which is lipid; of the lipid, about 30 percent is phospholipid and 40 percent cholesterol and compounds derived from it. Because of their lipid content, the lipoproteins have the lowest density (mass per unit volume) of all proteins and are usually classified as low- and high-density lipoproteins (LDL and HDL).

Coloured lipoproteins are formed by the combination of protein with carotenoids. Crustacyanin, the pigment of lobsters, crayfish, and other crustaceans, contains astaxanthin, which is a compound derived from carotene.

Pigments of the eye

Among the most interesting of the coloured lipoproteins are the pigments of the retina of the eye. They contain retinal, which is a compound derived from carotene and which is formed by the oxidation of vitamin A. In rhodopsin, the red pigment of the retina, the aldehyde group (−CHO) of retinal forms a covalent bond with an amino ($-NH_2$) group of opsin, the protein carrier. Colour vision is mediated by the presence of several visual pigments in the retina that differ from rhodopsin either in the structure of retinal or in that of the protein carrier.

Metalloproteins. Proteins in which heavy metal ions are bound directly to some of the side chains of histidine, cysteine, or some other amino acid are called metalloproteins. Two metalloproteins, transferrin and ceruloplasmin, occur in the globulin fractions of blood serum; they act as carriers of iron and copper, respectively. Transferrin has a molecular weight of 84,000 and consists of two identical subunits, each of which contains one ferric ion (Fe^{3+}) that seems to be bound to tyrosine. Several genetic variants of transferrin are known to occur in man. Another iron protein, ferritin, which contains 20 to 22 percent iron, is the form in which iron is stored in animals; it has been obtained in crystalline form from liver and spleen. A molecule consisting of 20 subunits, its molecular weight is approximately 480,000. The iron can be removed by reduction from the ferric (Fe^{3+}) to the ferrous (Fe^{2+}) state. The iron-free protein, apoferritin, is synthesized in the body before the iron is incorporated.

Green plants and some photosynthetic and nitrogen-fixing bacteria (*i.e.*, bacteria that convert atmospheric nitrogen, N_2, into amino acids and proteins in their own bodies) contain various ferredoxins. They are small proteins containing 50 to 100 amino acids (see Table 2) and a chain of iron and disulfide units (FeS_2), in which some of the sulfur atoms are contributed by cysteine; others are sulfide ions (S^{2-}). The number of FeS_2 units per ferredoxin molecule varies from five in the ferredoxin of spinach to 10 in the ferredoxin of certain bacteria. Ferredoxins act as electron carriers in photosynthesis and in nitrogen fixation.

Function of ceruloplasmin

Ceruloplasmin is a copper-containing globulin with a molecular weight of 151,000; the molecule consists of eight subunits, each containing one copper ion. Ceruloplasmin is the principal carrier of copper in organisms, although copper can also be transported by the iron-containing globulin transferrin. Another copper-containing protein, erythrocuprein (molecular weight 64,000), has been isolated from red blood cells; it has also been found in the liver and the brain. The molecule, which consists of four subunits with a molecular weight of 16,000 each, contains four copper and four zinc ions. Because of their copper content, ceruloplasmin and erythrocuprein may have some catalytic activity in oxidation-reduction reactions. Another copper-containing protein, hemocyanin, is described below (see *Respiratory proteins*).

Many animal enzymes contain zinc ions, which are usually bound to the sulfur of cysteine. Horse kidneys contain the protein metallothionein, which contain zinc and cadmium; both are bound to sulfur. A vanadium-protein complex (homovanadin) has been found in surprisingly high amounts in yellowish-green cells (vanadocytes) of tunicates, which are marine invertebrates.

Heme proteins and other chromoproteins. Although the heme proteins contain iron, they are usually not classified as metalloproteins, because their prosthetic group is an iron-porphyrin complex in which the iron is bound very firmly. The intense red or brown colour of the heme proteins is not caused by iron but by porphyrin, a complex cyclic structure. All porphyrin compounds absorb light intensely at or close to 410 nanometres. Porphyrin consists of four pyrrole rings (five-membered closed structures containing one nitrogen and four carbon atoms) linked to each other by methine groups (−CH=). The iron atom is kept in the centre of the porphyrin ring by interaction with the four nitrogen atoms. The iron atom can combine with two other substituents; in oxyhemoglobin, one substituent is a histidine of the protein carrier, the other is an oxygen molecule. In some heme proteins, the protein is also bound covalently to the side chains of porphyrin. Heme proteins are described below (see *Respiratory proteins* and *Oxidoreductases*).

Little is known about the structure of the chromoprotein melanin, a pigment found in dark skin, dark hair, and melanotic tumours. It is probably formed by the oxidation of tyrosine, which results in the formation of red, brown, or dark-coloured derivatives.

Green chromoproteins called biliproteins are found in many insects, such as grasshoppers, and also in the eggshells of many birds. The biliproteins are derived from the bile pigment biliverdin, which in turn is formed from porphyrin; biliverdin contains four pyrrole rings and three of the four methine groups of porphyrin. Large amounts of biliproteins, the molecular weights of which are about 270,000, have been found in red and blue-green algae; the red protein is called phycoerythrin, the blue one phycocyanobilin. Phycocyanobilin consists of eight subunits with a molecular weight of 28,000 each; about 89 percent of the molecule is protein with a large amount of carbohydrate.

Nucleoproteins. When a protein solution is mixed with a solution of a nucleic acid, the phosphoric acid component of the nucleic acid combines with the positively charged ammonium groups ($-NH_3^+$) of the protein to form a protein–nucleic acid complex. The nucleus of a cell contains predominantly deoxyribonucleic acid (DNA) and the cytoplasm predominantly ribonucleic acid (RNA); both parts of the cell also contain protein. Protein–nucleic acid complexes, therefore, form in living cells. It has not yet been definitely established whether the protein–nucleic acid complexes isolated from biological material are indeed formed during the life of the organism or whether they are artifacts produced during the isolation procedure.

The only nucleoproteins for which some evidence for specificity exists are nucleoprotamines, nucleohistones, and some RNA and DNA viruses. The nucleoprotamines are the form in which protamines occur in the sperm cells of fish; the histones of the thymus and of pea seedlings and other plant material apparently occur predominantly as nucleohistones. Both nucleoprotamines and nucleohistones contain only DNA.

Proteins of the tobacco mosaic virus

Some of the simplest viruses consist of a specific RNA, which is coated by protein. One of the best known RNA viruses, tobacco mosaic virus (TMV), has the shape of a rod. RNA comprises only 5.1 percent of the mass of the virus. The complete sequence of the virus protein, which consists of about 2,130 identical peptide chains, each containing 158 amino acids, has been determined. The protein is arranged in a spiral around the RNA core.

DNA has been found in most bacterial viruses (bacteriophages) and in some animal viruses. As in TMV, the core of DNA is surrounded by protein. Phage protein is a mixture of enzymes and therefore cannot be considered as the protein portion of only one nucleoprotein.

Respiratory proteins. *Hemoglobin.* Hemoglobin is the oxygen carrier in all vertebrates and some invertebrates. In oxyhemoglobin (HbO_2), which is bright red, the ferrous ion (Fe^{2+}) is bound to the four nitrogen atoms of porphyrin; the other two substituents are an oxygen molecule and the histidine of globin, the protein component of hemoglobin. Deoxyhemoglobin (deoxy-Hb), as its name implies, is oxyhemoglobin minus oxygen (*i.e.*, reduced hemoglobin); it is purple in colour. Oxidation of the ferrous ion of hemoglobin yields a ferric compound, methemoglobin, sometimes called hemiglobin or ferrihemoglobin. The oxy-

C.S.N.L.S.T.C.V.L.S.A.Y.W.K.D.L.N.N.Y.H.R.F.S.G.M.G.F.G.P.E.T.P($CONH_2$)

Formula 7: The amino acid sequence of human calcitonin. At the left end the line represents the disulfide bond. At the right end ($CONH_2$) indicates that the C terminal proline is present as prolinamide.

gen of oxyhemoglobin can be displaced by carbon monoxide, for which hemoglobin has a much greater affinity, preventing oxygen from reaching the body tissues.

The hemoglobins of all mammals, birds, and many other vertebrates are tetramers of two α- and two β-chains (see Table 2). The molecular weight of the tetramer is 64,500; the molecular weight of the α- and β-chains is approximately 16,100 each, and the four subunits are linked to each other by noncovalent interactions. If hemin (the ferric porphyrin component) is removed from globin (the protein component), two molecules of globin, each consisting of one α- and one β-chain, are obtained; the molecular weight of globin is 32,200. In contrast to hemoglobin, globin is an unstable protein that is easily denatured. If native globin is incubated with a solution of hemin at pH values of 8 to 9, native hemoglobin is reconstituted. Both the hemoglobin of the lamprey and the myoglobin, the red pigment of mammalian muscles, are monomers with a molecular weight of 16,000.

Mammalian hemoglobins

The mammalian hemoglobins differ from each other in their amino acid composition and therefore in their secondary and tertiary structure. Rat and horse hemoglobin crystallize very easily, but those of man, cattle, and sheep, because they are more soluble, are difficult to crystallize. The shape of hemoglobin crystals varies in different species; moreover, decomposition and denaturation occur at different rates in different species. It was also found that the blood of newborn children contains two different hemoglobins, about 20 percent of an adult hemoglobin (hemoglobin A) and 80 percent of a fetal hemoglobin (hemoglobin F). Hemoglobin F persists in the child for the first seven months of life. The same hemoglobin F has also been found in the blood of patients suffering from thalassemia, an anemia that occurs in the countries of southern Europe. Hemoglobin F contains, as does hemoglobin A, two α-chains; the two β-chains, however, have been replaced by two quite different γ-chains. When the technique of electrophoresis was first applied to the hemoglobin of blacks suffering from sickle cell anemia in 1949, a new hemoglobin (hemoglobin S) was discovered. More than 100 different human hemoglobins now are known. They differ from normal hemoglobin A in the amino acid composition of either the α- or the β-chain.

The hemoglobins of some of the lowest fishes are monomers containing one iron atom per molecule. Hemoglobin-like respiratory proteins have been found in some invertebrates. The red hemoglobin of insects, mollusks, and protozoans is called erythrocruorin. It differs from vertebrate hemoglobin by its high molecular weight.

Although green plants contain no hemoglobin, a red protein, called leg-hemoglobin, has been discovered in the root nodules of leguminous plants. It seems to be produced by the nitrogen-fixing bacteria of the root nodules and may be involved in the reduction of atmospheric nitrogen to ammonia and amino acids.

Other respiratory proteins. A green respiratory protein, chlorocruorin, has been found in the blood of the marine worm *Spirographis*. It has the same high molecular weight as erythrocruorin, but differs from hemoglobin in its prosthetic group. A red metalloprotein, hemerythrin, acts as a respiratory protein in marine worms of the phylum Sipuncula. The molecule consists of eight subunits with a molecular weight of 13,500 each. Hemerythrin contains no porphyrins and therefore is not a heme protein.

A metalloprotein containing copper is the respiratory protein of crustaceans (shrimps, crabs, etc.) and of some gastropods (snails). The protein, called hemocyanin, is pale yellow when not combined with oxygen, and blue when combined with oxygen. The molecular weights of hemocyanins vary from 300,000 to 9,000,000. Each animal investigated thus far apparently has a species-specific hemocyanin.

Hemocyanin

PROTEIN HORMONES

Some hormones that are products of endocrine glands are proteins or peptides; others are steroids. (The origin of hormones, their physiological role, and their mode of action are dealt with below in the section *Hormones*.) None of the hormones has any enzymatic activity. Each has a target organ in which it elicits some biological action; *e.g.*, secretion of gastric or pancreatic juice, production of milk, production of steroid hormones. The mechanism by which the hormones exert their effects is not fully understood. Cyclic adenosine monophosphate is involved in the transmittance of the hormonal stimulus to the cells whose activity is specifically increased by the hormone.

Functional aspects of hormones

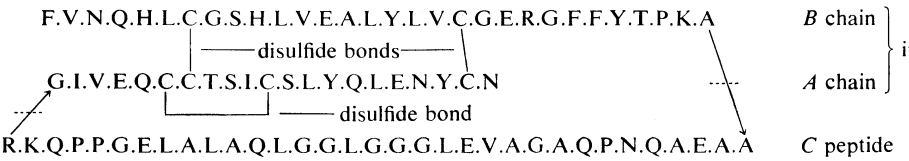
Hormones of the thyroid gland. Thyroglobulin, the active groups of which are two molecules of the iodine-containing compound thyroxine (see Figure 1), has a molecular weight of 670,000.

Thyroglobulin also contains thyroxine with two and three iodine atoms instead of the four shown in Figure 1, and trosine, with one and two iodine atoms. Injection of the hormone causes an increase in metabolism; lack of it results in a slowdown. Another hormone, calcitonin, which lowers the calcium level of the blood, occurs in the thyroid gland. The structure of human calcitonin is given in Formula 7 (see Figure 1 for structures of amino acids corresponding to the one-letter codes).

The amino acid sequences of calcitonin from pig, beef, and salmon differ from human calcitonin in some amino acids. All of them, however, have the half-cystines and the prolinamide in the same position. Porcine calcitonin has been synthesized in the laboratory.

The parathyroid hormone (parathormone), produced in small glands that are embedded in or lie behind the thyroid gland, is essential for maintaining the calcium level of the blood. Its lack results in the disease hypocalcemia. Bovine parathormone has a molecular weight of 8,500; it contains no cystine or cysteine and is rich in aspartic acid, glutamic acid, or their amides.

Hormones of the pancreas. Although the structure of insulin has been known since 1949, repeated attempts to synthesize it gave very poor yields because of the failure of the two peptide chains to combine forming the correct disulfide bridge. The ease of the biosynthesis of insulin is explained by the discovery in the pancreas of proinsulin, from which insulin is formed. The single peptide chain of proinsulin loses a peptide consisting of 33 amino acids and called the connecting peptide, or C peptide, during its conversion to insulin. The structure of porcine proinsulin is shown in Formula 8.



Formula 8: The amino acid sequence of porcine proinsulin. The arrows indicate the direction from the N terminus of the *B* chain to the C terminus of the *A* chain. In the *C* peptide the N terminus is at the right (A=alanine) end and the C terminus is at the left end in order to show the connection with the *A* and *B* chains. Insulin is released when the two peptide bonds at the ends of the *C* peptide (broken lines) are hydrolyzed (broken by the introduction of a water molecule).

In aqueous solutions insulin exists predominantly as a complex of six subunits, each of which contains an *A*- and a *B*-chain. The insulins of several species have been isolated and analyzed; their amino acid sequences differ somewhat, but all apparently contain the same disulfide bridges between the two chains.

Opposite functions of insulin and glucagon

Although the injection of insulin lowers the blood sugar, administration of glucagon, another pancreas hormone, raises the blood sugar level. Glucagon consists of a straight peptide chain of 29 amino acids. Its structure, which is free of cystine and isoleucine, is given in Formula 9. It has been synthesized; the synthetic product has the full biological activity of natural glucagon.

Pituitary hormones. The pituitary gland has an anterior lobe, a posterior lobe, and an intermediate portion; they differ in cellular structure and in the structure and action of the hormones they form. The posterior lobe produces

H.S.Q.G.T.F.T.S.D.Y.S.K.Y.L.D.S.R.R.A.Q.D.F.V.Q.W.L.M.N.T

Formula 9: Amino acid sequence of the pancreas hormone glucagon.

two similar hormones, oxytocin and vasopressin. The former causes contraction of the pregnant uterus; the latter raises the blood pressure. Both are octapeptides formed by a ring of five amino acids (the two cystine halves count as one amino acid) and a side chain of three amino acids. The structure of oxytocin is given in Formula 10. The two cystine halves are linked to each other by a disulfide bond, and the C terminal amino acid is glycinamide. The structure has been established and confirmed. Human vasopressin (Formula 10) differs from oxytocin in that isoleucine is replaced by phenylalanine and leucine by arginine. Porcine vasopressin contains lysine instead of arginine.

A Cys.Tyr.Ile.GluN.Asn.Cys.Pro.Leu.Gly($CONH_2$)

B Cys.Tyr.Phe.GluN.Asn.Cys.Pro.Arg.Gly($CONH_2$)

Formula 10: Amino acid sequence of the two similar hormones of the posterior lobe of the pituitary gland. (A) Oxytocin. (B) Human vasopressin. The solid line represents the disulfide bond between the two halves of cystine.

The intermediate part of the pituitary gland produces the melanocyte-stimulating hormone (MSH), which causes expansion of the pigmented melanophores (cells) in the skin of frogs and other batrachians. Two hormones, called α-MSH and β-MSH, have been prepared from hog pituitary glands. α-MSH consists of 13 amino acids (see Formula 11); its N terminal serine is acetylated (*i.e.*, the acetyl group, CH_3CO, of acetic acid is attached), and its C terminal valine residue is present as valinamide. β-MSH contains in its 18 amino acids many of those occurring in α-MSH (see Formula 11).

The anterior pituitary lobe produces several protein hormones—a thyroid-stimulating hormone, molecular weight 28,000; a lactogenic hormone, molecular weight 22,500; a growth hormone, molecular weight 21,500; a luteinizing hormone, molecular weight 30,000; and a follicle-stimulating hormone, molecular weight 29,000. The thyroid-stimulating hormone (TSH, thyrotropin) consists of α and β subunits with a composition similar to the subunits of luteinizing hormone. When separated, neither of the two subunits has hormonal activity; when combined, however, they regain about 50 percent of the original activity. The lactogenic hormone (prolactin) from sheep pituitary glands contains 190 amino acids. Their sequence has been elucidated; a similar peptide chain of 188 amino acids that has been synthesized not only has 10 percent of the biological activity of the natural hormone but also some activity of the growth hormone. The amino acid sequence of the growth hormone (somatotropic hormone) is also known; it seems to stimulate the synthesis of RNA and in this way to accelerate growth. The luteinizing hormone (LH) consists of two subunits, each with a molecular weight of approximately 15,000; when separated, the subunits recombine spontaneously. LH is a mucoprotein containing about 12 percent carbohydrate. The urine of pregnant women contains chorionic gonadotropin, the presence of which makes possible early diagnosis of pregnancy. The amino acid sequence is known. The sequence of 160 of its 190 amino acids is identical with those of the growth hormone; 100 of these also occur in the same sequence as in lactogenic hormone. The different pituitary hormones and the chorionic gonadotropin thus may have been derived from a common substance that, during evolution, underwent differentiation.

Derivation of pituitary hormones

Peptides with hormonelike activity. Small peptides have been discovered that, like hormones, act on certain target organs. One peptide, angiotensin (angiotonin or hypertensin), is formed in the blood from angiotensinogen by the action of renin, an enzyme of the kidney. It is an octapeptide and increases blood pressure. Similar peptides include bradykinin, which stimulates smooth muscles; gastrin, which stimulates secretion of hydrochloric acid and pepsin in the stomach; secretin, which stimulates the flow of pancreatic juice; and kallikreine, the activity of which is similar to bradykinin.

IMMUNOGLOBULINS AND ANTIBODIES

Antibodies, proteins that combat foreign substances in the body, are associated with the globulin fraction of the immune serum (see IMMUNITY). As stated previously, when the serum globulins are separated into α-, β-, and γ- fractions, antibodies are associated with the γ-globulins. Antibodies can be purified by precipitation with the antigen (*i.e.*, the foreign substance) that caused their formation, followed by separation of the antigen-antibody complex. Antibodies prepared in this way consist of a mixture of many similar antibody molecules, which differ in molecular weight, amino acid composition, and other properties. The same differences are found in the γ-globulins of normal blood serums. It is believed that the γ-globulin of normal blood serum is a mixture of thousands of different γ-globulins, each of which occurs in amounts too small for isolation. Because the physical and chemical properties of normal γ-globulins are the same as those of antibodies, the γ-globulins are frequently called immunoglobulins. They may be considered to be antibodies against unknown antigens. If solutions of γ-globulin are resolved by gel filtration through dextran, the first fraction has a molecular weight of 800,000. This fraction is called IgM or γM; Ig is an abbreviation for immunoglobulin and M for macroglobulin. The next two fractions are IgA (γA) and IgG (γG), with molecular weights of about 300,000 and 150,000 respectively. Two other immunoglobulins, known as IgD and IgE, have also been detected in much smaller amounts in some immune sera.

γ-globulins of blood serums

The bulk of the immunoglobulins is found in the IgG fraction, which also contains most of the antibodies. The IgM molecules are apparently pentamers—aggregates of five of the IgG molecules. Electron microscopy shows their five subunits to be linked to each other by disulfide bonds in the form of a pentagon. The IgA molecules are found principally in milk and in secretions of the intestinal mucosa. Some of them contain, in addition to a dimer of IgG, a "secretory piece," the structure of which

(CH_3CO)S.Y.S. M.E.H.F.R.W.G.K.P.V($CONH_2$) porcine α-MSH, melanocyte-stimulating hormone

D.S.G.P.Y.K.M.E.H.F.R.W.G.S.P.P.K.D porcine β-MSH

A.E.K.K.D.E.G.P.Y.K.M.E.H.F.R.W.G.S.P.P.K.D human β-MSH

S.Y.S. M.E.H.F.R.W.G.K.P.V.G.K.K.R.R.P.V.K.V.Y.P.D.G.A.E.D.Q.L.A.E.A.F.P.L.E.F porcine β-corticotropin

Formula 11: The amino acid sequence of hormones produced by the intermediate part of the pituitary gland. The amino acid sequence M.E.H.F.R.W.G. occurs in all melanocyte-stimulating hormones and in adrenocorticotropic hormones (corticotropins).

is not yet known. The IgM and IgA immunoglobulins and antibodies contain 10 to 15 percent carbohydrate; the carbohydrate content of the IgG molecules is 2 to 3 percent.

IgG molecules treated with the enzyme papain split into three fragments of almost identical molecular weight of 50,000. Two of these, called Fab fragments, are identical; the third is abbreviated Fc. Reduction to sulfhydryl groups of some of the disulfide bonds of IgG results in the formation of two heavy, or *H*, chains (molecular weight 55,000) and two light, or *L*, chains (molecular weight 22,000). They are linked by disulfide bonds in the order *L*—*H*—*H*—*L*. Each *H* chain contains four intrachain disulfide bonds, each *L* chain contains two. The structure of antibodies and normal immunoglobulins of the IgG type is shown in Figure 6.

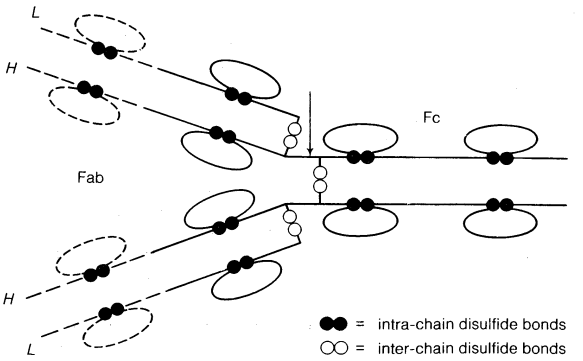


Figure 6: Diagram of an IgG immunoglobulin. Two heavy chains (*H*) and two light chains (*L*) are linked to each other by inter-chain disulfide bonds. Intra-chain disulfide bonds cause loops to form in the 12 peptide portions, each of which contains about 110 amino acid residues. The 12 peptide regions have cystine residues at similar positions and other similarities in their amino acid sequences. The broken lines represent variable portions and the solid lines represent constant portions of the chains. Specific sites that bind antigens are formed by the variable portions. The vertical arrow indicates cleavage of the IgG molecule into two Fab fragments and one Fc fragment by the action of the enzyme papain.

Antibody preparations of the IgG type, even after removal of IgM and IgA antibodies, are heterogeneous. The *H* and *L* chains consist of a large number of different *L* chains and a variety of *H* chains. Pure IgG, IgM, and IgA immunoglobulins, however, occur in the blood serum of patients suffering from myelomas, which are malignant tumours of the bone marrow. The tumours produce either an IgG, an IgM, or an IgA protein, but rarely more than one class. A protein called the Bence-Jones protein, which is found in the urine of patients suffering from myeloma tumours, is identical with the *L* chains of the myeloma protein. Each patient has a different Bence-Jones protein; no two of the more than 100 Bence-Jones proteins that have been analyzed thus far are identical. It is thought that one lymphoid cell among hundreds of thousands becomes malignant and multiplies rapidly, forming the mass of a myeloma tumour that produces one γ-globulin.

Bence-Jones proteins

Analyses of the Bence-Jones proteins have revealed that the *L* chains of man and other mammals are of two quite different types, kappa (κ) and lambda (λ). Both consist of approximately 220 amino acids. The N–terminal halves of κ- and λ-chains are variable, differing in each Bence-Jones protein. The C–terminal halves of these same *L* chains have a constant amino acid sequence of either the κ- or the λ-type. The fact that one half of a peptide chain is variable and the other half invariant is contradictory to the view that the amino acid sequence of each peptide chain is determined by one gene (see GENETICS AND HEREDITY: *The gene*). Evidently, two genes, one of them variable, the other invariant, fuse to form the gene for the single peptide chain of the *L* chains. Whereas the normal human *L* chains are always mixtures of the κ- and λ-types, the *H* chains of IgG, IgM, and IgA are different. They have been designated as gamma (γ), mu (μ), and alpha (α) chains, respectively. The N-terminal quarter of the *H* chains has a variable amino acid sequence; the C-terminal three-quarters of the *H* chains have a constant amino acid sequence, as indicated in Figure 6.

Some of the amino acid sequences in the *L* and *H* chains are transmitted from generation to generation. As a result, the constant portion of the human *L* chains of the κ-type has in position 191 either valine or leucine. They correspond to two alleles (character-determining portions) of a gene; the two types are called allotypes. The valine-containing genetic type has been designated as $InV(a^+)$, the leucine-containing type as $InV(b^+)$. Many more allotypes, called Gm allotypes, have been found in the gamma chains of the human IgG immunoglobulins; more than 20 Gm allotypes are now known. Certain combinations of Gm types occur; the combination of Gm types 5, 6, and 11 has been found in Caucasians and Negroes but not in Chinese; the combination of 1, 2, and 17 has not been found in Negroes; and the combination of 1, 4, and 17 has not been found in Caucasians. Allotypes have also been discovered to occur in a number of other animals, including rabbits and mice.

It is understandable from the occurrence of a large number of allotypes that antibodies, even if produced in response to a single antigen, are mixtures of different allotypes. The existence of several classes of antibodies, of different allotypes, and of adaptation of the variable portions of antibodies to different regions of an antigen molecule results in a multiplicity of antibody molecules even if only a single antigen is administered. For this reason it has not yet been possible to unravel the amino acid sequence in the variable portion of antibody molecules. Much of the amino acid sequence in the constant regions of the *L* and *H* chains of man and rabbit immunoglobulins, however, has been resolved.

(F.Ha./Ed.)

ENZYMES

Practically all of the numerous and complex biochemical reactions that take place in animals, plants, and microorganisms are regulated by enzymes. These catalytic proteins are efficient and specific—that is, they accelerate the rate of one kind of chemical reaction of one type of compound, and they do so in a far more efficient manner than man-made catalysts. They are controlled by activators and inhibitors that initiate or block reactions. All cells contain enzymes, which usually vary in number and composition, depending on the cell type; an average mammalian cell, for example, is approximately one one-billionth (10^{-9}) the size of a drop of water and generally contains about 3,000 enzymes.

The existence of enzymes was established in the middle of the 19th century by scientists studying the process of fermentation. Their role as catalysts of all living things followed rapidly. Developments before 1850 included (in 1833) the separation from malt of the enzyme amylase, which converts starch into sugar, and (in 1836) the isolation from the stomach wall of animals of a component of gastric juice that could partially digest food in a test tube, the enzyme pepsin.

Enzymes were known for many years as ferments, a term derived from the Latin word for yeast. In 1878 the name enzyme, from the Greek words meaning "in yeast," was introduced; since the late 19th century it has been universally used.

Role of enzymes in metabolism. Some enzymes help to break down large nutrient molecules, such as proteins, fats, and carbohydrates, into smaller molecules. This process occurs during the digestion of foodstuffs in the stomach and intestines of animals. Other enzymes guide the smaller, broken-down molecules through the intestinal wall into the bloodstream. Still other enzymes promote the formation of large, complex molecules from the small, simple ones to produce cellular constituents. Enzymes are also responsible for numerous other functions, which include the storage and release of energy, the course of reproduction, the processes of respiration, and vision. They are indispensable to life.

Each enzyme is able to promote only one type of chemical reaction. The compounds on which the enzyme acts are called substrates. Enzymes operate in tightly organized metabolic systems called pathways. A seemingly simple

Metabolic pathways

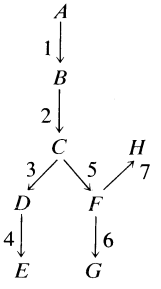
biological phenomenon—the contraction of a muscle, for example, or the transmission of a nerve impulse—actually involves a large number of chemical steps in which one or more chemical compounds (substrates) are converted to substances called products; the product of one step in a metabolic pathway serves as the substrate for the succeeding step in the pathway.

The role of enzymes in metabolic pathways can be illustrated diagrammatically. The chemical compound represented by *A* (see diagram) is converted to product *E* in a series of enzyme-catalyzed steps, in which intermediate compounds represented by *B, C,* and *D* are formed in succession. They act as substrates for enzymes represented by 2, 3, and 4. Compound *A* may also be converted by another series of steps, some of which are the same as those in the pathway for the formation of *E,* to products represented by *G* and *H.*

The letters represent chemical compounds; numbers represent enzymes that catalyze individual reactions. The relative heights represent the thermodynamic energy of the compounds; *e.g.,* compound *A* is more energy-rich than *B, B* more energy-rich than *C.* Compounds *A, B,* etc., change very slowly in the absence of a catalyst but do so rapidly in the presence of catalysts 1, 2, 3, etc.

The regulatory role of enzymes in metabolic pathways can be clarified by using a simple analogy: that between the compounds, represented by letters in the diagram, and a series of connected water reservoirs on a slope. Similarly, the enzymes represented by the numbers are analogous to the valves of the reservoir system. The valves control the flow of water in the reservoir; that is, if only valves 1, 2, 3, and 4 are open, the water in *A* flows only to *E,* but, if valves 1, 2, 5, and 6 are open, the water in *A* flows to *G.* In a similar manner, if enzymes 1, 2, 3, and 4 in the metabolic pathway are active, product *E* is formed, and, if enzymes 1, 2, 5, and 6 are active, product *G* is formed. The activity or lack of activity of the enzymes in the pathway therefore determines the fate of compound *A; i.e.,* it either remains unchanged or is converted to one or more products. In addition, if products are formed, the activity of enzymes 3 and 4 relative to that of enzymes 5 and 6 determines the quantity of product *E* formed compared with product *G.*

Both the flow of water and the activity of enzymes obey the laws of thermodynamics; hence, water in reservoir *F*



cannot flow freely to *H* by opening valve 7, because water cannot flow uphill. If, however, valves 1, 2, 5, and 7 are open, water flows from *F* to *H,* because the energy conserved during the downhill flow of water through valves 1, 2, and 5 is sufficient to allow it to force the water up through valve 7. In a similar way, enzymes in the metabolic pathway cannot convert compound *F* directly to *H* unless energy is available; enzymes are able to utilize energy from energy-conserving reactions in order to catalyze reactions that require energy. During the enzyme-catalyzed oxidation of carbohydrates to carbon dioxide and water, energy is conserved in the form of an energy-rich compound, adenosine triphosphate (ATP). The energy in ATP is utilized during an energy-consuming process such as the enzyme-catalyzed contraction of muscle.

Because the needs of cells and organisms vary, not only the activity but also the synthesis of enzymes must be regulated; *e.g.,* the enzymes responsible for muscular activity in a leg muscle must be activated and inhibited at appropriate times. Some cells do not need certain enzymes; a liver cell, for example, does not need a muscle enzyme. A bacterium does not need enzymes to metabolize substances that are not present in its growth medium. Some enzymes, therefore, are not formed in certain cells, others are synthesized only when required, and still others are found in all cells. The formation and activity of enzymes are regulated not only by genetic mechanisms but also by organic secretions (hormones) from endocrine glands and by nerve impulses. Small molecules also play an important role (see below *Enzyme flexibility and allosteric control*).

The significance of enzymes in disease

If an enzyme is defective in some respect, disease may occur. The enzymes represented by the numbers 1 to 4 in the diagram must function during the conversion of the starting substance *A* to the product *E.* If one step is blocked because an enzyme is unable to function, product *E* may not be formed; if *E* is necessary for some vital function, disease results. Many inherited diseases of man result from a deficiency of one enzyme. Some of these are listed in Table 3. The disease called albinism, for example, results from an inherited lack of ability to synthesize the enzyme tyrosinase, which catalyzes one step in the pathway by which the pigment for hair and eye colour is formed (see also METABOLISM).

Other functions. Enzymes play an increasingly important role in medicine. The enzyme thrombin is used to promote the healing of wounds. Other enzymes are used to diagnose certain kinds of disease, to cause the remission of some forms of leukemia—a disease of the blood-forming organs—and to counteract unfavourable reactions in people who are allergic to penicillin. The enzyme lysozyme, which destroys cell walls, is used to kill bacteria. Research concerning medical applications of enzymes may lead to their use as preventives of tooth decay and as anticoagulants in the treatment of thrombosis, a disease characterized by the formation of a clot, or plug, in a blood vessel. Enzymes may eventually be used to control enzyme deficiencies and abnormalities resulting from diseases.

Table 3: Enzymes Identified with Hereditary Diseases

disease name	defective enzyme
Albinism	tyrosinase
Phenylketonuria	phenylalanine hydroxylase
Fructosuria	fructokinase
Methemoglobinemia	methemoglobin reductase
Galactosemia	galactose-1-phosphate uridyl transferase

It might also be noted in passing that enzymes are used in industrial processes involving the preparation of certain chemical compounds and the tanning of leather; they are valuable in analytical procedures involving the detection of very small quantities of specific substances. Enzymes are necessary in such food-related industries as cheese making, the brewing of beer, the aging of wine, and the baking of bread. Enzymes also may be used to clean clothes. For industrial use of enzymes, see FOOD PROCESSING: *Baking and bakery products* and BEVERAGE PRODUCTION: *Brewing; Wine making.*

General properties. *Classification and nomenclature.* The first enzyme name, proposed in 1833, was diastase. Sixty-five years later, it was suggested that all enzymes be named by adding "-ase" to a root indicative of the nature of the substrate of the enzyme. Although enzymes are no longer named in such a simple manner, with the exception of a few—*e.g.,* pepsin, trypsin, chymotrypsin, papain—most enzyme names do end in "-ase."

Any systematic classification of enzymes should be based on a common property or quality that varies sufficiently to be useful as a distinguishing feature. In this regard, three properties of enzymes could serve as a basis for enzyme classification—the exact chemical nature of the enzyme, the chemical nature of the substrate, and the nature of the reaction catalyzed. Adequate information about the detailed chemical structures of more than a few enzymes does not yet exist. In addition, although, as indicated above, early attempts at enzyme classification were based on the nature of broad groups of substrates (*e.g.,* enzymes called carbohydrases act on carbohydrates), close functional similarities among enzymes in different groups were often obscured. By general agreement, enzymes now

Table 4: Classification of Some Enzymes

systematic name*		trivial name	reaction catalyzed	biological role
code number†	name‡			
1.1.1.1	alcohol: NAD oxidoreductase	alcohol dehydrogenase	alcohol + NAD → acetaldehyde NADH	alcoholic fermentation
1.1.1.27	L-lactate: NAD oxidoreductase	lactic dehydrogenase	lactate + NAD → pyruvate + NADH	carbohydrate metabolism
2.7.1.40	ATP: pyruvate phospho transferase	pyruvate kinase	pyruvic acid + ATP → phosphoenol-pyruvic acid + ADP	carbohydrate metabolism
3.1.1.7	acetylcholine: acetyl hydrolase	acetylcholinesterase	acetylcholine + H_2O → acetate + choline	nerve-impulse conduction
3.4.4.13	peptide peptido hydrolase	thrombin	peptide bond + $H_2O \rightarrow -\overset{O}{\overset{\|}{C}}-OH + H_3^+N-$	blood-clotting mechanism

*Based on recommendations (1964) of the International Union of Biochemistry. †The numbering system is as follows: the first number places the enzyme in one of six general groups—1, oxidoreductases; 2, transferases; 3, hydrolases; 4, lyases; 5, isomerases and 6, ligases. The second number places the enzyme in a subclass based on substrate type or reaction type; *e.g.*, the enzyme may act on molecules with —CHOH groups. The third number places the enzyme in a subsubclass, which specifies the reaction type more fully; *e.g.*, NAD coenzyme required. The fourth number is the serial number of the enzyme in its subsubclass. ‡NAD and NADH represent the oxidized and reduced forms of nicotinamide adenine dinucleotide (NAD), respectively; ATP and ADP represent adenosine triphosphate and adenosine diphosphate, respectively.

are classified according to their substrates and the nature of the reaction they catalyze.

Systematic and trivial names

In an attempt to devise a rational system of enzyme nomenclature, two names are given to an enzyme. One, known as the systematic name, is based on logical principles but is often long and awkward; the other, "trivial" name is short and generally used but not usually exact or systematic (see Table 4). In the scheme of systematic nomenclature, six main groups of enzymatic reactions are recognized; each catalyzes one reaction type and is subdivided on the basis of detailed definitions of the reaction catalyzed and of the substrate involved in the reaction. Enzymes that catalyze reactions in which hydrogen is transferred belong to the group known as oxidoreductases; those that catalyze the introduction of the elements of water at a specific site in a molecule are called hydrolases. The other four groups of reactions are the transferases—which catalyze reactions in which substances other than hydrogen are transferred—the lyases, the isomerases, and the ligases. Oxidoreductases and transferases account for about 50 percent of the approximately 1,000 enzymes recognized thus far. Table 4 lists a few enzymes, their trivial names, their systematic names, and their biological roles.

Chemical nature. Little was known about the chemical nature of enzymes until the beginning of the 20th century, although scientists were almost convinced that they were proteins. In 1926, the enzyme urease was the first to be crystallized and clearly identified as a protein. Within the next few years, the digestive enzymes pepsin, trypsin, and chymotrypsin were shown to be proteins. Since that time, hundreds of enzymes, all of them proteins, have been prepared and characterized by chemical methods. Much of the knowledge of protein chemistry has, in fact, resulted from studies involving enzymes and from attempts to understand their nature and mode of action.

Although some enzymes consist of a single chain of the amino acids (*i.e.*, simple organic molecules containing nitrogen), most enzymes are composed of more than one chain. Each chain is called a subunit. Many enzymes have two, four, or six subunits, and some consist of as many as 12 to 60 subunits. In many cases, the subunits have identical structures; in others, however, several different types of subunit chains are involved.

With the exception of proteins that act as structural elements, most of the proteins in physiologically active tissues such as kidney and liver are enzymes. Regardless of the exact amount of enzymatic protein in an organism, it is clear that hundreds of different enzymes must be present in each tissue to account for the myriad reactions comprising metabolism.

Cofactors. Although some enzymes consist only of protein, many are complex proteins; *i.e.*, they have a protein component and a so-called cofactor. A complete enzyme is called a holoenzyme; if the cofactor is removed, the protein, no longer enzymatically active, is called the apoenzyme. A cofactor may be a metal—such as iron, copper, or magnesium—a moderately sized organic molecule called a prosthetic group, or a special type of substrate molecule known as a coenzyme. The cofactor may aid in the catalytic function of an enzyme, as do metals and prosthetic groups, or take part in the enzymatic reaction, as do coenzymes.

Holoenzyme

Coenzymes

A coenzyme serves as a type of substrate in certain enzymatic reactions and thus reacts in the exact proportions (*i.e.*, stoichiometrically) required for reaction, rather than in catalytic quantities. A coenzyme may, for example, assume the role of a hydrogen acceptor, as does nicotinamide adenine dinucleotide (NAD), which accepts hydrogen from the substrate, or a chemical-group donor, as does adenosine triphosphate (ATP), which donates phosphoric acid to the substrate. After ATP has donated a phosphoric acid molecule to the substrate, the phosphoric acid can be reacquired in a second stoichiometric reaction catalyzed by a second enzyme. The catalytic nature of a coenzyme is apparent only when it couples the activities of two enzymes in this way. Coenzymes thus are the links, or shuttles, in metabolic pathways that enable substances—*e.g.*, hydrogen, phosphoric acid—to be exchanged.

The nature of enzyme-catalyzed reactions. *The nature of catalysis.* In a chemical reaction—for example, one in which substance *A* is converted into product *B*—a point of equilibrium eventually is reached at which no further chemical change occurs; *i.e.*, the rate of conversion of *A* to *B* equals the rate of conversion of *B* to *A*. The so-called thermodynamic-equilibrium constant expresses this chemical equilibrium. A catalyst may be defined as a substance that accelerates a chemical reaction but is not consumed in the process. The amount of catalyst has no relationship to the quantity of substance altered; very small amounts of enzymes are very efficient catalysts. Because the presence of an enzyme accelerates the rate of conversion of a compound to a product, it accelerates the approach to equilibrium; it does not, however, influence the equilibrium point attained.

The molecules in the watery medium of the cell are in constant thermal motion but, because they are more or less stable compounds, they would react only occasionally to form products in the absence of enzymes. There exists an energy barrier to the reaction of a molecule. The energy required to overcome the barrier to reaction is called the energy of activation. A reaction proceeds to equilibrium only if the molecules have sufficient energy of activation to form an activated complex, from which products can be derived. Enzymes greatly increase the chances for reactions by their ability to make large numbers of specific molecules more reactive (*i.e.*, unstable) by forming intermediate compounds with them. The unstable intermediates quickly break down to form stable products, and the enzymes, unchanged by the reaction, are able to catalyze the formation of additional products.

The role of the active site. That the compound on which an enzyme acts (substrate) must combine in some

way with it before catalysis can proceed is an old idea, now supported by much experimental evidence. The combination of substrate molecules with enzymes involves collisions between the two. Enzymes are large molecules, the molecular weights of which (based on the weight of a hydrogen atom as 1) range from several thousand to several million. The substrates on which enzymes act usually have molecular weights of several hundred. Because of the difference in size between the two, only a fraction of the enzyme is in contact with the substrate; the region of contact is called the active site. Usually, each subunit of an enzyme has one active site capable of binding substrate.

The characteristics of an enzyme derive from the sequence of amino acids, which determine the shape of the enzyme (*i.e.*, the structure of the active site) and hence the specificity of the enzyme.

Figure 7A shows the way in which the substrate fits into the active site of an enzyme. A small portion of the structure of the enzyme is represented by the various dark lines and symbols. The triangles labeled *A* and *B* represent the amino acids at the active site that actually catalyze the reaction in which the bond (indicated by the indentation) in the substrate molecule is broken as a product forms. The circles in the enzyme structure represent amino acids that function in attracting certain regions of the substrate, so that it is held in the proper position to allow the reaction at the active site to proceed.

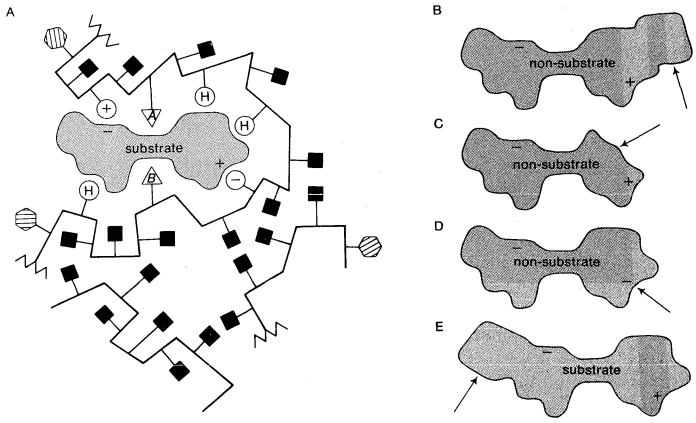


Figure 7: The role of active site in lock and key fit of a substrate—the key—to an enzyme—the lock (see text).

Forces attracting enzyme and substrate

The forces that attract the substrate to the surface of an enzyme may be of a physical or a chemical nature. Electrostatic bonds may occur between oppositely charged groups—the circles containing plus and minus signs on the enzyme are attracted to their opposites in the substrate molecule. Such electrostatic bonds can occur with groups that are completely positively or negatively charged (*i.e.*, ionic groups) or with groups that are partially charged (*i.e.*, dipoles). The attractive forces between substrate and enzyme may also involve so-called hydrophobic bonds, in which the oily, or hydrocarbon, portions of the enzyme (represented by H-labelled circles) and the substrate are forced together in the same way as oil droplets tend to coalesce in water.

Modifications in the structure of the amino acids at or near the active site usually affect the enzyme's activity, because these amino acids are intimately involved in the fit and attraction of the substrate to the enzyme surface. The characteristics of the amino acids near the active site determine whether or not a substrate molecule will fit into the site. A molecule that is too bulky in the wrong places, as indicated by the arrow in Figure 7B, cannot fit into the active site and thus cannot react with the enzyme. In a similar manner, a molecule lacking essential attractive forces (indicated by the arrow in 7C) or the appropriately charged regions (indicated by the arrow in 7D) might not be bound to the enzyme. On the other hand, a molecule with a bulky group (7E) at a position such that it does not interfere with the binding of the molecule to the enzyme or with the function of the active site is able to serve as a substrate for the enzyme. The idea of a fit between substrate and enzyme, called the "key–lock" hypothesis, was proposed by a German chemist, Emil Fischer, in 1899 and explains one of the most important features of enzymes, their specificity. In most of the enzymes studied thus far, a cleft, or indentation, into which the substrate fits (see 7A), is found at the active site.

The specificity of enzymes. Since the substrate must fit into the active site of the enzyme before catalysis can occur, only properly designed molecules can serve as substrates for a specific enzyme; in many cases, an enzyme will react with only one naturally occurring molecule. Two oxidoreductase enzymes from Table 4 will serve to illustrate the principle of enzyme specificity. One (alcohol dehydrogenase) acts on alcohol, the other (lactic dehydrogenase) on lactic acid; the activities of the two, even though both are oxidoreductase enzymes, are not interchangeable—*i.e.*, alcohol dehydrogenase will not catalyze a reaction involving lactic acid or vice versa, because the structure of each substrate differs sufficiently to prevent its fitting into the active site of the alternative enzyme. Enzyme specificity is essential because it keeps separate the many pathways, involving hundreds of enzymes, that function during metabolism.

Figure 7 illustrates enzyme specificity. As mentioned above, the molecules in B, C, and D cannot serve as substrates for the enzyme because they are either too large (B), too small (C), or too repulsive because of charge (D) to bind to the enzyme's active site. Molecules with structural differences that do not affect the active site (see arrow, Figure 7E) are able to react with the enzyme and are substrates.

Not all enzymes are as highly specific as the example in Figure 7; digestive enzymes such as pepsin and chymotrypsin, for example, are able to act on almost any protein, as they must if they are to act upon the varied types of proteins consumed as food. On the other hand, thrombin, which reacts only with the protein fibrinogen, is part of a very delicate blood-clotting mechanism and thus must act only on one compound in order to maintain the proper functioning of the system.

When enzymes were first studied, it was thought that most of them were "absolutely specific"—*i.e.*, that they would react with only one compound. In most cases, however, a molecule other than the natural substrate can be synthesized in the laboratory; it is enough like the natural substrate to react with the enzyme. Use of these synthetic substrates has been valuable in understanding enzymatic action. It must be remembered, however, that, in the living cell, many enzymes are absolutely specific for the compounds found there.

All enzymes isolated thus far are specific for the type of chemical reaction they catalyze—*i.e.*, oxidoreductases do not catalyze hydrolase reactions, and hydrolases do not catalyze reactions involving oxidation and reduction. An enzyme therefore catalyzes a specific chemical reaction but may be able to do so on several similar compounds.

The mechanism of enzymatic action. An enzyme attracts substrates to its active site, catalyzes the chemical reaction by which products are formed, and then allows

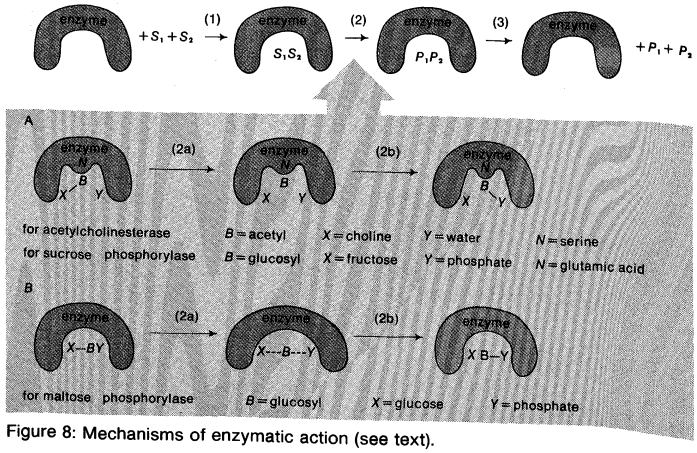


Figure 8: Mechanisms of enzymatic action (see text).

the products to dissociate—*i.e.*, separate from the enzyme surface. This sequence is shown in Figure 8. The combination formed by an enzyme and its substrates is called the enzyme–substrate complex. When two substrates and one enzyme are involved, the complex is called a ternary complex; one substrate and one enzyme are called a binary complex. The substrates are attracted to the active site by electrostatic and hydrophobic forces, which are called noncovalent bonds because they are physical attractions and not chemical bonds (see above *The role of the active site*).

Enzyme–substrate complex

In Figure 8, two substrates (S_1 and S_2) bind to the active site of the enzyme during step (1) and react to form products (P_1 and P_2) during step (2). The products dissociate from the enzyme surface in step (3), releasing the enzyme. The enzyme, unchanged by the reaction, is able to react with additional substrate molecules in this manner many times per second to form products. The step in which the actual chemical transformation occurs is of great interest, and, although much is known about it, it is not yet fully understood. In general there are two types of enzymatic mechanisms, one in which a so-called covalent intermediate forms (see below) and one in which none forms (Figure 8B). Figure 8A and B show the catalytic events that occur during an enzyme-catalyzed reaction.

Figure 8A shows the mechanism by which a covalent intermediate—*i.e.*, an intermediate with a chemical bond between substrate and enzyme—forms. One substrate, $B-X$ (in 8A), reacts with the group N on the enzyme surface to form an enzyme-B intermediate compound. The intermediate compound then reacts with the second substrate, Y, to form the products $B-Y$ and X.

Many enzymes catalyze reactions by this type of mechanism. Acetylcholinesterase (Table 4 and Figure 8A) is used as a specific example in the sequence described below. The two substrates (S_1 and S_2) for acetylcholinesterase are acetylcholine (the $B-X$ of Figure 8A) and water (the Y of Figure 8A). After acetylcholine ($B-X$) binds to the enzyme surface, a chemical bond forms between the acetyl moiety (B) of acetylcholine and the group N (part of the amino acid serine) on the enzyme surface. The result of the formation of this bond, called an acyl–serine bond, is one product, choline (X), and the enzyme-B intermediate compound (an acetyl–enzyme complex). The water molecule (Y) then reacts with the acyl–serine bond to form the second product, acetic acid ($B-Y$), which dissociates from the enzyme. Acetylcholinesterase is regenerated and is again able to react with another molecule of acetylcholine. This kind of reaction, involving the formation of an intermediate compound on the enzyme surface, is generally called a double displacement reaction.

Double displacement reaction

Sucrose phosphorylase (see Figure 8A) acts in a similar way. The substrate for sucrose phosphorylase is sucrose, or glucosyl-fructose ($B-X$), and the group N on the enzyme surface is a chemical group called a carboxyl group (COOH). The enzyme-B intermediate, a glucosyl–carboxyl compound, reacts with phosphate (Y) to form glucosyl-phosphate ($B-Y$). The other product (X) is fructose.

In double displacement reactions, the covalent intermediate between enzyme and substrate apparently influences the reaction to proceed more rapidly. Because the enzyme is unaltered at the end of the reaction, it functions as a true catalyst, even though it is temporarily altered during the enzymatic process.

Although many enzymes form a covalent intermediate, the mechanism is not essential for catalysis; some reactions proceed as illustrated in Figure 8B. One substrate (Y) reacts directly with the second substrate ($X-B$), in a so-called single displacement reaction. The B moiety, which is transformed in the chemical reaction, is involved in only one reaction and does not form a bond with a group on the enzyme surface. In Figure 8B, the enzyme maltose phosphorylase directly affects the bonds of the substrates ($B-X$ and Y), which, in this case, are maltose (glucosyl-glucose) and phosphate, to form the products, glucose (X) and glucosylphosphate ($B-Y$).

Covalent intermediates between part of a substrate and an enzyme occur in many enzymatic reactions, and various amino acids—serine, cysteine, lysine, and glutamic acid—are involved.

The rate of enzymatic reactions. *The Michaelis-Menten hypothesis.* If the velocity of an enzymatic reaction is represented graphically as a function of the substrate concentration (S), the curve obtained in most cases is a hyperbola (see Figure 9). The mathematical expression of this curve, shown in the equation below, was developed in 1913 by two German biochemists, L. Michaelis and M.L. Menten. In the equation, V_M is the maximal velocity of the reaction, and K_M is called the Michaelis constant,

$$\text{velocity} = \frac{V_M(S)}{K_M + (S)} .$$

The shape of the curve is a logical consequence of the active-site concept; *i.e.*, the curve flattens at the maximum velocity (V_M), which occurs when all the active sites of the enzyme are filled with substrate. The fact that the velocity approaches a maximum at high substrate concentrations provides support for the assumption that an intermediate enzyme–substrate complex forms. At the point of half the maximum velocity ($\frac{V_M}{2}$ in Figure 9), the substrate concentration in moles per litre (M) is equal to the Michaelis constant, which is a rough measure of the affinity of the substrate molecule for the surface of the enzyme. K_M values usually vary from about 10^{-8} to 10^{-2} M, and V_M from 10^5 to 10^9 molecules of product formed per molecule of enzyme per second. The value for V_M is referred to as the turnover number when expressed as moles of product formed per mole of enzyme per minute. The binding of molecules that inhibit or activate the protein surface usually results in similar types.

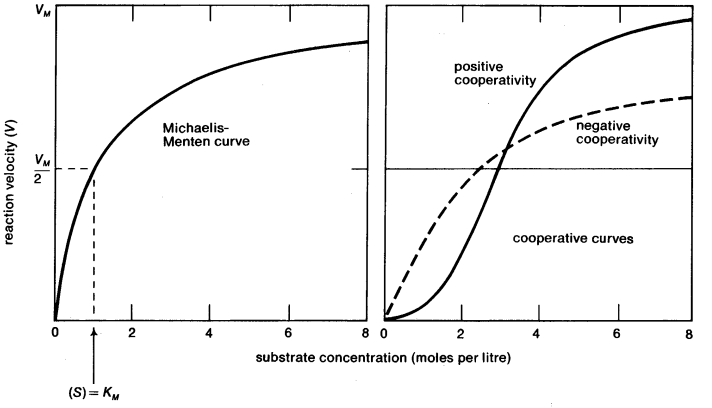


Figure 9: Curves representing enzyme action (see text).

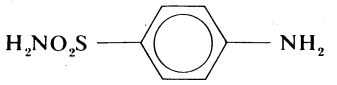
The efficiency of enzymes

Enzymes are more efficient than man-made catalysts operating under the same conditions. Because many enzymes with different specificities occur in a cell, adequate space exists only for a few enzyme molecules catalyzing one specific reaction. Each enzyme, therefore, must be very efficient. One molecule of the enzyme catalase, for example, can produce 10^{12} molecules of oxygen per second. The catalytic groups at the active site of an enzyme act 10^6 to 10^9 times more effectively than do analogous groups in a nonenzymatic reaction.

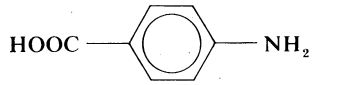
The reason for the great efficiency of enzymes is not completely understood. It results in part from the precise positioning of the substrates and the catalytic groups at the active site, which serves to increase the probability of collision between the reacting atoms. In addition, the environment at the active site may be favourable for reaction—that is, acidic and basic groups may act together more effectively there, or some strain may be induced in the substrate molecules so that their bonds are broken more easily, or the orientation of the reacting substrates may be optimal at the enzyme surface. The theories that have been formulated to account for the high catalytic efficiency of enzymes, although reasonable, still remain to be proved.

Inhibition of enzymes. Some molecules very similar to the substrate for an enzyme may be bound to the active site but be unable to react. Such molecules cover the active site and thus prevent the binding of the actual substrate to the site. This inhibition of enzyme action is of

a competitive nature, because the inhibitor molecule actually competes with the substrate for the active site. The inhibitor sulfanilamide (see below), for example, is similar enough to a substrate (*p*-aminobenzoic acid) of an enzyme involved in the metabolism of folic acid that it binds to the enzyme but cannot react. It covers the active site and prevents the binding of *p*-aminobenzoic acid. This enzyme is essential in certain disease-causing bacteria but is not essential to man; large amounts of sulfanilamide therefore kill the microorganism but do not harm man. Inhibitors such as sulfanilamide are called anti-metabolites. Sulfanilamide and similar compounds that kill a pathogen without harming its host are now widely used in chemotherapy.



sulfanilamide
(inhibitor)



p-aminobenzoic acid
(substrate)

Some inhibitors prevent, or block, enzymatic action by reacting with groups at the active site. The nerve gas diisopropyl fluorophosphate, for example, reacts with the serine at the active site of acetylcholinesterase to form a covalent bond. The nerve-gas molecule involved in bond formation prevents the active site from binding the substrate, acetylcholine, thereby blocking catalysis and nerve action. Iodoacetic acid similarly blocks a key enzyme in muscle action by forming a bulky group on the amino acid cysteine, which is found at the enzyme's active site. This process is called irreversible inhibition.

Some inhibitors modify amino acids other than those at the active site, resulting in loss of enzymatic activity. The inhibitor causes changes in the shape of the active site. Some amino acids other than those at the active site, however, can be modified without affecting the structure of the active site; in these cases, enzymatic action is not affected.

Such chemical changes parallel natural mutations. Inherited diseases frequently result from a change in an amino acid at the active site of an enzyme, thus making the enzyme defective. In some cases, an amino acid change alters the shape of the active site to the extent that it can no longer react; such diseases are usually fatal. In others, however, a partially defective enzyme is formed, and an individual may be very sick but able to live.

Denaturation of enzymes

Effects of temperature. Enzymes function most efficiently within a physiological temperature range. Since enzymes are protein molecules, they can be destroyed by high temperatures. An example of such destruction, called protein denaturation, is the curdling of milk when it is boiled. Increasing temperature has two effects on an enzyme. First, the velocity of the reaction increases somewhat, because the rate of chemical reactions tends to increase with temperature; second, the enzyme is increasingly denatured. Increasing temperature thus increases the metabolic rate only within a limited range. If the temperature becomes too high, enzyme denaturation destroys life. Low temperatures also change the shapes of enzymes. With enzymes that are cold-sensitive, the change causes loss of activity. Both excessive cold and heat are therefore damaging to enzymes.

The degree of acidity or basicity of a solution, which is expressed as pH, also affects enzymes. As the acidity of a solution changes—*i.e.,* the pH is altered—a point of optimum acidity occurs, at which the enzyme acts most efficiently. Although this pH optimum varies with temperature and is influenced by other constituents of the solution containing the enzyme, it is a characteristic property of enzymes. Because enzymes are sensitive to changes in acidity, most living systems are highly buffered; *i.e.,* they have mechanisms that enable them to maintain a constant acidity. This acidity level, or pH, is about 7 in most organisms. Some bacteria function under moderately acidic or basic conditions; and the digestive enzyme pepsin acts in the acid milieu of the stomach. There is no known organism that can survive in either a very acidic or a very basic environment.

Enzyme flexibility and allosteric control. *The induced-fit theory.* The key–lock hypothesis (see above *The nature of enzyme-catalyzed reactions*) does not fully account for enzymatic action; *i.e.,* certain properties of enzymes cannot be accounted for by the simple relationship between enzyme and substrate proposed by the key–lock hypothesis. A theory called the induced-fit theory retains the key–lock idea of a fit of the substrate at the active site but postulates in addition that the substrate must do more than simply fit into the already preformed shape of an active site. Rather, the theory states, the binding of the substrate to the enzyme must cause a change in the shape of the enzyme that results in the proper alignment of the catalytic groups on its surface. This concept has been likened to the fit of a hand in a glove, the hand (substrate) inducing a change in the shape of the glove (enzyme). Although some enzymes appear to function according to the older key–lock hypothesis, most apparently function according to the induced-fit theory.

During step 1 in Figure 10, which illustrates the induced-fit theory, the substrate approaches the enzyme surface and induces a change in its shape that results in the correct alignment of the catalytic groups (indicated by triangles *A* and *B*). In the case of the digestive enzyme carboxypeptidase, the binding of the substrate causes a tyrosine molecule at the active site to move by as much as 15 angstroms. Circles *C* and *D* in the figure represent substrate-binding groups on the enzyme that are essential for catalytic activity. During step 2 the catalytic groups at the active site react with the substrate to form products. The products separate from the enzyme surface during step 3, and the enzyme is able to repeat the sequence.

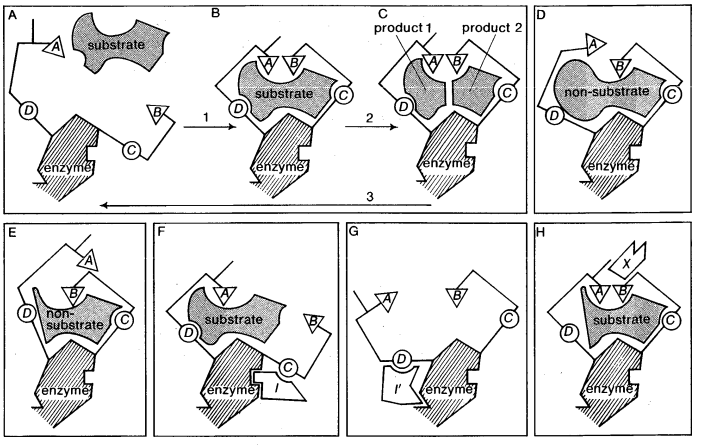


Figure 10: Induced-fit binding of a substrate to an enzyme surface and allosteric effects (see text).

Nonsubstrate molecules that are too bulky (Figure 10D) or too small (Figure 10E) alter the shape of the enzyme so that a misalignment of catalytic groups *A* and *B* occurs; such molecules are not able to react even if they are attracted to the active site.

The induced-fit theory explains a number of anomalous properties of enzymes; for example, "noncompetitive inhibition" (Figure 10F), in which a compound inhibits the reaction of an enzyme but does not prevent the binding of the substrate. In this case, the inhibitor compound *I* attracts the binding group *C* so that the catalytic group *B* is too far away from the substrate to react. The site at which the inhibitor binds to the enzyme is not the active site and is called an allosteric site. The inhibitor changes the shape of the active site to prevent catalysis without preventing binding of the substrate.

Figure 10G shows the effect of an inhibitor (*I'*), which distorts the active site by affecting the essential binding group *D;* as a result, the enzyme can no longer attract the

substrate. In Figure 10H, a so-called activator molecule, *X*, affects the active site so that a nonsubstrate molecule is properly aligned and hence can react with the enzyme; *X* is called an allosteric activator of the reaction. Such activators can affect both binding and catalytic groups at the active site.

The importance of enzyme flexibility

Enzyme flexibility is extremely important because it provides a mechanism for regulating enzymatic activity. As shown in Figure 10F and G, the orientation at the active site can be disrupted by the binding of an inhibitor at a site other than the active site. Moreover, the enzyme can be activated by molecules that induce a proper alignment of the active site for a substrate that alone cannot induce this alignment (Figure 10H).

As mentioned above, the sites that bind inhibitors and activators are called allosteric sites to distinguish them from active sites. Allosteric sites are in fact regulatory sites able to activate or inhibit enzymatic activity by influencing the shape of the enzyme. When the activator or inhibitor dissociates from the enzyme, it returns to its normal shape. Thus, the flexibility of the protein structure allows the operation of a simple, reversible control system similar to a thermostat.

Types of allosteric control. Allosteric control can operate in many ways; two examples serve to illustrate some general effects. A pathway consisting of ten enzymes is involved in the synthesis of the amino acid histidine. When a cell contains enough histidine, synthesis stops—an appropriate economy move by the cell. Synthesis is stopped by the inhibition of the first enzyme in the pathway by the product, histidine; the mechanism is similar to that of Figure 10G. The inhibition of an enzyme by a product is called feedback inhibition; *i.e.,* a product many steps removed from an initial enzyme blocks its action. Feedback inhibition occurs in many pathways in all living things.

Allosteric control can also be achieved by activators. The hormone adrenaline (epinephrine) acts in this way. When energy is needed, adrenaline is released and activates, by allosteric activation, the enzyme adenyl cyclase. This enzyme catalyzes a reaction in which the compound cyclic adenosine monophosphate (cyclic AMP) is formed from ATP. Cyclic AMP in turn acts as an allosteric activator of enzymes that speed the metabolism of carbohydrate to produce energy. This type of allosteric regulation also is widespread in biological systems.

Thus, a combination of allosteric activation and inhibition allows the production of energy or materials when they are needed and shuts off production when the supply is adequate.

Allosteric control is a rapid method of regulating products continuously needed by living things. Yet some cells have no need for certain enzymes, and it would be wasteful for the cell to synthesize them. In this case, certain molecules, called repressors, prevent the synthesis of unneeded enzymes. The repressors are proteins that bind to DNA and prevent the first step in the process resulting in protein synthesis. If certain metabolites are added to cells that need an enzyme, enzyme synthesis occurs—*i.e.,* it is induced. Addition of galactose to a growth medium containing *Escherichia coli* bacteria, for example, induces the synthesis of the enzyme beta galactosidase. The bacteria thus can synthesize this galactose-metabolizing enzyme when it is needed and prevent its synthesis when it is not. The way in which the synthesis of enzymes is induced or repressed in mammalian systems is less understood but is believed to be similar.

Different types of cells in complex organisms have different enzymes, even though they have the same DNA content. The enzymes actually synthesized are the ones needed in a specific cell and vary not only for different types of cells—*e.g.,* nerve, muscle, eye, and skin cells—but also for different species.

In an enzyme consisting of several subunits, or chains, alteration in the shape of one chain as a result of the influence either of a substrate molecule or of allosteric inhibitors or activators may change the shape of a neighbouring chain. As a result, the binding of a second molecule of substrate occurs in a different way from the binding of the first, and the third is different from the second. This phenomenon, called cooperativity, is characteristic of allosteric enzymes.

Cooperativity

Cooperativity (see Figure 9) is reflected by a sigmoid curve, as compared to the hyperbolic curve of Michaelis–Menten. An enzyme of several subunits that exhibits cooperativity is far more sensitive to control mechanisms than is an enzyme of one subunit and hence one active site.

The first example of cooperativity was observed in hemoglobin, which is not an enzyme but behaves like one in many ways. The absorption of oxygen in the lungs and its deposition in the tissues is far more efficient because the subunits of hemoglobin show positive cooperativity, so-called because the first molecule of substrate makes it easier for the next to bind.

Negative cooperativity (also illustrated in Figure 9), in which the binding of one molecule makes it less easy for the next to bind, also occurs in living things. Negative cooperativity makes an enzyme less sensitive to fluctuations in concentrations of metabolites and may be important for enzymes that must be present in the cell at relatively constant levels of activity.

Some enzymes are closely associated aggregates of several enzyme units; the pyruvate dehydrogenase system, for example, contains five different enzymes, has a total molecular weight of 4,000,000, and consists of four different types of chains. Apparently, the enzymes in cells may be organized by forming complex units, by being absorbed on a cell wall, or by being isolated by membranes in special compartments. Since a pathway involves the stepwise modification of chemical compounds, aggregations of the enzymes in a given pathway facilitate their function in a manner similar to an industrial assembly line.

(D.E.K./Ed.)

CARBOHYDRATES

The term carbohydrate, which means watered carbon, represents a class of naturally occurring compounds and derivatives formed from them. In the early part of the 19th century, substances such as wood, starch, and linen were found to be composed mainly of molecules containing atoms of carbon (C), hydrogen (H), and oxygen (O), and to have the general formula $C_6H_{12}O_6$; other organic molecules with similar formulas were found to have a similar ratio of hydrogen to oxygen. The general formula $C_x(H_2O)_x$ is commonly used to represent many carbohydrates.

Carbohydrates are probably the most abundant and widespread organic substances in nature. Essential constituents of all living things, carbohydrates are formed by green plants from carbon dioxide and water during the process of photosynthesis. Carbohydrates serve organisms as energy sources and as essential structural components; in addition, part of the structure of nucleic acids, which contain genetic information, consists of carbohydrate.

General features

CLASSIFICATION AND NOMENCLATURE

Although a number of classification schemes have been devised for carbohydrates, the division into four major groups—monosaccharides, disaccharides, oligosaccharides, and polysaccharides—used here is among the most common. Most monosaccharides, or simple sugars, are found in grapes, other fruits, and honey. Although they can contain from three to nine carbon atoms, the most common representatives consist of five or six joined together to form a chainlike molecule. Three of the most important simple sugars, glucose—also known as dextrose, grape sugar, and corn sugar—fructose (fruit sugar), and galactose, have the same molecular formula, $(C_6H_{12}O_6)$, but, because their atoms have different structural arrangements, the sugars have different characteristics; *i.e.,* they are isomers. Slight changes in structural arrangements are detectable by living

Basic molecular structures

things and influence the biological significance of isomeric compounds. It is known, for example, that the degree of sweetness of various sugars differs according to the arrangement of the hydroxyl groups (—OH) that compose part of the molecular structure; a direct correlation that may exist between taste and any specific structural arrangement, however, has not yet been established—that is, it is not yet possible to predict the taste of a sugar by knowing its specific structural arrangement. The energy in the chemical bonds of glucose indirectly supplies most living things with a major part of the energy that is necessary for them to carry on their activities. Galactose, which is rarely found as a simple sugar, is usually combined with other simple sugars in order to form larger molecules.

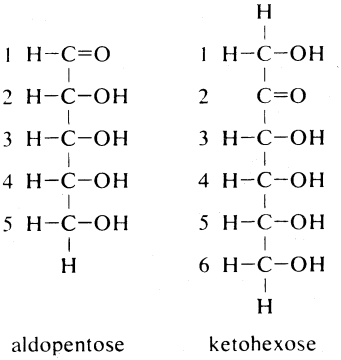
Two molecules of a simple sugar that are linked to each other form a disaccharide, or double sugar. The disaccharide sucrose, or table sugar, consists of one molecule of glucose and one molecule of fructose; the most familiar sources of sucrose are sugar beets and cane sugar. Milk sugar, or lactose, and maltose are also disaccharides. Before the energy in disaccharides can be utilized by living things, the molecules must be broken down into their respective monosaccharides.

Oligosaccharides, which consist of three to six monosaccharide units, are rather infrequently found in natural sources, although a few plant derivatives have been identified.

Polysaccharides (the term means many sugars) represent most of the structural and energy-reserve carbohydrates found in nature. Large molecules that may consist of as many as 10,000 monosaccharide units linked together, polysaccharides vary considerably in size, in structural complexity, and in sugar content; several hundred distinct types have thus far been identified. Cellulose, the principal structural component of plants, is a complex polysaccharide comprising many glucose units linked together; it is the most common polysaccharide. The starch found in plants and the glycogen found in animals also are complex glucose polysaccharides. Starch (from the old English word *stercan* meaning "to stiffen") is found mostly in seeds, roots, and stems, where it is stored as an available energy source for plants. Plant starch may be processed into such foods for man as bread, or it may be consumed directly—as are potatoes. Glycogen, which consists of branching chains of glucose molecules, is formed in the liver and muscles of higher animals and is stored as an energy source.

Names for monosaccharides

The generic nomenclature ending for the monosaccharides is -ose; thus, the term pentose (pent = five) is used for monosaccharides containing five carbon atoms, and hexose (hex = six) is used for those containing six. In addition, because the monosaccharides contain a chemically reactive group that is either an aldehydo group (H–C=O) or a keto group (R–C=O), they are frequently referred to as aldopentoses or ketopentoses or aldohexoses or ketohexoses; in the examples below, the aldehydo group is at position 1 of the aldopentose, the keto group is at position 2 of the ketohexose. Glucose is an aldohexose—*i.e.*, it contains six carbon atoms, and the chemically reactive group is an aldehydo group.



aldopentose ketohexose

BIOLOGICAL SIGNIFICANCE

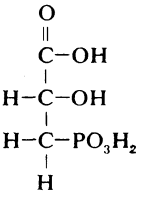
The importance of carbohydrates to living things can hardly be overemphasized. The energy stores of most animals and plants are both carbohydrate and lipid in nature; carbohydrates are generally available as an immediate energy source, whereas lipids act as a long-term energy resource and tend to be utilized at a slower rate. Glucose, the prevalent uncombined, or free, sugar circulating in the blood of higher animals, is essential to cell function. The proper regulation of glucose metabolism is of paramount importance to survival.

The ability of ruminants, such as cattle, sheep, and goats, to convert the polysaccharides present in grass and similar feeds into protein provides a major source of protein for man. A number of medically important antibiotics, such as streptomycin, are carbohydrate derivatives. The cellulose in plants is used to manufacture paper, wood for construction, and fabrics.

Role in the biosphere. The essential process in the biosphere, the portion of the Earth in which life can occur, that has permitted the evolution of life as it now exists is the conversion by green plants of carbon dioxide from the atmosphere into carbohydrates, using light energy from the Sun. This process, called photosynthesis, results in both the release of oxygen gas into the atmosphere and the transformation of light energy into the chemical energy of carbohydrates. The energy stored by plants during the formation of carbohydrates is used by animals to carry out mechanical work and to perform biosynthetic activities.

Photosynthesis

All green plants apparently photosynthesize in the same way, yielding as an immediate product the compound 3-phosphoglyceric acid; the formula, in which P represents phosphorus, is illustrated below.



3-phosphoglyceric acid

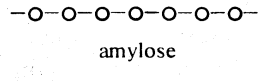
This compound then is transformed into cell-wall components such as cellulose, varying amounts of sucrose, and starch—depending on the plant type—and a wide variety of polysaccharides, other than cellulose and starch, that function as essential structural components. For a detailed discussion of the process of photosynthesis, see PHOTOSYNTHESIS.

Role in human nutrition. The total caloric, or energy, requirement for an individual depends on age, occupation, and other factors but generally ranges between 2,000 and 4,000 calories per 24-hour period (one calorie, as this term is used in nutrition, is the amount of heat necessary to raise the temperature of 1,000 grams of water from 15° to 16° C [59° to 61° F]; in other contexts this amount of heat is called the kilocalorie). Carbohydrate that can be used by man produces four calories per gram as opposed to nine calories per gram of fat and four per gram of protein. In areas of the world where nutrition is marginal, a high proportion (approximately one to two pounds) of an individual's daily energy requirement may be supplied by carbohydrate, with most of the remainder coming from a variety of fat sources.

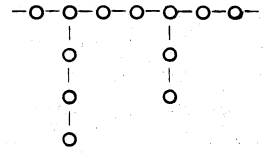
Although carbohydrates may compose as much as 80 percent of the total caloric intake in the human diet, for a given diet, the proportion of starch to total carbohydrate is quite variable, depending upon the prevailing customs. In the Far East and in areas of Africa, for example, where rice or tubers such as manioc provide a major food source, starch may account for as much as 80 percent of the total carbohydrate intake. In a typical Western diet, 33 to 50 percent of the caloric intake is in the form of carbohydrate. Approximately half (*i.e.*, 17 to 25 percent) is represented by starch; another third by table sugar (sucrose) and milk sugar (lactose); and smaller percentages by monosaccharides such as glucose and fructose, which are

common in fruits, honey, syrups, and certain vegetables such as artichokes, onions, and sugar beets. The small remainder consists of bulk, or indigestible carbohydrate, which comprises primarily the cellulosic outer covering of seeds and the stalks and leaves of vegetables. (See also NUTRITION.)

Role in energy storage. Starches, the major plant-energy-reserve polysaccharides used by man, are stored in plants in the form of nearly spherical granules that vary in diameter from about three to 100 micrometres (about .0001 to .004 inch). Most plant starches consist of a mixture of two components, amylose and amylopectin (see diagrams). As the diagrams show, the glucose molecules composing amylose have a straight-chain, or linear, structure; amylopectin has a branched-chain structure and is a somewhat more compact molecule. Several thousand glucose units may be present in a single starch molecule (each small circle represents one glucose molecule).



amylose



amylopectin

Starch content of plants

In addition to granules, many plants have large numbers of specialized cells, called parenchymatous cells, the principal function of which is the storage of starch; examples of plants with these cells include root vegetables and tubers. The starch content of plants varies considerably; the highest concentrations are found in seeds and in cereal grains, which contain up to 80 percent of their total carbohydrate as starch. The amylose and amylopectin components of starch occur in variable proportions; most plant species store approximately 25 percent of their starch as amylose and 75 percent as amylopectin. This proportion can be altered, however, by selective-breeding techniques, and some varieties of corn have been developed that produce up to 70 percent of their starch as amylose, which is more easily digested by man than is amylopectin.

In addition to the starches, some plants (*e.g.*, the Jerusalem artichoke and the leaves of certain grasses, particularly rye grass) form storage polysaccharides composed of fructose units rather than glucose. Although the fructose polysaccharides can be broken down and used to prepare syrups, they cannot be digested by higher animals.

Starches are not formed by animals; instead, they form a closely related polysaccharide, glycogen. Virtually all vertebrate and invertebrate animal cells, as well as those of numerous fungi and protozoans, contain some glycogen; particularly high concentrations of this substance are found in the liver and muscle cells of higher animals. The overall structure of glycogen, which is a highly branched molecule consisting of glucose units, has a superficial resemblance to that of the amylopectin component of starch, although the structural details of glycogen are significantly different. Under conditions of stress or muscular activity in animals, glycogen is rapidly broken down to glucose, which is subsequently used as an energy source. In this manner, glycogen acts as an immediate carbohydrate reserve. Furthermore, the amount of glycogen present at any given time, especially in the liver, directly reflects an animal's nutritional state; *i.e.*, when adequate food supplies are available, both glycogen and fat reserves of the body increase, but when food supplies decrease or when the food intake falls below the minimum energy requirements, the glycogen reserves are depleted quite rapidly, while those of fat are used at a slower rate.

Role in plant and animal structure. Whereas starches and glycogen represent the major reserve polysaccharides of living things, most of the carbohydrate found in nature occurs as structural components in the cell walls of plants. Carbohydrates in plant cell walls generally consist of several distinct layers, one of which contains a higher concentration of cellulose than the others. The physical and chemical properties of cellulose are strikingly different from those of the amylose component of starch.

In most plants, the cell wall is about 0.5 micrometres thick and contains a mixture of cellulose, pentose-containing polysaccharides (pentosans), and an inert (chemically unreactive) plastic-like material called lignin. The amounts of cellulose and pentosan may vary; most plants contain between 40 and 60 percent cellulose, although higher amounts are present in the cotton fibre.

Polysaccharides also function as major structural components in animals. Chitin, which is similar to cellulose, is found in insects and other arthropods. Other complex polysaccharides predominate in the structural tissues of higher animals.

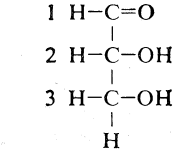
Structural arrangements and properties

STEREOISOMERISM

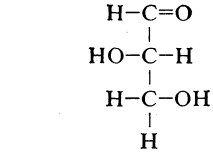
Isomers of glyceraldehyde

Studies by the German chemist Emil Fischer in the late 19th century showed that carbohydrates, such as fructose and glucose, with the same molecular formulas but with different structural arrangements and properties (*i.e.*, isomers) can be formed by relatively simple variations of their spatial, or geometric, arrangements. This type of isomerism, which is called stereoisomerism, exists in all biological systems; and, among carbohydrates, the simplest example is provided by the three-carbon aldose sugar glyceraldehyde. There is no way by which the structures of the two isomers of glyceraldehyde (see the formulas below, which are the so-called Fischer projection formulas that are commonly used to distinguish between such isomers) can be made identical, excluding breaking and reforming the linkages, or bonds, of the hydrogen (—H) and hydroxyl (—OH) groups attached to the carbon at position 2. The isomers are, in fact, mirror images akin to right and left hands; the term enantiomorphism is frequently employed for such isomerism. The chemical and physical properties of enantiomers are identical except for the property of optical rotation.

As explained above, optical rotation is the rotation of the plane of polarized light. Polarized light is light that has been separated into two beams that vibrate at right angles to each other; solutions of substances that rotate the plane of polarization are said to be optically active, and the degree of rotation is called the optical rotation of the solution. In the case of the isomers of glyceraldehyde, the magnitudes of the optical rotation are the same, but the direction in which the light is rotated—generally designated as plus, or *d* for dextrorotatory (to the right), or as minus, or *l* for levorotatory (to the left)—is opposite; *i.e.*, a solution of D-(*d*)-glyceraldehyde causes the plane of polarized light to rotate to the right, and a solution of L-(*l*)-glyceraldehyde rotates the plane of polarized light to the left. Fischer projection formulas for the two isomers of glyceraldehyde are given below (see *Configuration*, below, for explanation of D and L).



D-(*d*)-glyceraldehyde

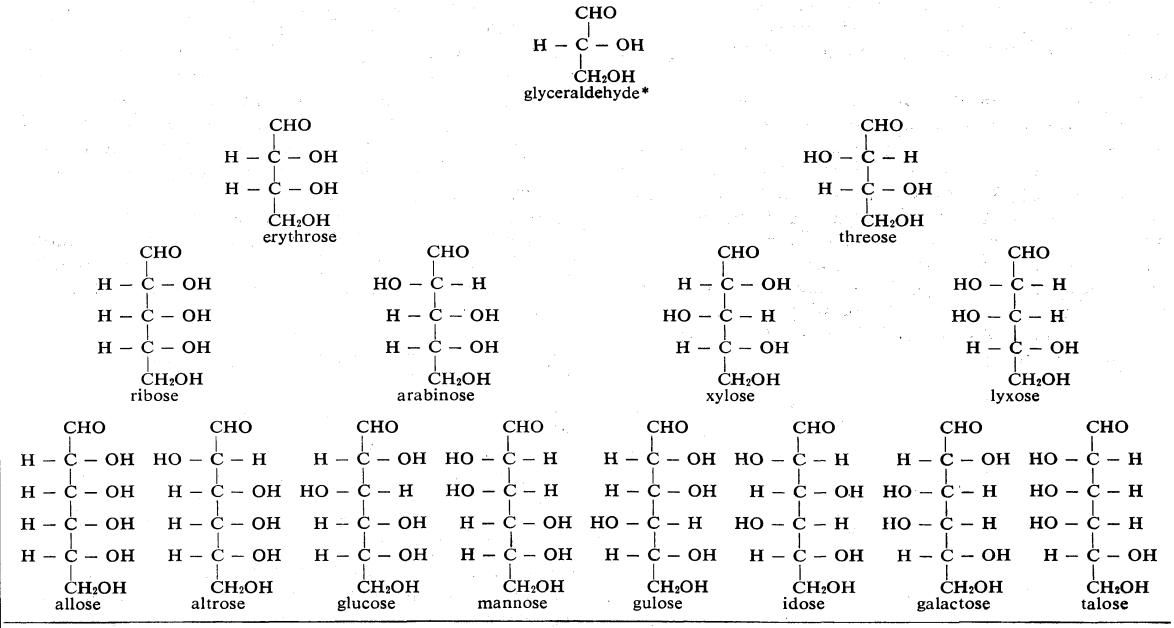


L-(*l*)-glyceraldehyde

CONFIGURATION

Molecules, such as the isomers of glyceraldehyde—the atoms of which can have different structural arrangements—are known as asymmetrical molecules. The number of possible structural arrangements for an asymmetrical molecule depends on the number of centres of asymmetry; *i.e.*, for n (any given number of) centres of asymmetry, 2^n different isomers of a molecule are possible. An asymmetrical centre in the case of carbon is defined as a carbon atom to which four different groups are attached. In the three-carbon aldose sugar, glyceraldehyde, the asymmetri-

Table 5: Aldoses, up to Aldohexoses, of the D-Series



*Aldehydo carbon (—CHO) is carbon atom 1; the remaining carbon atoms are numbered in succession. The hydroxyl groups only are shown for the asymmetrical carbon atoms. The D-configuration is derived from the asymmetrical carbon atom most remote from the aldehydo end of the molecule; *i.e.*, the hydroxyl group at carbon 3 in 4 carbon sugars; 4 in pentoses, and 5 in hexoses projects to the right. The L-series are corresponding mirror images; *i.e.*, the hydroxyl group projects to the left.

cal centre is located at the central carbon atom. The four different groups attached to the atom are:

(1) $H-\overset{|}{C}=O$; (2) $H-$; (3) $-OH$; and (4) $H-\overset{|}{C}-OH$ (H).

The position of the hydroxyl group (—OH) attached to the central carbon atom—*i.e.*, whether —OH projects from the left or the right—determines whether the molecule rotates the plane of polarized light to the left, or to the right. Since glyceraldehyde has one asymmetrical centre, n is one in the relationship 2^n, and there thus are two possible isomers. Sugars containing four carbon atoms have two asymmetrical centres; hence, there are four possible isomers (2^2). Similarly, sugars with five carbon atoms have three asymmetrical centres, and thus have eight isomers (2^3). Keto sugars have one less asymmetrical centre for a given number of carbon atoms than do aldehydo sugars.

A convention of nomenclature, devised in 1906, states that the form of glyceraldehyde whose asymmetrical carbon atom has a hydroxyl group projecting to the right (see Fischer projection formulas) is designated as of the D-configuration; that form, whose asymmetrical carbon atom has a hydroxyl group projecting to the left, is designated as L. All sugars that can be derived from D-glyceraldehyde—*i.e.*, hydroxyl group attached to the asymmetrical carbon atom most remote from the aldehydo or keto end of the molecule projects to the right—are said to be of the D-configuration; those sugars derived from L-glyceraldehyde are said to be of the L-configuration. See Table 5 for aldoses—*i.e.*, sugars containing an aldehydo group ($\diagdown C=O$) —of the D-configuration.

The configurational notation D or L is independent of the sign of the optical rotation of a sugar in solution. It is common, therefore, to designate both, as for example, D-(*l*)-fructose or D-(*d*)-glucose; *i.e.*, both have a D-configuration at the centre of asymmetry most remote from the aldehydo end (in glucose) or keto end (in fructose) of the molecule, but fructose is levorotatory, and glucose is dextrorotatory—hence the latter has been given the alternative name dextrose. Although the initial assignments of configuration for the glyceraldehydes were made on purely arbitrary grounds, studies that were carried out nearly half a century later established them as correct in an absolute spatial sense. In biological systems, only the D or L form may be utilized.

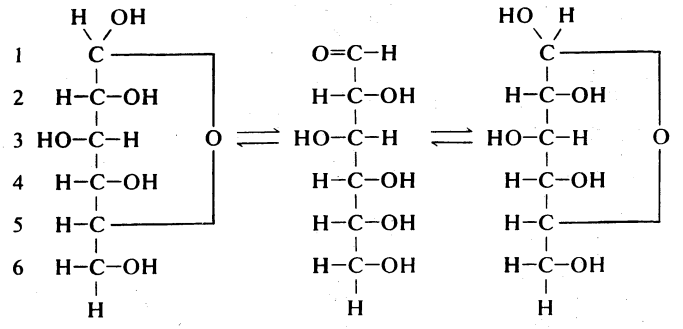
When more than one asymmetrical centre is present in a molecule, as is the case with sugars having four or more carbon atoms, a series of DL pairs exists, and they are functionally, physically, and chemically distinct; thus, although D-xylose and D-lyxose (see Table 5) both have five carbon atoms and are of the D-configuration, the spatial arrangement of the asymmetrical centres (at carbon atoms 2, 3, and 4) is such that they are not mirror images.

DL pairs

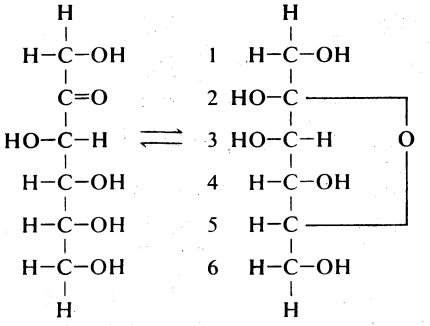
HEMIACETAL AND HEMIKETAL FORMS

Although optical rotation has been one of the most frequently determined characteristics of carbohydrates since its recognition in the late 19th century, the rotational behaviour of freshly prepared solutions of many sugars differs from that of solutions that have been allowed to stand. This phenomenon, termed mutarotation, is demonstrable even with apparently identical sugars and is caused by a type of stereoisomerism involving formation of an asymmetrical centre at the first carbon atom (aldehydo carbon) in aldoses and the second one (keto carbon) in ketoses.

Most pentose and hexose sugars, therefore, do not exist as linear, or open-chain, structures in solution, as indicated for the aldoses in Table 5, but form cyclic, or ring, structures termed hemiacetal or hemiketal forms, respectively. As illustrated for glucose and fructose, the cyclic structures are formed by the addition of the hydroxyl group (—OH) from either the fourth, fifth, or sixth carbon atom (in the diagram, the numbers 1 through 6 represent the positions of the carbon atoms) to the carbonyl group ($\diagdown C=O$) at position 1 in glucose or 2 in fructose. A five-membered ring is illustrated for the ketohexose, fructose; a six-membered ring is illustrated for the aldohexose, glucose. In either case, the cyclic forms are in equilibrium with (*i.e.*, the rate of conversion from one form to another is stable) the open-chain structure—a free aldehyde if the solution contains glucose, a free ketone if it contains fructose; each form has a different optical rotation value. Since the forms are in equilibrium with each other, a constant value of optical rotation is measurable; the two cyclic forms represent more than 99.9 percent of the sugar in the case of a glucose solution.



α-D-glucose (hemiacetal form) | D-glucose (linear, open-chain, or free form) | β-D-glucose (hemiacetal form)



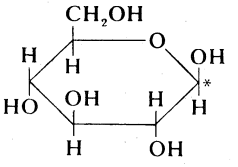
D-fructose (linear, open-chain, or free form) | α-D-fructose (hemiketal form)

By definition, the carbon atom containing the aldehydo () or keto group () is termed the anomeric carbon atom; similarly, carbohydrate stereoisomers that differ in configuration only at this carbon atom are called anomers. When a cyclic hemiacetal or hemiketal structure forms, the structure with the new hydroxyl group projecting on the same side as that of the oxygen involved in forming the ring is termed the alpha anomer (see hemiacetal forms for glucose and diagram); that with the hydroxyl group projecting on the opposite side from that of the oxygen ring is termed the beta anomer (see diagram). The spatial arrangements of the atoms in these

Anomeric carbon atom

alpha isomer | beta isomer

cyclic structures are better shown (glucose is used as an example) in the representation devised by the British organic chemist Walter Norman (later Sir Norman) Haworth about 1930; they are still in widespread use. In the formulation the asterisk indicates the position of the anomeric carbon atom; the carbon atoms, except at position 6, usually are not labelled. The large number of asymmetrical carbon atoms and the consequent number of possible isomers considerably complicates the structural chemistry of carbohydrates.



Haworth formulation of β-D-glucose

Classes of carbohydrates

MONOSACCHARIDES

Sources. The most common naturally occurring monosaccharides are D-glucose, D-mannose, D-fructose, and D-galactose among the hexoses, and D-xylose and L-arabinose among the pentoses. In a special sense, D-ribose and 2-deoxy-D-ribose are ubiquitous because they form the carbohydrate component of ribonucleic acid (RNA) and deoxyribonucleic acid (DNA), respectively; these sugars are present in all cells as components of nucleic acids. Sources of some of the naturally occurring monosaccharides are listed in Table 6.

Table 6: Some Naturally Occurring Monosaccharides

sugar	sources
L-arabinose	mesquite gum; wheat bran
D-ribose	all living cells as component of ribonucleic acid
D-xylose	corncobs; seed hulls; straw
D-ribulose	one derivative, is an intermediate in photosynthesis
2-deoxy-D-ribose	as constituent of deoxyribonucleic acid
D-galactose	lactose; agar; gum arabic; brain glycolipids
D-glucose	sucrose; cellulose; starch; glycogen
D-mannose	seeds; ivory nut
D-fructose	sucrose; artichokes; honey
L-fucose	marine algae; seaweed
L-rhamnose	poison-ivy blossom; oak bark
D-mannoheptulose	avocado
D-altroheptulose	numerous plants

D-xylose, found in most plants in the form of a polysaccharide called xylan, is prepared from corncobs, cottonseed hulls, or straw by chemical breakdown of xylan. D-galactose, a common constituent of both oligosaccharides and polysaccharides, also occurs in carbohydrate-containing lipids, called glycolipids, which are found in the brain and other nervous tissues of most animals. Galactose is generally prepared by acid hydrolysis (breakdown involving water) of lactose, which is composed of galactose and glucose. Since the biosynthesis of galactose in animals occurs through intermediate compounds derived directly from glucose, animals do not require galactose in the diet. In fact, in most human populations (Caucasoid peoples being the major exception) the majority of people do not retain the ability to manufacture the enzyme necessary to metabolize galactose after they reach the age of four, and many individuals possess a hereditary defect known as galactosemia and never have the ability to metabolize galactose.

D-glucose

D-glucose (from the Greek word *glykys*, meaning "sweet"), the naturally occurring form, is found in fruits, honey, blood, and, under abnormal conditions, in urine. It is also a constituent of the two most common naturally found disaccharides, sucrose and lactose, as well as the exclusive structural unit of the polysaccharides cellulose, starch, and glycogen. Generally, D-glucose is prepared from either potato starch or cornstarch.

D-fructose, a ketohexose, is one of the constituents of the disaccharide sucrose and is also found in uncombined form in honey, apples, and tomatoes. Fructose, generally considered the sweetest monosaccharide, is prepared by sucrose hydrolysis and is metabolized by man.

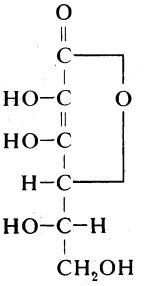
Chemical reactions. The reactions of the monosaccharides can be conveniently subdivided into those associated with the aldehydo or keto group and those associated with the hydroxyl groups.

Reactions involving the aldehydo or keto group

The relative ease with which sugars containing a free or potentially free aldehydo or keto group can be oxidized to form products has been known for a considerable time and once was the basis for the detection of these so-called reducing sugars in a variety of sources. For many years, analyses of blood glucose and urinary glucose were carried out by a procedure involving the use of an alkaline copper compound. Because the reaction has undesirable

features—extensive destruction of carbohydrate structure occurs, and the reaction is not very specific (*i.e.,* sugars other than glucose give similar results) and does not result in the formation of readily identifiable products—blood and urinary glucose now are analyzed by using the enzyme glucose oxidase, which catalyzes the oxidation of glucose to products that include hydrogen peroxide. The hydrogen peroxide then is used to oxidize a dye present in the reaction mixture; the intensity of the colour is directly proportional to the amount of glucose initially present. The enzyme, glucose oxidase, is highly specific for β-D-glucose.

In another reaction, the aldehydo group of glucose (H–C=O) reacts with alkaline iodine to form a class of compounds called aldonic acids. One important aldonic acid is ascorbic acid (vitamin C, see structure), an essential dietary component for man and guinea pigs. The formation of similar acid derivatives does not occur with the keto sugars.



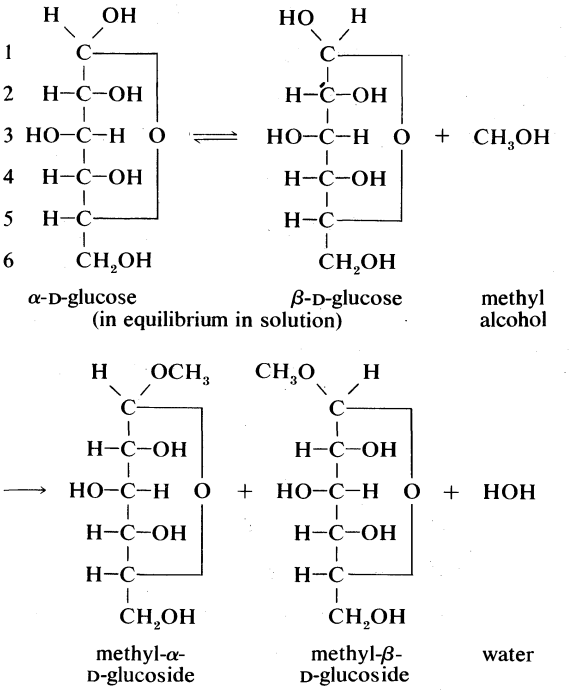
ascorbic acid, vitamin C (L-gulonolactone-2,3-enediol)

Either the aldehydo or the keto group of a sugar may be reduced (*i.e.,* hydrogen added) to form an alcohol; compounds formed in this way are called alditols, or sugar alcohols. The product formed as a result of the reduction of the aldehydo carbon of D-glucose is called sorbitol (D-glucitol). D-glucitol also is formed when L-sorbose is reduced. The reduction of mannose results in mannitol, that of galactose in dulcitol.

Sugar alcohols that are of commercial importance include sorbitol (D-glucitol), which is commonly used as a sweetening agent, and D-mannitol, which is also used as a sweetener, particularly in chewing gums, because it has a limited water solubility and remains powdery and granular on long storage.

Formation of glycosides

The hydroxyl group that is attached to the anomeric carbon atom (*i.e.,* the carbon containing the aldehydo or keto group) of carbohydrates in solution has unusual reactivity, and derivatives, called glycosides, can be formed; glycosides formed from glucose are called glucosides. It is not possible for equilibration between the α- and β-anomers of a glycoside in solution (*i.e.,* mutarotation) to occur. The reaction by which a glycoside is formed (see below) involves the hydroxyl group (—OH) of the anomeric carbon atom (numbered 1) of both α and β forms of D-glucose—α and β forms of D-glucose are shown in equilibrium in the reaction sequence—and the hydroxyl group of an alcohol (methyl alcohol in the reaction sequence); methyl α-D-glucosides and β-D-glucosides are formed as products, as is water.



α-D-glucose (in equilibrium in solution) β-D-glucose methyl alcohol

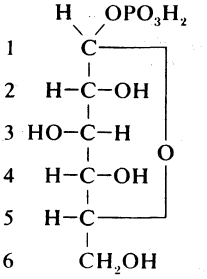
methyl-α-D-glucoside methyl-β-D-glucoside water

Among the wide variety of naturally occurring glycosides are a number of plant pigments, particularly those red, violet, and blue in colour; these pigments are found in flowers and consist of a pigment molecule attached to a sugar molecule, frequently glucose. Plant indican (from *Indigofera* species), composed of glucose and the pigment indoxyl, was important in the preparation of indigo dye before synthetic dyes became prevalent. Of a number of heart-muscle stimulants that occur as glycosides, digitalis is still used. Other naturally occurring glycosides include vanillin, which is found in the vanilla bean, and amygdalin (oil of bitter almonds); a variety of glycosides found in mustard have a sulfur atom at position 1 rather than oxygen.

Important antibiotics

A number of important antibiotics are glycosides; the best known are streptomycin and erythromycin. Glucosides—*i.e.,* glycosides formed from glucose—in which the anomeric carbon atom (at position 1) has phosphoric acid linked to it, are extremely important biological compounds.

For example, α-D-glucose-1-phosphate (see formula), is an intermediate product in the biosynthesis of cellulose, starch, and glycogen; similar glycosidic phosphate derivatives of other monosaccharides participate in the formation of naturally occurring glycosides and polysaccharides.



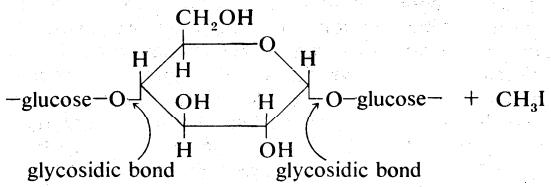
α-D-glucose-1-phosphate

The hydroxyl groups other than the one at the anomeric carbon atom can undergo a variety of reactions, several of which deserve mention. Esterification, which consists of reacting the hydroxyl groups with an appropriate acidic compound, results in the formation of a class of compounds called sugar esters. Among the common ones are the sugar acetates, in which the acid is acetic acid. Esters of phosphoric acid and sulfuric acid are important biological compounds; glucose-6-phosphate, for example, plays a central role in the energy metabolism of most living cells, and D-ribulose 1,5-diphosphate is important in photosynthesis.

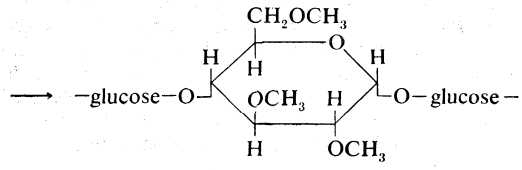
Formation of methyl ethers

Treatment of a carbohydrate with methyl iodide or similar agents under appropriate conditions results in the formation of compounds in which the hydroxyl groups are converted to methyl groups (—CH_3). Called methyl ethers, these compounds are employed in structural studies of oligosaccharides and polysaccharides because their formation does not break the bonds, called glycosidic bonds, that

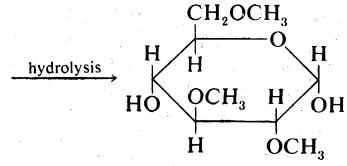
link adjacent monosaccharide units. In the reaction sequence shown, a segment of a starch molecule, consisting of three glucose units, is indicated; the Haworth formulation used to represent one of the glucose units shows the locations of the glycosidic bonds and the —OH groups. When complete etherification of the starch molecule is carried out, using methyl iodide, methyl groups become attached to the glucose molecules at the three positions shown in the methylated segment of the starch molecule; note that the glycosidic bonds have not been broken by the reaction with methyl iodide. When the methylated starch molecule then is broken down (hydrolyzed), hydroxyl groups are located at the positions in the molecule previously involved in linking one sugar molecule to another, and a methylated glucose, in this case named 2,3,6 tri-*O*-methyl-D-glucose, forms. The linkage positions (in the example, at carbon atoms 1 and 4; the carbon atoms are numbered in the structure of the methylated glucose),



segment of a starch molecule — methyl iodide



methylated segment of starch molecule

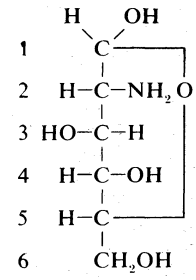


methylated glucose

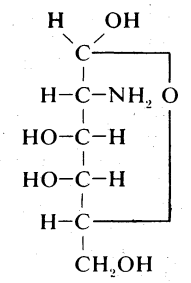
which are not methylated, in a complex carbohydrate can be established by analyzing the locations (in the example, at carbon atoms 2, 3, and 6) of the methyl groups in the monosaccharides. This technique is useful in determining the structural details of polysaccharides, particularly since the various methylated sugars are easily separated by techniques involving gas chromatography, in which a moving gas stream carries a mixture through a column of a stationary liquid or solid, the components thus being resolved.

When the terminal group (CH_2OH) of a monosaccharide is oxidized chemically or biologically, a product called a uronic acid is formed. Glycosides that are derived from D-glucuronic acid (the uronic acid formed from D-glucose) and fatty substances called steroids appear in the urine of animals as normal metabolic products; in addition, foreign toxic substances are frequently converted in the liver to glucuronides before excretion in the urine. D-glucuronic acid also is a major component of connective tissue polysaccharides, and D-galacturonic acid and D-mannuronic acid, formed from D-galactose and D-mannose, respectively, are found in several plant sources.

Other compounds formed from monosaccharides include those in which one hydroxyl group, usually at the carbon at position 2 (see formulas for D-glucosamine and D-galactosamine), is replaced by an amino group ($-NH_2$); these compounds, called amino sugars, are widely distributed in nature. The two most important ones are glucosamine (2-amino-2-deoxy-D-glucose) and galactosamine (2-amino-2-deoxy-D-galactose).



D-glucosamine

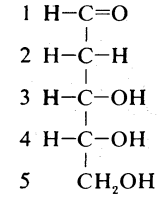


D-galactosamine

Neither amino sugar is found in the uncombined form. Both occur in animals as components of glycolipids or polysaccharides; *e.g.*, the primary structural polysaccharide (chitin) of insect outer skeletons and various blood-group substances.

Deoxy sugars

In a number of naturally occurring sugars, known as deoxy sugars, the hydroxyl group at a particular position is replaced by a hydrogen atom. By far the most important representative is 2-deoxy-D-ribose (see formula), the pentose sugar found in deoxyribonucleic acid (DNA); the hydroxyl group at the carbon atom at position 2 has been replaced by a hydrogen atom.



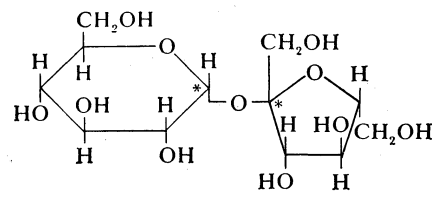
2-deoxy-D-ribose

Other naturally occurring deoxy sugars are hexoses, of which L-rhamnose (6-deoxy-L-mannose) and L-fucose (6-deoxy-L-galactose) are the most common; the latter, for example, is present in the carbohydrate portion of blood-group substances and in red-blood-cell membranes.

DISACCHARIDES AND OLIGOSACCHARIDES

Disaccharides are a specialized type of glycoside in which the anomeric hydroxyl group of one sugar has combined with the hydroxyl group of a second sugar with the elimination of the elements of water. Although an enormous number of disaccharide structures are possible, only a limited number are of commercial or biological significance.

Sucrose and trehalose. Sucrose, or common table sugar, has a world production amounting to well over 10,000,000 tons annually. The unusual type of linkage between the two anomeric hydroxyl groups of glucose and fructose (see formula, in which the asterisk indicates anomeric carbon atom) means that neither a free aldehydo group (on the glucose moiety) nor a free keto group (on the fructose moiety) is available to react unless the linkage between the monosaccharides is destroyed; for this reason, sucrose is known as a nonreducing sugar. Sucrose solutions do



sucrose

not exhibit mutarotation, which involves formation of an asymmetrical centre at the aldehydo or keto group. If the linkage between the monosaccharides composing sucrose is broken, the optical rotation value of sucrose changes from positive to negative; the new value reflects the composite rotation values for D-glucose, which is dextrorotatory (+52°), and D-fructose, which is levorotatory (−92°). The change in the sign of optical rotation from

Invert sugar

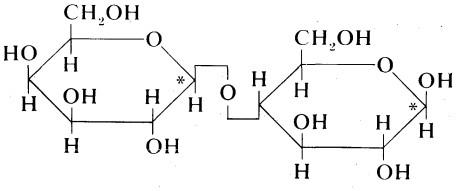
positive to negative is the reason sucrose is sometimes called invert sugar.

The commercial preparation of sucrose takes advantage of the alkaline stability of the sugar, and a variety of impurities are removed from crude sugarcane extracts by treatment with alkali. After this step, syrup preparations are crystallized to form table sugar. Successive "crops" of sucrose crystals are "harvested," and the later ones are known as brown sugar. The residual syrupy material is called either cane final molasses or blackstrap molasses; both are used in the preparation of antibiotics, as sweetening agents, and in the production of alcohol by yeast fermentation.

Sucrose is formed following photosynthesis in plants by a reaction in which sucrose phosphate first is formed.

The disaccharide trehalose is similar in many respects to sucrose but is much less widely distributed. It is composed of two molecules of α-D-glucose and is also a nonreducing sugar. Trehalose is present in young mushrooms and in the resurrection plant (*Selaginella*); it is of considerable biological interest because it is also found in the circulating fluid (hemolymph) of many insects. Since trehalose can be converted to a glucose phosphate compound by an enzyme-catalyzed reaction that does not require energy, its function in hemolymph may be to provide an immediate energy source, a role similar to that of the carbohydrate storage forms (*i.e.*, glycogen) found in higher animals.

Lactose and maltose. Lactose is one of the sugars (sucrose is another) found most commonly in human diets throughout the world; it composes about 5 percent or more of the milk of all mammals. Lactose consists of two aldohexoses—β-D-galactose and glucose—linked so that the aldehydo group at the anomeric carbon of glucose is free to react (see structural formula, in which the asterisk indicates position of anomeric carbon atoms); *i.e.*, lactose is a reducing sugar.



β-lactose

A variety of metabolic disorders related to lactose may occur in infants; in some cases, they are the result of a failure to metabolize properly the galactose portion of the molecule.

Importance of maltose

Although not found in uncombined form in nature, the disaccharide maltose is biologically important because it is a product of the enzymatic breakdown of starches during digestion. Maltose consists of α-D-glucose linked to a second glucose unit in such a way that maltose is a reducing sugar. Maltose, which is readily hydrolyzed to glucose and can be metabolized by animals, is employed as a sweetening agent and as a food for infants whose tolerance for lactose is limited. Table 7 lists the component sugars of, the linkage between, and the occurrence of a number of disaccharides and oligosaccharides.

Table 7: Representative Disaccharides and Oligosaccharides

common name	component sugars	linkages	sources
Cellobiose	glucose, glucose	β 1 → 4†	hydrolysis of cellulose
Gentiobiose	glucose, glucose	β 1 → 6	plant glycosides, amygdalin
Isomaltose	glucose, glucose	α 1 → 6	hydrolysis of glycogen, amylopectin
Raffinose*	galactose, glucose, fructose	α 1 → 6, α 1 → 2	sugarcane, beets, seeds
Stachyose*	galactose, galactose, glucose, fructose	α 1 → 6, α 1 → 6, α 1 → 2	soybeans, jasmine, twigs, lentils

*Note that raffinose and stachyose are galactosyl sucroses. †The linkage joins carbon atom 1 (in the β configuration) of one glucose molecule and carbon atom 4 of the second glucose molecule; the linkage may also be abbreviated β-1, 4.

POLYSACCHARIDES

Polysaccharides, or glycans, may be classified in a number of ways; the following scheme is frequently used. Homopolysaccharides are defined as polysaccharides formed from only one type of monosaccharide. Homopolysaccharides may be further subdivided into straight-chain and branched-chain representatives, depending upon the arrangement of the monosaccharide units. Heteropolysaccharides are defined as polysaccharides containing two or more different types of monosaccharides; they may also occur in both straight-chain and branched-chain forms. In general, extensive variation of linkage types (see Table 8) does not occur within a polysaccharide structure, nor are there many polysaccharides composed of more than three or four different monosaccharides; most contain one or two.

HOMOPOLYSACCHARIDES

In general, homopolysaccharides have a well-defined chemical structure, although the molecular weight of an individual amylose or xylan molecule may vary within a particular range, depending on the source; molecules from a single source also may vary in size, because most polysaccharides are formed biologically by an enzyme-catalyzed process lacking genetic information regarding size. Several naturally occurring homopolysaccharides are listed in Table 8.

Cellulose and xylans

The basic structural component of most plants, cellulose, is widely distributed in nature. It has been estimated that nearly 10,000,000,000 tons of cellulose are synthesized yearly as a result of photosynthesis by higher plants. The proportion of cellulose to total carbohydrate found in plants may vary in various types of woods from 30 to 40 percent, and to more than 98 percent in the seed hair of the cotton plant. Cellulose, a large, linear molecule composed of 3,000 or more β-D-glucose molecules, is insoluble in water.

The chains of glucose units composing cellulose molecules are frequently aligned within the cell-wall structure of a plant to form fibre-like or crystalline arrangements. This alignment permits very tight packing of the chains and promotes their structural stability but also makes structural analysis difficult. The relationships between cellulose and other polysaccharides present in the cell wall are not well established; in addition, the presence of unusual chemical linkages or nonglucose units within the cellulose structure has not yet been established with certainty.

During the preparation of cellulose, raw plant material is treated with hot alkali; this treatment removes most of the lignin, the hemicelluloses, and the mucilaginous components. The cellulose then is processed to produce papers and fibres. The high resistance of cellulose to chemical or enzymatic breakdown is important in the manufacture of paper and cloth. Cellulose also is modified chemically for other purposes; *e.g.*, compounds such as cellulose acetate are used in the plastics industry, in the production of photographic film, and in the rayon-fibre industry. Viscose rayon is produced from an ester of cellulose, and cellulose nitrate is employed in the lacquer and explosives industries.

The noteworthy biological stability of cellulose is dramatically illustrated by trees, the life-span of which may be several thousand years. Enzymes capable of breaking down cellulose are generally found only among several species of bacteria and molds. The apparent ability of termites to utilize cellulose as an energy source depends on the presence in their intestinal tracts of protozoans that can break it down. Similarly, the single-celled organisms present in the rumina of sheep and cattle are responsible for the ability of these animals to utilize the cellulose present in typical grasses and other feeds.

Xylans are almost as ubiquitous as cellulose in plant-cell walls and contain predominantly β-D-xylose units linked as in cellulose (see Table 8). Some xylans contain other sugars, such as L-arabinose, but they form branches and are not part of the main chain. Xylans are of little commercial importance.

Starch

The term starch refers to a group of plant reserve polysaccharides consisting almost exclusively of a linear compo-

Table 8: Representative Homopolysaccharides

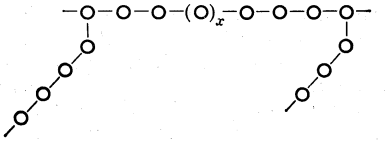
homopolysaccharide	sugar component	linkage	function	sources
Cellulose	glucose	β, 1 → 4	structural	throughout plant kingdom
Amylose	glucose	α, 1 → 4	food storage	starches, especially corn, potatoes, rice
Chitin	N-acetylglucosamine	β, 1 → 4	structural	insect and crustacean skeleton
Inulin	fructose	β, 2 → 1	food storage	artichokes, chicory
Xylan	xylose	β, 1 → 4	structural	all land plants
Glycogen	glucose	α, 1 → 4, 6 ← 1, α	food storage	liver and muscle cells of all animals
Amylopectin	glucose	α, 1 → 4, 6 ← 1, α	food storage	starches, especially corn, potatoes, rice
Dextran	glucose	α, 1 → 6, 4 ← 1, α	unknown	primarily bacterial
Agar*	galactose	α, 1 → 3	structural	seaweeds

*May contain sulfate groups.

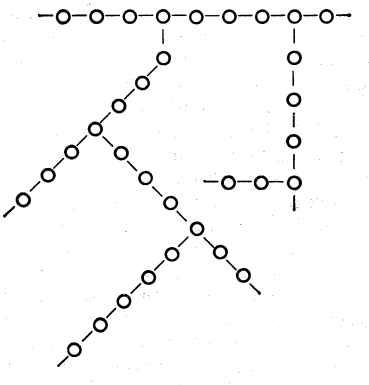
nent (amylose) and a branched component (amylopectin). Man's use of starch as an energy source depends on his ability to convert it completely to individual glucose units; the process is initiated by the action of enzymes called amylases, synthesized by the salivary glands in the mouth, and continues in the intestinal tract. The primary product of amylase action is maltose, which is hydrolyzed to two component glucose units as it is absorbed through the walls of the intestine.

A characteristic reaction of the amylose component of starch is the formation with iodine of a complex compound with a characteristic blue colour. About one iodine molecule is bound for each seven or eight glucose units, and at least five times that many glucose units are needed in an amylose chain to permit the effective development of the colour.

The amylopectin component of starch is structurally similar to glycogen in that both are composed of glucose units linked together in the same way, but the distance between branch points (see schematic diagrams, in which —O— represents one glucose unit) is greater in amylopectin than in glycogen, and the former may be thought of as occupying more space per unit weight.



schematic amylopectin structure



schematic glycogen structure

The applications of starches other than as foods are limited. Starches are employed in adhesive manufacture, and starch nitrate has some utility as an explosive.

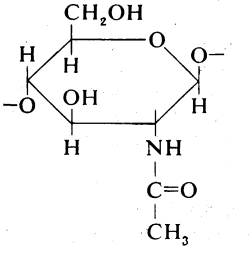
Glycogen

Glycogen, which is found in all animal tissues, is the primary animal storage form of carbohydrate and, indirectly, of rapidly available energy. The distance between branch points in a glycogen molecule is only five or six units (see schematic diagram above), which results in a compact treelike structure. The ability of higher animals to form and break down this extensively branched structure is essential to their well-being; in conditions known as glycogen storage diseases, these activities are abnormal, and the asymmetrical glycogen molecules that are formed have severe, often fatal, consequences. Glycogen synthesis and breakdown are controlled by substances called hormones.

Large molecules—*e.g.*, pectins and agars—composed of galactose or its uronic-acid derivative (galacturonic acid) are important because they can form gels. Pectins, which are predominantly galacturonans, are produced from citrus fruit rinds; they are used commercially in the preparation of jellies and jams. Agar is widely employed in biological laboratories as a solidifying agent for growth media for microorganisms and in the bakery industry as a gelling agent; it forms a part of the diet of people in several areas of the Far East.

Dextrans, a group of polysaccharides composed of glucose, are secreted by certain strains of bacteria as slimes. The structure of an individual dextran varies with the strain of microorganism. Dextrans can be used as plasma expanders (substitutes for whole blood) in cases of severe shock. In addition, a dextran derivative compound is employed medically as an anticoagulant for blood.

Chitin is structurally similar to cellulose, but the repeating sugar is 2-deoxy-2-acetamido-D-glucose (*N*-acetyl-D-glucosamine, see structural formula) rather than glucose.



N-acetyl-D-glucosamine

Sometimes referred to as animal cellulose, chitin is the major component of the outer skeletons of insects, crustaceans, and other arthropods, as well as annelid and nematode worms, mollusks, and coelenterates. The cell walls of most fungi also are predominantly chitin, which comprises nearly 50 percent of the dry weight of some species. Since chitin is nearly as chemically inactive as cellulose and easily obtained, numerous attempts, none of which has thus far been successful, have been made to develop it commercially. The nitrogen content of the biosphere, however, is stabilized by the ability of soil microorganisms to degrade nitrogen-containing compounds such as those found in insect skeletons; these microorganisms convert the nitrogen in complex molecules to a form usable by plants. If such microorganisms did not exist, much of the organic nitrogen present in natural materials would be unavailable to plants.

Heteropolysaccharides. In general, heteropolysaccharides (heteroglycans) contain two or more different monosaccharide units. Although a few representatives contain three or more different monosaccharides, most naturally occurring heteroglycans contain only two different ones and are closely associated with lipid or protein. The complex nature of these substances has made detailed structural studies extremely difficult. The major heteropolysaccharides include the connective-tissue

Table 9: Representative Heteropolysaccharides

heteropolysaccharide	component sugars	functions	distribution
Hyaluronic acid	D-glucuronic acid and *N*-acetyl-D-glucosamine	lubricant, shock absorber, water binding	connective tissue, skin
Chondroitin-4-sulfate*	D-glucuronic acid and *N*-acetyl-D-galactosamine-4-*O*-sulfate	calcium accumulation, cartilage and bone formation	cartilage
Heparin*	D-glucuronic acid, *L*-iduronic acid, *N*-sulfo-D-glucosamine	anticoagulant	mast cells, blood
Gamma globulin*	*N*-acetyl-hexosamine, D-mannose, D-galactose	antibody	blood
Blood group substance*	D-glucosamine, D-galactosamine, *L*-fucose, D-galactose	blood group specificity	cell surfaces, especially red-blood cells

*Covalently linked to protein; the proportion of protein to carbohydrate in such complex molecules varies from about 10% protein in the case of chondroitin-4-sulfate to better than 95% for gamma globulin.

polysaccharides, the blood-group substances, glycoproteins (combinations of carbohydrates and proteins) such as gamma globulin, and glycolipids (combinations of carbohydrates and lipids), particularly those found in the central nervous system of animals and in a wide variety of plant gums (see Table 9).

Connective tissue heteropolysaccharides

The most important heteropolysaccharides are found in the connective tissues of all animals and include a group of large molecules that vary in size, shape, and interaction with other body substances. They have a structural role, and the structures of individual connective-tissue polysaccharides are related to specific animal functions; hyaluronic acid, for example, the major component of joint fluid in animals, functions as a lubricating agent and shock absorber.

The connective-tissue heteropolysaccharides contain acidic groups (uronic acids or sulfate groups) and can bind both water and inorganic metal ions. They can also play a role in other physiological functions; *e.g.*, in the accumulation of calcium before bone formation. Ion-binding ability also appears to be related to the anticoagulant activity of the heteropolysaccharide heparin (see Table 9).

The size of the carbohydrate portion of glycoproteins such as gamma globulin or hen-egg albumin is usually between five and 10 monosaccharide units; several such units occur in some glycoprotein molecules. The function of the carbohydrate component has not yet been established except for glycoproteins associated with cell surfaces; in this case, they appear to act as antigenic determinants—*i.e.*, they are capable of inducing the formation of specific antibodies.

PREPARATION AND ANALYSIS

In general, monosaccharides are prepared by breakdown with acids of the polysaccharides in which they occur. Sugars usually are difficult to obtain in crystalline form, and the crystallization process usually is begun by "seeding" a concentrated solution of the sugar with crystals. The techniques employed for separation of monosaccharides depend to some extent on their physical and chemical properties; chromatographic procedures are often used.

Oligosaccharides and polysaccharides are prepared from natural sources by techniques that take advantage of size, alkaline stability, or some combination of these and other properties of the molecule of interest. It should be noted that preparation of an oligosaccharide or polysaccharide usually results in a range of molecular sizes of the desired molecule. The purity of a carbohydrate preparation, which is frequently based on an analysis of its composition, is more easily established for monosaccharides and disaccharides than for large, insoluble molecules such as cellulose.

Analytical techniques. A variety of organic chemical analytical techniques are generally applicable to studies involving carbohydrates. Optical rotation, for example, once was frequently used to characterize carbohydrates. The ability to measure the rotation of the plane of polarized light transmitted through a solution containing a carbohydrate depends on finding a suitable solvent; water usually is used, with light at a wavelength of 589 mμ (millimicrons). Optical rotation is no longer widely used to characterize monosaccharides. The magnitude and sign of the optical rotation of glycosides, however, is useful in assigning configuration (α or β) to the hydroxyl group at the anomeric centre; glycosides of the α-configuration generally have rotations of higher magnitude than do the same glycosides of the β-configuration. Optical rotation is not a completely additive property; a trisaccharide composed of three glucose residues, for example, does not have a rotation three times that of one glucose molecule. Sugar alcohols cannot form ring structures; their rotation values are extremely small, suggesting a relationship between ring structure and the ability of a carbohydrate to rotate the plane of polarized light. Certain types of reactions (*e.g.*, glycoside hydrolysis) can be monitored by measuring the change in optical rotation as a function of time. This technique is frequently used to examine the breakdown of disaccharides or oligosaccharides to monosaccharide units, especially if a large change in the net optical rotation may be expected, as occurs in the hydrolysis of sucrose.

Optical rotation

Spectroscopic techniques. Several other optical techniques used in chemistry have been applied to the analysis of carbohydrates. Infrared spectroscopy, used to measure vibrational and rotational excitation of molecules, and nuclear magnetic-resonance spectroscopy, which measures the excitation of certain components of molecules in a magnetic field induced by radio-frequency radiation, are valuable, although the similarity of the functional groups (*i.e.*, the hydroxyl groups) limits use of the former technique for most sugars. Proton magnetic-resonance spectroscopy, nuclear magnetic resonance applied to protons (H atoms), is employed to identify the relative spatial arrangements of individual hydrogen atoms in a molecule. When they are precisely placed, the corresponding positions of the hydroxyl groups attached to the same carbon atom can be deduced. An extension of this technique utilizes the resonance spectroscopy of carbon-13, a nonradioactive isotope of carbon, so that ring structures can be established with great accuracy. Both the proton and carbon magnetic resonance methods are best applied to monosaccharides; they are less valuable in studying polysaccharides because an individual hydrogen atom in a large molecule is too small for accurate detection.

Identification of subunits. The study of polysaccharide structure usually focusses on the chemical composition, the linkage between the monosaccharide units, and the size and shape of the molecule. The last two properties can be ascertained by techniques usually applied to large molecules; *e.g.*, the most accurate molecular weight method measures the sedimentation properties of the molecule in an applied gravitational field (*e.g.*, the rate at which a solid material is deposited from a state of suspension or solution in a liquid). Indications of the shape of polysaccharide molecules in solution are obtained from viscosity measurements, in which the resistance of the molecules to flow (viscosity) is equated with the end-to-end length of the molecule; the viscosity of hyaluronic acid, for example, shows a marked dependence on both concentration of the acid and the salt content of the solution, and, under conditions approximating those found in biological systems, the molecule may be thought of as occupying a great deal of space. Alternatively, the compact nature of a glycogen molecule of equal molecular weight results in its accommodation to a much smaller space.

The identification of sugars in a mixture resulting from

the hydrolytic breakdown of a heteropolysaccharide is most often carried out by chromatography of the mixture on paper, silica gel, or cellulose. Ready separations can be achieved between pentoses, hexoses and, for example, deoxy sugars; closely related compounds such as D-glucose and D-galactose also can be separated using chromatographic techniques. The linkage positions in polysaccharides are usually determined using the methylation procedure described previously. The various monosaccharide methyl ethers are separated by gas–liquid chromatography.

Detailed statements about polysaccharide structure and function are limited by the statistical nature of some measurements (*e.g.,* branching frequency), the biological variability of parameters such as size and molecular weight, and incomplete information about associative interactions in living things. (E.A.D.)

LIPIDS

Lipids are a diverse group of organic compounds found in plants, in animals, and in microorganisms. They are greasy to the touch and insoluble in water but soluble in alcohol, ether, and other organic solvents. Lipids comprise one of the three large classes of foods and, with proteins and carbohydrates, are components of all living cells. The proportion of lipids in foodstuffs varies; it is 0.2 percent in white potatoes and 70 percent in some nut kernels. Fats such as olive oil and cod-liver oil contain a mixture of fatty substances, lipids called triglycerides. Triglycerides compose almost 90 percent of the adipose, or fat, tissue of animals. Contrary to popular opinion, adipose tissue is an energy source and can be used when needed. Triglycerides are sometimes called nature's storehouse of energy because, on a weight basis, they contain more than twice as much energy as do carbohydrates and proteins.

As noted previously, lipids made by cells are important not only because they serve as an energy source but also because they form structural components. Lipids such as lecithin and cephalin, which are soluble in both water and fats, serve a vital role in the cell by binding water-soluble compounds such as proteins to lipid-soluble substances. Lecithin is an important structural component of the cell membrane, where it maintains continuity between the water and lipid phases inside and outside the cell. Certain enzymes depend upon their attachment to lipids such as lecithin to perform their function properly.

Variation in lipid structure

The structure of lipids varies from simple chainlike molecules consisting of hydrogen, carbon, and oxygen to complex ring, or cyclic, structures with side chains of varying composition and complexity. Many naturally occurring lipids are associated with proteins, in combinations called lipoproteins. Lipids dealt with in the following section include neutral lipids (or triglycerides; *i.e.,* fatty-acid esters of the alcohol glycerol), phosphoglycerides (or phospholipids; *i.e.,* fatty-acid esters of glycerol and phosphoric acid or one of its derivatives), and sphingolipids (complex lipids containing compounds—sphingosine or phytosphingosine—other than glycerol). The characteristics of fatty acids, which are components of many lipids, also are described. Lipoproteins are dealt with briefly. Sterols and carotenoids, also included here as lipids, are considered in detail in the related article CHEMICAL COMPOUNDS: *Isoprenoids; Steroids.* Other related articles include METABOLISM and CELLS: *Biological membrane.*

General features

FUNCTION AND IDENTIFICATION OF LIPIDS

Lipids as food reserves in cells. The most important role of the fatty-acid components of neutral lipids in plant and animal tissues is to provide a fuel supply for cells; *i.e.,* neutral lipids comprise a reserve supply of potential energy and are broken down, when needed, in such a manner that the energy liberated is employed to make an energy-rich compound called adenosine triphosphate (ATP), which in turn is utilized in energy-requiring cellular processes such as muscle contraction and the synthesis of cell constituents. The energy in a fatty-acid molecule is transformed into ATP by a process known as fatty-acid oxidation (or beta oxidation). For a detailed discussion of the mechanism of fatty-acid oxidation, see METABOLISM.

Triglycerides and fatty acids are formed during digestive processes in animals. After a mammal ingests a fatty meal, the fats are acted upon by digestive secretions containing the enzyme lipase, which breaks down at least part of the triglycerides. The breakdown products and the remaining intact triglycerides then are absorbed through the intestinal cell wall and are recombined, at least in part, to form triglycerides and phospholipids. These lipids, in the form of very small droplets (chylomicrons), are transported in blood and in chyle (a milky fluid from the small intestine) to points of utilization or storage in the body.

One function of bile salts in digestion is to promote the linkage of (*i.e.,* emulsify) lipid-soluble groups with water-soluble ones (such as those in enzymes) and also to increase the solubility of lipids in water. Both emulsification and solubilization are necessary because lipids are completely metabolized only at the lipid-water interface created by bile salts and by the salts of fatty acids (soaps), which are formed during the partial breakdown of lipids (see also DIGESTION AND DIGESTIVE SYSTEMS).

If an animal ingests more energy-rich substances (*e.g.,* fats, carbohydrates) than it can utilize, excess fatty acids combine with glycerol to form neutral lipids, which are stored in the animal; *e.g.,* in adipose tissue in mammals. If the energy requirements of the animal increase, the stored neutral lipids may then be broken down, each molecule forming three molecules of fatty acid and one molecule of glycerol. The three molecules of fatty acid combine with a protein (albumin) in mammalian blood plasma and are carried in the bloodstream to various tissues and organs that require energy. Neutral lipids probably also function as depots of concentrated energy in plant reproductive structures such as pollen grains and seeds; *i.e.,* as food reserves for developing embryos.

Neutral lipids as depot material

The types of neutral lipids in an individual animal may vary according to the animal species and the composition of fats in the food it consumes. Fats used by or stored in animal tissues come from two sources—diet and enzymatic synthesis. The lipids synthesized from carbohydrates or proteins are characteristic of the animal species, whereas those resynthesized from dietary fats are characteristic of the food ingested. Many animals require some lipids containing one or more specific fatty acids, usually linoleic, linolenic, and arachidonic, to prevent the development of an essential fatty-acid deficiency, which is manifested by skin lesions, scaliness, poor hair growth, and low growth rates. These fatty acids cannot be synthesized by the animal and must be supplied in the diet (see also NUTRITION: *Nutritional diseases and disorders*).

Lipids as structural components of cells. *Neutral lipids.* Subcutaneous deposits of neutral lipids insulate animals against cold because of the low rate of heat transfer in fats, a property especially important to animals commonly found in cold waters or cold climates; *e.g.,* whales, walruses, and bears. Neutral lipids also provide structural support or padding for organs.

Phosphoglycerides. Phosphoglycerides, or phospholipids, are important constituents of cell membranes and of the membranes of cell components such as mitochondria; they function in the conduction of nerve impulses, in the insulation of nerve cells, in certain enzyme-catalyzed reactions within cells, and in blood coagulation and the transport of lipids from the liver to other tissues and organs in mammals.

Phosphoglycerides contain acidic (negatively charged) and basic (positively charged) groups as well as fatty-acid groups. Because they have both charged, water-attracting (hydrophilic) groups and lipid-attracting, water-repelling (hydrophobic) groups, phosphoglycerides are moderately soluble in both water and lipids and serve, therefore,

an important role in the cell in binding both types of compounds together; lecithin is a naturally occurring phosphoglyceride of special importance in the cell membrane because its hydrophilic groups maintain continuity between the water outside and that inside the cell, and its hydrophobic groups dissolve lipid materials, allowing them to enter the cell (see CELLS: *Biological membrane*).

Phosphoglycerides in blood bodies called platelets function in the process of blood-clot formation. Blood coagulation in mammals involves the conversion of a soluble protein (fibrinogen) into an insoluble derivative (fibrin clot). This reaction in part is catalyzed by the enzyme thrombin, which is derived from another protein (prothrombin) by the action of the enzyme prothrombinase. Prothrombinase in turn is formed by the interaction of two blood-plasma proteins, a phosphoglyceride, and calcium. The exact role of the phosphoglyceride in this reaction, however, has not yet been established with certainty.

Lipids in nervous tissue

Sphingolipids. Lipids found in nervous tissue (especially the brain) and necessary for its normal function include the sphingolipids, among them sphingomyelins, cerebrosides, sulfatides, and gangliosides. Schwann cells, the source for the membranous sheath surrounding certain nerve fibres, release a lipid material (myelin) that contains sphingomyelins. Myelination of nerve fibres is necessary for the normal development of nerve tissue, and absence or alteration of the myelination process may result in severe mental retardation in man. Sphingolipids have not been studied as exhaustively as some of the other lipids because they are difficult to isolate and separate into homogeneous components.

Sterols and carotenoids. Sterols include cholesterol in vertebrates and saponins and digitaloids in plants; cholesterol, with phospholipids, plays a role in membrane structure and an important role in the synthesis of numerous other biologically active sterols, commonly referred to as steroids—*e.g.*, bile acids, certain hormones, vitamin D. Carotenoids, widely distributed in living things, are pigments (coloured molecules) and function in the photosynthetic process in plants. These substances are discussed in detail in CHEMICAL COMPOUNDS: *Steroids; Isoprenoids.*

PREPARATION AND ANALYSIS

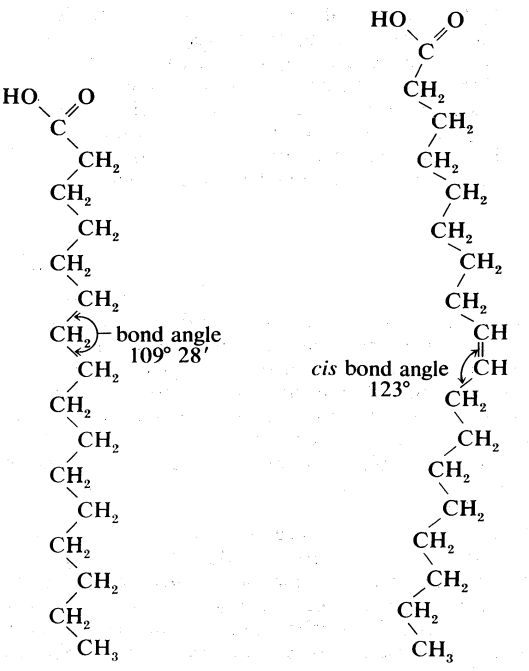
The isolation and identification of lipids is a challenging task. The procedures involve the separation of lipids from other cellular constituents using organic solvents (*e.g.*, chloroform, methanol), in which the lipids dissolve. Two solvents usually are used to separate lipids from other cellular constituents. One solvent, usually methanol, separates the lipoproteins into lipid and protein components; the other, usually chloroform, dissolves the lipids. After this mixture is treated to remove low-molecular-weight compounds, such as inorganic salts, amino acids, and sugars, the lipids are separated by column chromatography into individual components. A typical chromatographic separation of the total lipids, however, does not necessarily result in separation of the components into pure compounds. Additional analytical techniques, such as thin-layer chromatography, mass spectrometry, infrared spectrometry, and optical activity, are needed to establish the identity of each component.

Characteristics and classes of fatty acids

COMPOSITION AND STRUCTURE

The fatty acids of the naturally occurring lipids have an even number of carbon (C) atoms because they are synthesized from acetyl groups, each of which contains two carbon atoms. Fatty acids with 16 (palmitic acid) and 18 (stearic acid) carbon atoms are most commonly found in nature, but the reasons for their abundance have not yet been established. Fatty acids constitute important components of lipids in plants, animals, and microorganisms. In most cases, they are not found in free form but, instead, are bound to other compounds to form fatty-acid-containing lipids; *e.g.*, neutral lipids (triglycerides), sterols, phosphoglycerides such as lecithin, and sphingolipids such as sphingomyelin. Two typical fatty acids are palmitic and oleic.

Fatty-acid-containing lipids



palmitic acid oleic acid

The most stable arrangement of methylene groups ($-CH_2-$ groups), which comprise the hydrocarbon moiety of the molecule, is represented in the formula for palmitic acid; the double bond ($-CH=CH-$) in the oleic acid molecule, whose structure differs from that of palmitic acid only in the double bond between carbon atoms 9 and 10 (the carbon in the carboxyl, or $-COOH$, group is number 1), changes the shape of the molecule from that of palmitic acid. The double bond, which indicates that the acid contains two fewer hydrogen atoms than its counterpart (palmitic acid), makes the oleic acid molecule almost symmetrical because it is located exactly in the middle of the hydrocarbon chain. An additional change in the shape of fatty-acid molecules occurs in those acids that contain two double bonds, such as linoleic acid. Fatty-acid molecules containing one or more double bonds ($-CH=CH-$) are called unsaturated fatty acids; those without double bonds are called saturated fatty acids. Double bonds give rise to a phenomenon called geometrical isomerism; this means that the positions of certain atoms in the acid molecule may lie on the same side of an imaginary plane through the molecule, in which case a bond is said to be *cis*, or on opposite sides of the plane, in which case the bond is said to be *trans*. The unsaturated fatty acids of mammalian tissues usually contain *cis* double bonds (see structural formulas for palmitic and oleic acids); in addition, fatty acids found in mammals usually consist of long, chainlike arrangements of the hydrocarbon portion of the molecule (*e.g.*, as in palmitic acid) rather than branched or cyclic structures. On the other hand, fatty acids with cyclic and branched structures occur in plants and bacteria.

PRINCIPAL CLASSES

Saturated fatty acids. Typical naturally occurring saturated fatty acids are chainlike (nonbranched) compounds with an even number of carbon atoms; *e.g.*, palmitic acid, stearic acid. Names given to fatty acids other than the more commonly used, or trivial, names are called systematic names (*n*-hexadecanoic acid for palmitic and *n*-octadecanoic acid for stearic); they provide information about either the correct geometric form of an unsaturated fatty acid or the exact location of substituted groups in the molecule.

Although palmitic acid and stearic acid are the major saturated fatty acids found in animal and plant tissues, significant amounts of other saturated fatty acids, such as myristic acid and lauric acid, occur in certain tissues, and lignoceric acid and behenic acid are found in high

concentrations in brain sphingolipids. Small amounts of fatty acids with an odd number of carbon atoms also are known; *e.g.*, pentadecanoic acid and heptadecanoic acid. A list of the most common saturated chainlike fatty acids (*i.e.*, those containing 12 or more carbon atoms), including their usual sources, is presented in Table 10. There are several important short-chain fatty acids; these include butyric acid (which has four carbon atoms, or C_4) and caproic acid (C_6), which are important constituents of milk lipids, and octanoic acid (C_8) and decanoic acid (C_{10}), which are present in palm oil.

Table 10: Some Naturally Occurring Saturated Fatty Acids

trivial name	systematic name	chain length*	typical sources
Lauric acid	*n*-dodecanoic acid	12	palm-kernel oil, nutmeg
Myristic acid	*n*-tetradecanoic acid	14	palm-kernel oil, nutmeg
Palmitic acid	*n*-hexadecanoic acid	16	olive oil, animal lipids
Stearic acid	*n*-octadecanoic acid	18	cocoa butter, animal lipids
Behenic acid	*n*-docosanoic acid	22	brain tissue, radish oil
Lignoceric acid	*n*-tetracosanoic acid	24	brain tissue, carnauba wax

*In carbon atoms.

Unsaturated fatty acids. Unsaturated fatty acids have structures similar to those of saturated ones but contain at least one or more olefinic, or double bond. Fatty acids with one double bond are called monounsaturated fatty acids; those with two or more are called polyunsaturated ones. The naturally occuring unsaturated fatty acids may be distinguished solely by the degree of unsaturation and usually contain one or two double bonds; fatty acids containing acetylenic, or triple bonds ($-C\equiv C-$), also have been found in nature.

Mono- and polyunsaturated fatty acids

Monounsaturated fatty acids (also called monoethenoic, monoenoic, ethylenic, olefinic, or alkenoic acids) contain two fewer hydrogen atoms (H) than do corresponding saturated acids and thus have one double bond (*i.e.*, $-CH_2-CH_2$ becomes $-CH{=}CH-$); monounsaturated acids comprise the largest group of unsaturated fatty acids. The major and most representative monounsaturated acids in animal and plant tissues, oleic acid and palmitoleic acid, have the following formulas:

$$CH_3(CH_2)_7CH = CH(CH_2)_7COOH$$

oleic acid

$$CH_3(CH_2)_5CH = CH(CH_2)_7COOH.$$

palmitoleic acid

These two widely distributed fatty acids exist in abundant quantities in nature and have several common features; *i.e.*, a double bond between carbon atoms 9 and 10 (counting from the $-COOH$, or carboxyl, end) and a *cis* configuration at this double bond. Fatty acids with a double bond at other positions are listed in Table 11; *e.g.*, nervonic acid has a double bond between carbon atoms 15 and 16, as indicated by the systematic name, *cis*-15-tetracosenoic acid.

Table 11: Types of Monounsaturated Long-Chain Fatty Acids

trivial name	systematic name	chain length*	typical source(s)
Palmitoleic acid	*cis*-9-hexadecenoic acid	16	marine algae, pine oil
Oleic acid	*cis*-9-octadecenoic acid	18	animal tissues, olive oil
Gadoleic acid	*cis*-9-eicosenoic acid	20	fish oils (cod, sardine)
Erucic acid	*cis*-13-docosenoic acid	22	rapeseed oil
Nervonic acid	*cis*-15-tetracosenoic acid	24	elasmobranch fishes, brain

*In carbon atoms.

Polyunsaturated acids (also called polyethenoic, polyenoic, alkapolyenoic, alkadienoic, or alkatrienoic acids) are found in much smaller amounts in naturally occurring lipids than are monounsaturated ones. The formulas of the three common polyunsaturated fatty acids are given below.

$$CH_3(CH_2)_4CH = CHCH_2CH = CH(CH_2)_7COOH$$

linoleic acid

$$CH_3CH_2CH = CHCH_2CH = CHCH_2CH = CH(CH_2)_7COOH$$

linolenic acid

$$CH_3(CH_2)_4CH = CHCH_2CH = CHCH_2CH = CHCH_2CH = CH(CH_2)_3COOH.$$

arachidonic acid

Linoleic acid and arachidonic acid comprise most of the polyunsaturated fatty acids found in animal tissues. The double bonds usually are located at specific positions, as in linoleic and arachidonic acids. The most important polyunsaturated acids, which contain 18 to 22 carbon atoms, are listed in Table 12.

Table 12: Types of Polyunsaturated Fatty Acids

trivial name	systematic name	chain length*	typical source(s)
Linoleic acid	*cis*-9-, *cis*-12-octadecadienoic acid	18	corn oil, animal tissues, bacteria
Linolenic acid	*cis*-9-, *cis*-12-, *cis*-15-octadecatrienoic acid	18	animal tissues
	5,8,11-eicosatrienoic acid	20	
	8,11,14-eicosatrienoic acid	20	brain tissue
	7,10,13-docosatrienoic acid	22	phospholipids
	8,11,14-docosatrienoic acid	22	
Arachidonic acid	5,8,11,14-eicosatetraenoic acid	20	liver, brain tissue
	4,7,10,13-docosatetraenoic acid	22	brain tissue
	4,7,10,13,16,19-docosahexaenoic acid	22	brain tissue

*In carbon atoms.

Most mammals cannot synthesize linoleic acid, but they are able to convert it to other unsaturated acids; *e.g.*, arachidonic acid and the eicosapolyenoic acids. The polyunsaturated fatty acids in mammals therefore are derived largely from the diet, *i.e.*, they are essential fatty acids.

Alkyl-, hydroxy-, and alicyclic-substituted fatty acids

Fatty acids with substituted groups. Alkyl-substituted, or alkylalkanoic, fatty acids have one alkyl group on the molecule. An alkyl group is a hydrocarbon group with the general formula C_nH_{2n+1}; the methyl group ($-CH_3$) is the alkyl group in the example below. Methyltetradecanoic acid is found in butterfat.

$$CH_3CH_2\underset{\displaystyle CH_3}{\underset{|}{C}}HCH_2(CH_2)_9COOH$$

12-methyltetradecanoic acid (anteiso)

An excellent source of alkyl-substituted fatty acids is the bacterium that causes human tuberculosis; the acid is called tuberculostearic acid, or D-(*l*)-10-methyloctadecanoic acid. Other saturated acids isolated from tubercle bacilli include phytomonic acid (a 10- or 11-methylnonadecanoic acid), phthioic acid (a mixture containing fatty acids with 23 to 31 carbon atoms), mycocerosic acid (a polymethylated fatty acid with 31 carbon atoms), and a number of alkyl branched-chain, unsaturated fatty acids.

Several hydroxy fatty acids (*i.e.*, containing one $-OH$ group) occur in various lipid sources. Ricinoleic acid and cerebronic acid are typical examples.

$$CH_3(CH_2)_5\underset{\displaystyle OH}{\underset{|}{C}}HCH_2CH = CH(CH_2)_7COOH$$

ricinoleic acid
(*d*-12-hydroxy-*cis*-9-octadecanoic acid)

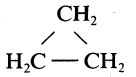
$$CH_3(CH_2)_{21}\underset{\displaystyle OH}{\underset{|}{C}}HCOOH$$

cerebronic acid
(2-hydroxytetracosanoic acid)

Ricinoleic acid occurs in castor oil, and cerebronic acid is present in brain tissue; α-hydroxynervonic acid (an un-

saturated fatty acid with 24 carbon atoms) also occurs in brain tissue. Very small amounts of other hydroxy fatty acids are found in animal and bacterial lipids.

Alicyclic-substituted fatty acids, which contain a cyclopropane group



attached at or near the centre of the molecule, occur in certain bacteria. One such acid, lactobacillic acid, comprises about 30 percent of the total fatty acids of the bacterium *Lactobacillus arabinosus*. Fatty acids containing cyclopropane groups have not yet been found in animal tissues. A series of compounds similar in structure has, however, been found in the chaulmoogra oils.

Several alicyclic acids, called prostaglandins, occur in seminal fluid and in certain glands of sheep and man. Hormone-like substances with widespread physiological effects, prostaglandins lower blood pressure, and stimulate the contraction of smooth muscles.

PHYSICAL AND CHEMICAL PROPERTIES

Solubility. Long-chain fatty acids (10 or more carbon atoms) are insoluble in water; short-chain fatty acids (two to eight carbon atoms) mix easily with water. The reason for the low solubility of long-chain fatty-acid molecules is that the methylene groups ($-CH_2-$) comprising the hydrocarbon chain are not ionized (*i.e.*, lack charge) and thus are nonpolar; they are more important in determining solubility properties of the acids than is the carboxyl ($-COOH$) component, which is ionized and therefore polar. Common table salt is about 1,000 times more soluble in water than is stearic acid; glucose is about 3,300 times more soluble. Fatty acids dissolve in nonpolar solvents or solvents that are less polar than water.

Reactivity of the carboxyl group. The low water solubility of fatty acids, which are members of a large group of organic acids known as carboxylic acids, makes difficult the determination of their properties as acids. Small amounts of fatty acids, which are weak acids compared with mineral acids such as hydrochloric and nitric, do not completely dissociate into ionized (charged) moieties in water. The carboxyl group ($-COOH$) is responsible for the acidic properties of fatty acids (see also CHEMICAL COMPOUNDS: *Carboxylic acids and their derivatives*).

Chlorides, esters, and amides

Most reactions at the carboxyl group of fatty acids involve the hydroxyl moiety ($-OH$) but are influenced by the carbonyl moiety ($-C{=}O$). Several reactions are typical of fatty acids; *e.g.*, formation of chlorides, esters, and amides. Fatty-acid chlorides, called acyl chlorides, are formed by the reaction of a fatty acid with the compound thionyl chloride ($SOCl_2$). Salts of fatty acids are formed in reactions with bases (*e.g.*, sodium hydroxide), in which case the products are called soaps, or with alcohols to form esters. If palmitic acid reacts with the alcohol methanol (CH_3OH), the product, called methyl palmitate, is the methyl ester of palmitic acid; the methyl esters of fatty acids are often used to determine the properties of the acids. The reaction of a fatty acid with ammonia in the presence of heat results in the formation of a product called an amide; the amides of fatty acids also are useful in characterizing them. Amides are found in nature.

Reactivity of the hydrocarbon chain. The most chemically reactive part of a hydrocarbon chain of a fatty acid is the double bond. Halogens (*e.g.*, chlorine, bromine, iodine) and their derivatives, called halides, react with unsaturated fatty acids to remove the double bonds, thereby forming saturated acids; this is known as an addition reaction. Halides such as hydriodic acid (HI) readily attack a double bond to form derivative compounds.

In the presence of a suitable catalyst (*e.g.*, platinum oxide, palladium on charcoal), hydrogen undergoes an addition reaction with unsaturated fatty acids; that is, hydrogen is added to the positions at which double bonds are found. This hydrogenation reaction results in the formation of saturated fatty acids. The double bonds of polyunsaturated fatty acids are hydrogenated in a specific order; linolenic acid (a fatty acid with 18 carbon atoms and three double bonds; Table 12), for example, is hydrogenated first at the bond between carbon atoms 12 and 13 to form a 9,15-octadecadienoic acid, which can then be further hydrogenated to form a saturated acid.

Bond alternation, or any change involving double bonds such as those found in unsaturated fatty acids, also occurs in living cells. The first type, called autoxidation, is a nonenzymatic process; the second, lipoxidation, is an enzyme-catalyzed reaction. Both involve oxygen.

Autoxidation

Autoxidation, or rancidification, which occurs slowly and spontaneously in air, involves the absorption of oxygen by an unsaturated fatty-acid molecule, with the formation of a compound called a hydroperoxide, which decomposes. Autoxidation may occur during the drying or handling of certain oils, under adverse conditions of fat storage, and in certain mammalian deficiencies; *i.e.*, vitamin E deficiency. The acids that undergo autoxidation usually are the polyunsaturated ones; *e.g.*, linoleic acid.

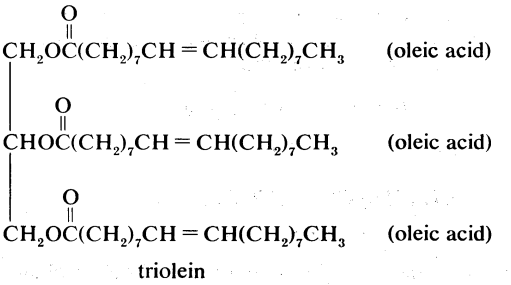
Lipoxidation contrasts with autoxidation in that an enzyme (lipoxidase) catalyzes the addition of oxygen to a double bond in linoleic acid. Lipoxidase is found in legume seeds (for example, soybeans) and in adipose (fat) tissue.

Derivatives of fatty acids and associated compounds

As was previously emphasized, naturally occurring fatty acids are normally found in cells bound to other substances to form triglycerides (neutral lipids), phosphoglycerides (phospholipids), and sphingolipids. Neutral lipids are found almost exclusively in the cytoplasmic compartment of cells, phosphoglycerides and sphingolipids almost exclusively in the membranous structures. For metabolic aspects of these lipids, see METABOLISM.

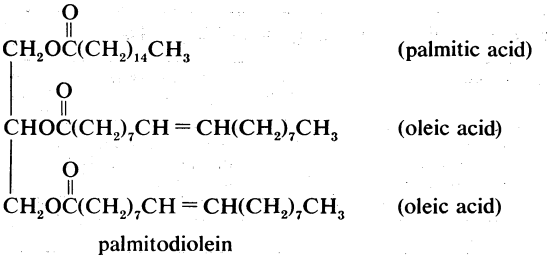
NEUTRAL LIPIDS

General features. Triglycerides, or neutral lipids, are compounds consisting of glycerol and three fatty acids; the structural formula of a typical triglyceride, triolein, is shown below.



triolein

Triolein is called a simple triglyceride since it contains only one type of fatty acid (oleic acid). Naturally occurring triglycerides usually are mixed triglycerides; *i.e.*, they contain more than one type of fatty acid. An example of a mixed triglyceride is palmitodiolein, the fatty-acid composition of which is, as the name indicates, one molecule of palmitic acid and two molecules of oleic acid.

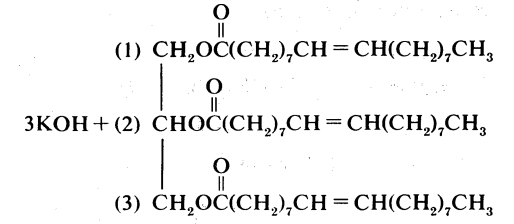


palmitodiolein

This triglyceride may have structural arrangements other than the one shown; *i.e.*, the fatty-acid molecules may be arranged with palmitic acid occupying any of the two possible different positions. (There are only two chemically different positions since the top and the bottom positions in the above formula are undistinguishable.)

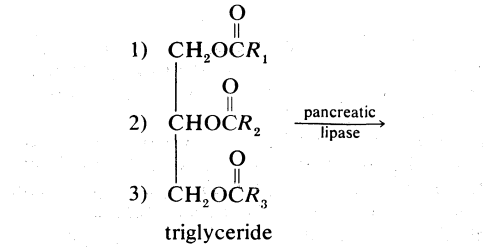
Physical and chemical characteristics. Triglycerides, called neutral lipids because they possess no charged groups, do not migrate in an electrical field and are insoluble in water. Information about the structure of naturally occurring mixed triglycerides containing different fatty acids may be obtained by chemical, enzymatic, and physical methods. Chemical and enzymatic techniques are summarized below.

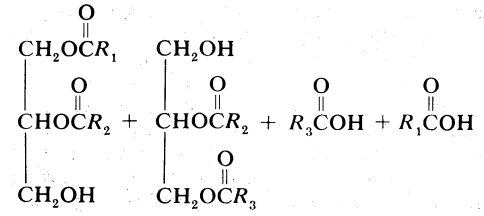
Triglyceride molecules undergo reactions with bases such as potassium hydroxide (KOH). The reaction, shown below, often called a saponification reaction, results in the formation of an alkali soap (potassium oleate) and glycerol.

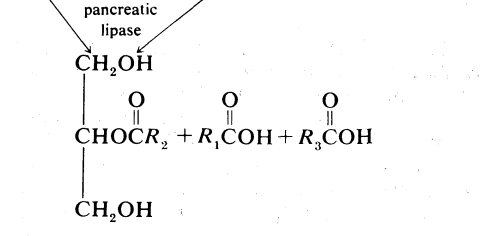


If acid is added to a solution containing the soap, free fatty acids are formed. The saponification reaction helps establish the general nature of the fatty acids associated with a specific triglyceride, but provides no information on their specific locations on the glycerol molecule [positions (1), (2), or (3) in the triolein formula above].

The distribution of fatty acids on the glycerol molecule is determined by using an enzyme (pancreatic lipase) that is found in mammals and plays a role in lipid digestion. Pancreatic lipase acts only at positions (1) and (3) of the triglyceride molecule in the sequence outlined below; the overall reaction occurs in two steps. The first products of the lipase reaction are two molecules of glycerol, each containing two fatty acids (labeled diglycerides I and II) and two molecules of free fatty acids (represented by the general formula *R*COOH) from positions (1) and (3) of the triglyceride. The lipase next reacts with the two diglycerides. A monoglyceride that contains one fatty acid at position (2), and free fatty acids from positions (1) and (3) of the diglycerides are the products of this second reaction. R_1, R_2, and R_3 in the reaction sequence represent hydrocarbon chains of fatty acid ester groups. The various products of the enzyme-catalyzed reactions may be isolated and analyzed to provide information about the chemical nature of the fatty acids at each of the three positions.

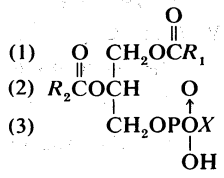






PHOSPHOGLYCERIDES

The most simple phosphoglyceride contains fatty acids combined with glycerophosphate; *i.e.*, glycerol, containing a molecule of phosphoric acid at position (3), as shown below in the general formula. R_1 and R_2 represent hydrocarbon chains of fatty-acid ester groups at positions (1) and (2) of glycerol; *X* at position (3) may be a hydrogen atom or an organic compound (*e.g.*, serine, inositol,

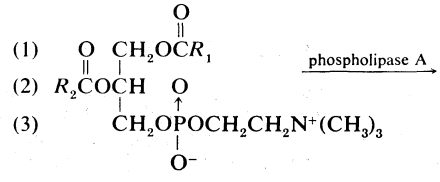


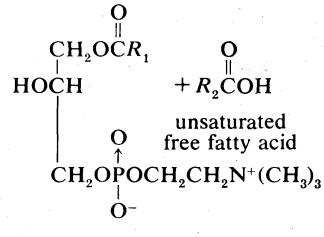
choline, glycerol, glycerophosphoric acid, or a diglyceride). Variations from the general structure occur; phospholipids called plasmalogens, for example, contain an unsaturated ether group in position (1) instead of a fatty-acid ester group, and other phospholipids called glyceryl ether phospholipids contain a saturated ether group in position (1). The properties and characteristics of several phosphoglycerides are summarized in the following sections.

The term phosphatidyl is used to describe the portion of the phosphoglyceride molecule containing the fatty-acid ester groups and glycerophosphate.

Phosphatidyl choline. Phosphatidyl choline, also known as lecithin, is the difatty-acid derivative of glyceryl-3-phosphorylcholine and is perhaps the most representative of all the phosphoglycerides. A lecithin molecule has a nonpolar portion (the two fatty-acid ester groups), which is insoluble in water, and a polar portion (phosphorylcholine), which is soluble in water. Since lecithin molecules contain both polar and nonpolar components, they can orient themselves in a specific way at oil–water or water–air interfaces. The nitrogen-containing component of lecithin (choline) is a strongly basic compound and has a positive charge; since phosphoric acid has a negative charge, lecithin has both acidic and basic groups and behaves as a zwitterion (*i.e.*, a dipolar ion).

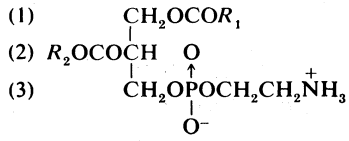
In animal tissues, the fatty-acid ester groups of lecithin are specific with regard to type and position; *e.g.*, reactions (see example below) using the enzyme phospholipase A





(from snake venom) as a catalyst show that lecithins from different sources have saturated fatty acids predominately at position (1) and unsaturated fatty acids at position (2). The action of this enzyme on lecithin results in the liberation of the unsaturated fatty acid at position (2); the remainder of the molecule is called a lysolecithin. R_1 and R_2 represent hydrocarbon chains of fatty-acid ester groups in the equation. The reaction proceeds most efficiently in a diethyl ether solution, which removes liberated fatty acids from the surfaces of the enzyme molecules. The free fatty acid represents the unsaturated fatty-acid ester group at position (2) in lecithin; the lysolecithin has a saturated fatty acid ester group at position (1). The specificity of the fatty-acid positions may act in metabolic utilization.

Phosphatidyl ethanolamine. The structure of phosphatidyl ethanolamine, found in most sources containing phosphatidyl choline, is represented below; R_1 and R_2 represent hydrocarbon chains of fatty-acid ester groups. The fatty acids in phosphatidyl ethanolamine, as in phosphatidyl choline, usually occupy specific positions;



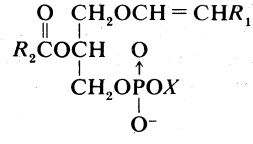
phosphatidyl ethanolamine

saturated fatty acids are located primarily at position (1), unsaturated fatty acids at position (2). Ethanolamine ($\overset{+}{N}H_3CH_2CH_2OH$) is a nitrogen-containing basic compound derived from the amino acid serine. This phosphoglyceride, unlike phosphatidyl choline, is not a zwitterion and exists in various ionic, or charged, forms.

Phosphatidyl serine. Although amino acids other than serine may occur in natural phospholipids, phosphatidyl serine is the most representative one. It is a minor chemical component of cells, compared with phosphatidyl choline and phosphatidyl ethanolamine.

Plasmalogens. Plasmalogens are naturally occurring phosphoglycerides that, under certain conditions, release an organic compound called an aldehyde. The substituent at the (1) position in the general formula below is called a vinyl ether ($-OCH{=}CHR_1$); that on position (2) is a fatty-acid ester, usually of an unsaturated fatty acid. The nitrogen base (*X*) in most plasmalogens is ethanolamine; choline-containing plasmalogens occur in lesser amounts.

Another phospholipid, similar in general structure to plasmalogens, is found in eggs, brain tissue, and red blood cells of cows; it contains a saturated ether ($-OCH_2CH_2R$) substituent instead of a vinyl ether.



general structure of plasmalogens

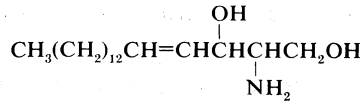
Phosphatidyl inositol. Phosphatidyl inositol, which is not present in large amounts in animals or plants, contains an alcohol (inositol) with a ring structure. Several similar compounds have been isolated from brain tissue, including phosphoglycerides containing inositol polyphosphates; *e.g.*, L-myo-inositol-1,4,5-triphosphate and L-myo-inositol-2,4,-5-triphosphate. More complex inositol-containing phospholipids may contain ethanolamine, tartaric acid, or the sugar galactose.

Minor phosphoglycerides. Not all of the glycerol-containing phospholipids are represented by the structures already described. Minor phosphoglycerides (*e.g.*, phosphatidic acids, phosphatidylglycerol, and bisphosphatidic acid), and usually are present in cells in very small amounts, less than 2 or 3 percent, and often occur in tissues as salts. Phosphatidic acids, which are glycerol phosphate molecules containing two molecules of fatty acid, are important in the biosynthetic pathways that result in the formation of phosphoglycerides and neutral lipids. Small amounts of phosphatidic acids occur in mammals, but they are abundant in plants (*e.g.*, cabbage). Cardiolipin, or diphosphatidylglycerol, which is found in heart muscle, usually contains only unsaturated fatty acids; it has been used in the diagnosis of syphilis. Phosphatidylglycerol, which occurs in a variety of sources, may be involved in both the synthesis and breakdown of phosphatidic acids.

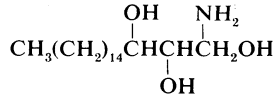
Phosphatidic acids

SPHINGOLIPIDS

The sphingolipids, a group of complex lipids, contain either sphingosine or the closely related compound phytosphingosine instead of glycerol.

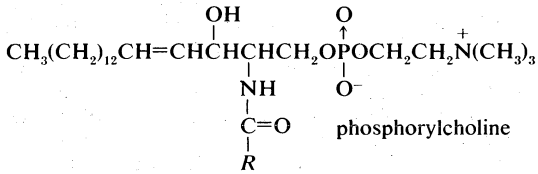


sphingosine



phytosphingosine

Sphingosine is found in the sphingolipids of animals, phytosphingosine in those of plants. Animal sphingolipids occur in high concentration in the brain, in the peripheral nerves, and in the myelin sheaths surrounding nerve fibres; they occur in lower amounts in several other tissues; *e.g.*, liver, blood plasma, kidney. Sphingomyelins are the most frequently encountered sphingolipids in animals; *R* represents the hydrocarbon chain of a fatty-acid ester group in the formula below.



a sphingomyelin

Other sphingolipids include glycosphingolipids, or glycolipids, which contain a sugar instead of the phosphoryl choline unit. A typical glycosphingolipid, present primarily in brain tissue, is the *O*-galactosido derivative of *N*-acyl sphingosine; it is generally called a cerebroside. The sugar component of the molecule is galactose, a hexose sugar (*i.e.*, containing six carbon atoms). The fatty acids found in cerebrosides usually contain 24 carbon atoms and have either a double bond or a hydroxyl group. Some cerebrosides contain sulfuric acid linked to galactose; these molecules are called sulfatides. Other glycosphingolipids are called gangliosides; they differ from the cerebrosides in the more complex chemical nature of the sugar group on the sphingosine molecule. The various ganglioside molecules usually differ in the number of sugar molecules.

Sialic acids

The neuraminic acid component, usually present as *N*-acetyl-*O*-*N*-diacetyl-, or *N*-glycolylneuraminic acid, is commonly known as a sialic acid. Naturally occurring sialic acids, which are found as components of gangliosides and other complex molecules, are widely distributed in animal tissues and in bacteria.

ASSOCIATED COMPOUNDS: STEROLS AND CAROTENOIDS

Widely distributed in nature, sterols and carotenoids usually are closely associated with the fatty-acid-containing lipids in cells. The most commonly occurring natural sterol in vertebrates is cholesterol, which is the starting substance, or precursor, for many other important biologically active sterols (often referred to as steroids). The carotenoids, which are sometimes referred to as tetraterpenes (C_{40}), are widely distributed in plants and microorganisms and to a lesser extent in animals, in which they

probably originate in dietary sources, since animals cannot synthesize them. A frequently encountered carotenoid is β-carotene, a precursor of vitamin A alcohol, which functions in mammalian vision. Other carotenoids function in photosynthesis in plants. These lipids are discussed in detail in CHEMICAL COMPOUNDS: *Isoprenoids; Steroids.*

Lipoproteins

Lipids in living cells normally are associated with proteins. Lipid–protein complexes, or lipoproteins, occur either as soluble components (*e.g.*, those in the blood plasma of mammals or in egg yolk) or as more insoluble types (*e.g.*, those in membranes of cells). The discussion that follows is concerned primarily with composition, structure, and biological behaviour of the soluble lipoproteins in blood plasma, which are relatively easy to isolate and purify and thus have been studied extensively.

One outstanding feature of the association between lipids and proteins is the lack of significant chemical bonds between them. Since lipids can be separated from the protein component, it is assumed that lipoproteins are held together by weak physical forces.

Lipoproteins isolated from various sources usually fall into one of two general groups. In one group are the highly ordered lipoproteins, which are characterized by high protein and low lipid content; the other group consists of disorganized lipoproteins, which are characterized by low protein and high lipid content. A large number of different lipoproteins have been isolated; their molecular weights and lipid contents vary over a wide range. A suitable characterization of a lipoprotein includes several criteria; *e.g.*, solubility characteristics, behaviour when spun in a centrifuge, and chemical composition.

PHYSICAL CHARACTERISTICS

The physical character of lipoproteins resembles that of proteins. Lipoproteins that differ with respect to physical properties can be separated by differences in solubility, in electrophoretic behaviour (migration of a charged particle in an electrical field), and in centrifugal behaviour. A well-established method for separating proteins (salt fractionation) is also useful for separating lipoproteins; it has been replaced to a great extent, however, by a more effective method, centrifugation, mentioned below. Another method now largely replaced by centrifugation involves the separation, using ethanol, or protein components in human blood plasma to obtain two major lipoprotein fractions, designated as α_1-lipoprotein and β-lipoprotein.

Centrifugation

Lipoproteins were first identified in human blood when it was spun in a centrifuge at extremely high speeds (ultracentrifugation). Lipoproteins separate from other serum constitutents after 20 to 24 hours of centrifugation. The lipoproteins gather as a uniform layer at the top of the centrifuge tube. If the density of a solution containing the serum lipoproteins is adjusted by the addition of a soluble material such as a salt, so that the solution at the bottom of the tube has a higher density than that at the top, and if the mixture then is spun in an ultracentrifuge, the lipoproteins become distributed according to the densities of the components. The degree of separation of various lipoproteins depends on the density adjustment of the solution before ultracentrifugation.

The physical characterization of low-density lipoproteins (*i.e.*, high lipid content) is often expressed in S_f values; S_f indicates low density lipoprotein flotation rate in a sodium chloride solution of density 1.063 grams per cubic centimetre at 26° C (79° F) in a unit centrifugal field and is described in Svedberg units (10^{-13} centimetre/second/dyne/gram).

CHEMICAL NATURE

The chemical nature of the three groups of human blood-plasma lipoproteins, the very low density lipoproteins (VLDL), the low-density lipoproteins (LDL_1), and the high-density lipoproteins (HDL_3), which differ from each other in various ways (*e.g.*, types and amounts of both lipids and proteins, immunochemical properties), are summarized in Table 13.

Table 13: Characteristics of Human Blood Plasma Lipoproteins
(a composite picture)

	HDL_3	LDL_1	VLDL
Lipid, in percent	37	48	90
Component lipids, in percent of total lipid			
Phosphatidyl choline	44	24	20
Sphingomyelin	20	16	56
Triglyceride	24	12	8
Cholesterol ester	6	12	8
Cholesterol	6	1	1
Free fatty acid	45	15	10
Protein, in percent	18	37	...
Water, in percent	84	185	500
Diameter, in angstrom units	1.13	1.04	0.91
Density, hydrated (grams per cubic centimetre)	300,000	2×10^6	8×10^6
Molecular weight (approx.)			

Variations in age and sex of individuals are reflected by differences in distribution of low-density lipoproteins (LDL_1) and high-density lipoproteins (HDL_3) in their blood serums. The chylomicrons occur as transient components produced during the intestinal absorption of lipids.

(D.J.H./Ed.)

NUCLEIC ACIDS

Nucleic acids are of interest because they provide the genetic material of the cell and, by directing the process of protein synthesis, determine its inherited characteristics. Nucleic acids are naturally occurring complex phosphorus compounds, acidic in character, and capable of being broken down chemically to yield phosphoric acid, sugars, and a mixture of organic bases (purines and pyrimidines). About 1868, nuclei isolated from pus cells were found to contain an unusual phosphorus compound, which was named nuclein. In later years, complex phosphorus-containing acid materials were also isolated from a wide variety of cells; these appeared to be chemically similar to nuclein and came to be called nucleic acids.

General considerations

CLASSIFICATION

There are two classes of nucleic acids: ribonucleic acids (RNA) and deoxyribonucleic acids (DNA). Both RNA and DNA are macromolecules that can be distinguished from each other by their base and sugar contents. DNA, a major constituent of chromosomes in the nuclei of all cells, is also found in other cellular components (*e.g.*, mitochondria). RNA is present in both the nucleus and cytoplasm of many cells. The bulk of cytoplasmic RNA is associated with ribosomes, small particles composed of RNA and proteins that are the site of protein synthesis.

BASIC COMPONENTS

Polynucleotides

Nucleic acids are polynucleotides, long chain compounds consisting of repeating structural units called nucleotides (Figure 11). They may be composed of more than 1,000,000 of these nucleotides. The nucleotides themselves consist of three subunits. Each of them contains a pentose (or five-carbon) sugar, a purine or pyrimidine base, and a phosphate residue. The pentose sugar is ribose in RNA and 2-deoxyribose in DNA. The major purine and pyrimidine bases are adenine, guanine, cytosine, and thymine (or uracil in RNA).

The nucleotides that serve as physiological building blocks in the synthesis of nucleic acids often are esters (when an acid and an alcohol react they form water and an ester) not of phosphoric acid but of condensed relatives of it, pyrophosphoric or triphosphoric acids.

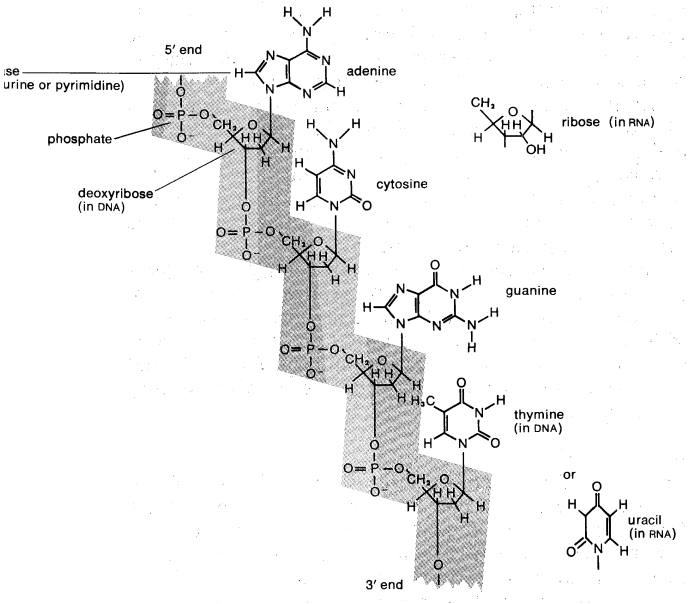


Figure 11: Portion of polynucleotide chain of deoxyribonucleic acid.

From J.D. Watson, *Molecular Biology of the Gene*, copyright © (1965), W.A. Benjamin, Inc., Menlo Park, California

Enzymatic breakdown of nucleic acids results in the release of nucleotide units in which a phosphate residue is attached to the deoxyribose sugar at one of two possible positions on the sugar molecule. These internucleotide linkages are known as phosphodiester bonds. The internucleotide linkage is the same in RNA.

It was the fact that adenine and thymine are present in approximately equal amounts in DNA, as are guanine and cytosine, together with information from X-ray crystallography of DNA that led Nobel Prize winners J.D. Watson and F.H.C. Crick to postulate that the DNA molecule consists of two chains or strands of polynucleotides coiled around each other to form a double helix, the bases of one strand being paired with complementary bases of the other by hydrogen bonds: adenine paired with thymine and cytosine with guanine. RNA has unequal proportions of the four bases and, in addition, contains unusual bases (*e.g.*, pseudouridine, various methylated purines).

Nucleic acids can be separated from other cellular constituents by treating a tissue with cold acid. Most of the contents of the tissue go into solution, but both RNA and DNA are insoluble in cold acid and can be removed from the mixture. RNA and DNA, in turn, can be separated by treatment with alkali; in this case, RNA is broken down by the alkali into nucleotide units, but DNA remains unchanged. If acid is added to this mixture, DNA precipitates (comes out of solution), and the RNA nucleotide units remain in solution.

The DNA found in cell nuclei (as chromosomes) and in other cell components usually is bound to proteins called histones. The protein bound to the DNA in sperm cells is known as protamine. Bacterial DNA is not associated with protein. DNA–protein complexes from animal tissues, known as deoxyribonucleoproteins, are molecules of very high molecular weight (from 10,000,000 to more than 100,000,000 times the weight of a hydrogen atom). Approximately 25 to 50 percent of the dry weight of this complex is nucleic acid.

Characteristics of DNA

PROPERTIES, STRUCTURE, AND BASE SEQUENCE

Physical properties. Many of the physical properties of DNA depend on the methods used to purify it. If, for example, fragmentation of DNA molecules by either mechanical means or enzymes is avoided, DNA preparations of very high molecular weight (*e.g.*, 120,000,000), corresponding to a chromosome or to the entire nucleic acid complement of a virus or a bacterium, can be obtained. Solutions of such preparations are viscous, and almost any manipulation (*e.g.*, stirring) can break the molecules into fragments of lower molecular weight.

Because of their purine and pyrimidine base content, DNA preparations absorb ultraviolet light. If a solution of DNA is heated to a critical temperature, it absorbs more ultraviolet light and becomes less viscous because the orderly helical structure of the molecules breaks down. DNA molecules become irregularly coiled structures during this process, which is called thermal denaturation of DNA. The "melting" temperature at which thermal denaturation takes place is dependent on the base composition of the DNA molecule. DNA molecules with a high content of guanine and cytosine, for example, are more stable at high temperatures than those high in adenine and thymine because the hydrogen bonding between guanine and cytosine is stronger. At temperatures slightly above the melting temperature, the two strands of DNA separate. If the DNA solution is cooled slowly, sometimes recombination of the strands and partial restoration of the double helix occurs.

If a concentrated salt solution is spun in a centrifuge at high speed for a long period of time, an equilibrium is eventually attained in which the concentration, and, therefore, the density, of the salt solution increases gradually from the top of the tube to the bottom. If the salt solution also contains DNA, these molecules migrate to the level in the tube in which the density of the salt solution equals their own. This is an important method for determining with great precision the density of nucleic acids and for separating DNA molecules of different densities. The density of DNA molecules with a high proportion of guanine and cytosine is slightly greater than that of molecules with a lower proportion of these bases. Denaturation also causes a slight increase in density. Although DNA usually is double-stranded, there are single-stranded forms (for example, the small bacterial virus designated φX 174), which do not show the denaturation phenomena of the double-stranded molecules.

Importance of hydrogen bonds

Structure. The long polynucleotide chains that comprise DNA molecules would form flexible threadlike molecules, instead of coils, were it not for cross-links (hydrogen bonds) between bases of each chain. Measurements indicate that the space between DNA chains agrees with values calculated for hydrogen bond linkage of a purine base to a pyrimidine base (the space is too small for two purines and too large for two pyrimidines). Three crosslinks occur between a guanine residue (purine base) and its complementary base, cytosine (a pyrimidine), in two complementary chains; on the other hand, two crosslinks occur between adenine (purine base) and its complementary base, thymine (a pyrimidine). Each purine base is linked to its complementary pyrimidine base in the opposite chain by either two or three hydrogen bonds.

Double helix

The sugar molecules (deoxyribose in DNA) are linked by phosphodiester bonds to form the backbone of each DNA chain. The bases, with their large hydrophobic (water-hating) surfaces, are stacked together on the inside of the molecule like a pile of coins. These two complementary chains coil around each other to form a double helix. In other words, DNA resembles a spiral staircase in which the

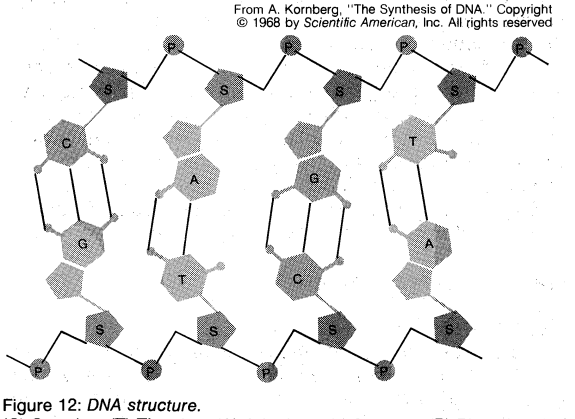


Figure 12: *DNA structure.*
(C) Cytosine. (T) Thymine. (A) Adenine. (G) Guanine. (P) Phosphate. (S) Deoxyribose (sugar).

bases are the steps and the sugar phosphate residues are the bannisters. The DNA molecule derives stability from hydrogen bonding between base pairs in complementary chains from hydrophobic "base-stacking forces" between neighbouring bases on the same chain. The DNA of mammalian chromosomes and DNA-containing viruses generally consist of long unbranched double-helical threads of the type described above. The DNA of many viruses (including φX 174) and of mitochondria are circular, and, in some cases, there is evidence that such cyclic double helices are twisted into a supercoiled form. A single break in either chain of the helix releases the molecule, and an untwisted structure results.

Base sequence. The biological function of DNA molecules, whether single- or double-stranded, cyclic or unbranched, is to provide a genetic message encoded in a sequence of purine and pyrimidine bases. Although there are ways to determine the sequences of the units in some macromolecules like proteins (see *Proteins*), it is extremely difficult to do so for DNA molecules, which generally are of much higher molecular weight. In addition, nucleic acids are composed of fewer kinds of units (*i.e.*, the purine and pyrimidine nucleotides) than are found in proteins (*i.e.*, about 20 amino acids). It is also more difficult to obtain a single molecule of pure DNA than it is to purify proteins. Progress has, however, been made in the base sequence determination of the relatively small DNA of the φX 174 virus, which contains about 5,500 nucleotide units; thus far it has been possible to obtain only statistical evidence abut the order in which these nucleotides are arranged. The usual experimental approach takes advantage of the fact tha the purine bases of DNA are removed by acid, while the pyrimidine bases retain their original positions in the molecule. Pyrimidine nucleotides have been shown to occur frequently in clusters.

LABORATORY SYNTHESIS

Experiments with living cells have tended to confirm the hypothesis that DNA is synthesized by a process in which the two strands of an existing molecule separate while new complementary strands are synthesized using the old strands as templates, or patterns. For a detailed discussion of the biosynthesis of DNA, see GENETICS AND HEREDITY: *The gene.* The product of DNA synthesis in the cell is the formation of two molecules, each containing one strand from the original DNA and a newly synthesized complementary strand.

Enzymes purified from mammalian and bacterial cells catalyze the synthesis of DNA in the test tube. The reaction takes place only in the presence of the nucleoside triphosphates of the four bases (adenine, cytosine, guanine, and thymine) and preformed DNA, which serves as a template. That the product has the same nucleotide sequence as the template has been demonstrated by the technique of "nearest neighbour frequency analysis." One of the four necleotides contains radioactive phosphorus. During DNA synthesis these radioactive phosphorus residues form internucleotide links with nucleotides next to them in the newly forming chain. The newly synthesized DNA is degraded with enzymes that split the radioactive phosphorus from the molecule to which it was originally attached. The result is that the radioactive phosphorus becomes part of the structure of the nucleotide nearest the necleotide originally labelled. In this way it is possible to determine the frequency with which particular pairs of nucleotides are neighbours in polynucleotide chains. Such nearest neighbour frequencies vary markedly from one DNA source to another and are specific for each kind of DNA. Generally, the DNA produced by the DNA-synthesizing enzyme in the test tube cannot be shown to possess the biological activity of the template.

Nearest neighbour frequency analysis

When single-stranded circular DNA of the virus φX 174 is used as a template for DNA synthesis, however, a complementary (or −) strand is formed along the original (or +) strand. The free ends of the new (or −) strand can be joined by an enzyme to form a cyclic double-stranded molecule. If the cyclic double-stranded molecules are broken with another enzyme, random single breaks occur in a proportion of the molecules in either strand. The broken molecules can be eliminated leaving a number of intact synthetic cyclic (−) strands. When these are tested, they are found to be capable of infecting susceptible cells and of giving rise to a new generation of normal virus particles.

Characteristics of RNA

PROPERTIES, STRUCTURE, AND BASE SEQUENCE

Physical properties. Most DNA molecules share a common genetic function and structural pattern. RNA molecules, however, perform several functions in the cell, and their properties vary correspondingly. Three types of RNA have been studied chemically: ribosomal RNA (rRNA), transfer (soluble or adaptor) RNA (tRNA), and viral RNA. Messenger RNA, despite its biological importance, has not yet been the subject of extensive chemical investigation. The RNA's of some viruses (*e.g.*, reovirus, wound tumour virus) have properties in common with DNA, including a double helical structure and a critical melting temperature. Other viral RNA's (*e.g.*, tobacco mosaic virus) and ribosomal RNA, however, have less sharp melting characteristics and lack a complete double helical structure; instead the polynucleotide chains are folded back on themselves to provide small regions of base pairing and, therefore, hydrogen bonding and some helical sections. The small tRNA molecules are thought to be folded similarly into a cloverleaf structure, with regions of hydrogen bonding between base pairs forming hairpin-like helical sections.

The molecular weights of RNA molecules vary according to their type and source. Ribosomal particles, for example, contain a large fraction of the RNA in a cell; these particles can be split into two unequal subunits. The main component of ribosomal RNA from the larger of the subunits has a molecular weight of about 1,000,000; the RNA from the smaller subunit has a molecular weight of about 800,000. Ribosomes also contain RNA's of lower molecular weight (about 60,000). Transfer RNA is the smallest type of RNA molecule; its molecular weight varies from about 25,000 to 30,000. The molecular weights of messenger RNA molecules vary over a wide range.

Structure and base sequence. Research involving RNA structure has centred around tRNA (the type of RNA to which amino acids are attached during protein biosynthesis), the smallest nucleic acids (70 to 80 nucleotide units). The different tRNA molecules (each of the approximately 20 amino acids has at least one specific tRNA) in the cell can be separated and purified. In addition to their small size, tRNA molecules are distinguished from other forms of nucleic acids by their relatively high content of minor nucleotides (*e.g.*, pseudouridine, methylated bases). The relative ease with which these minor nucleotides can be identified has been useful in sequence determinations. By using enzymes that catalyze different specific reactions for various lengths of time on one type of tRNA, fragments of the polynucleotide are produced and can be separated and analyzed to determine the sequence of nucleotides in the intact molecule.

LABORATORY SYNTHESIS

It has been suggested that RNA is synthesized in the living cell by a process in which one of the two strands of a DNA molecule is used as a template (or pattern) for the formation of a complementary strand of RNA. For specifics regarding the biosynthesis of RNA, see GENETICS AND HEREDITY: *The gene.* Enzymes found in microorganisms and in animal tissues can catalyze the synthesis of RNA in the test tube. This synthesis occurs provided that a mixture of the proper nucleoside triphosphates and a small quantity of DNA are present. The composition of the newly synthesized RNA corresponds exactly to that of the DNA template (with uracil replacing thymine), and nearest neighbour sequence analysis indicates a similar correspondence in base sequence. In certain cases, a hybrid double helical molecule, which may form from template DNA and the RNA product, indicates an exact correspondence in structure between primer and product. In animal and bacterial cells infected with RNA viruses, the RNA of the virus acts as a template for the synthesis of more RNA.

(J.N.D./R.Y.T./Ed.)

VITAMINS

The substances commonly known as vitamins are diverse in chemical structure and function. Originally defined as organic compounds obtainable in a normal diet and capable of maintaining life and promoting growth, vitamins are distinct from carbohydrates, fats, and proteins in function, as well as in the quantities in which organisms require them. In general terms, vitamins are organic substances that usually are separated into water-soluble *(e.g.,* the B vitamins, vitamin C) and fat-soluble (*e.g.,* vitamins A, D, E, K) groups; small quantities (from 0.00002 percent to 0.005 percent of a diet) are necessary for normal health and growth in higher forms of animal life. A number of compounds (*e.g.,* choline, carnitine) once grouped with vitamins no longer are considered vitamins (see below *Vitamin-like substances*). If a vitamin is absent from the diet or is not properly absorbed by an organism, a specific deficiency disease may develop.

The term vitamin originated from "vitamine," a word first used in 1911 to designate a group of compounds considered vital for life; each was thought to have a nitrogen-containing component known as an amine. The final *e* of vitamine was dropped when it was discovered that not all of the vitamins contain nitrogen, and, therefore, not all are amines. The term accessory food factor sometimes is used instead of vitamin to refer to these substances.

Vitamins as essential nutrients

Since they generally cannot be synthesized by an animal (or, if synthesized, the amounts are insufficient to meet body needs) and must be obtained from the diet or from some synthetic source, vitamins are called essential nutrients. The requirements for some of the B vitamins may be met in part by bacterial synthesis in the intestines of some mammals. The amino acid tryptophan can be converted to nicotinic acid (or niacin, a water-soluble vitamin) and thus can serve as a source for part of the nicotinic acid required by an animal. Vitamin C (also a water-soluble vitamin) can be synthesized by some organisms in sufficient amounts so that the dietary requirement is eliminated; vitamin C usually is considered a vitamin, however, because it must be included in the diet of man. Vitamins are distinct from many other compounds, which, although indispensable for proper animal functions, can be synthesized in adequate quantities.

General characteristics

Definitions of provitamin, antivitamin

A provitamin is similar in structure to a specific vitamin and can be converted to it by a few metabolic reactions *(e.g.,* beta-carotene can be converted to vitamin A; 7-dehydrocholesterol can be converted to vitamin D_3). The amino acid tryptophan is called a precursor of the vitamin nicotinic acid because the conversion pathway is less direct than that of a provitamin. Antivitamins are compounds that prevent the normal function of certain vitamins. Antivitamins may act by binding a vitamin *(e.g.,* the antivitamin avidin binds the vitamin biotin), by destroying a vitamin (*e.g.,* the antivitamin thiaminase destroys thiamine), or by inhibiting the coenzyme function of a vitamin. A coenzyme is, as noted previously, a heat stable compound that combines with the protein component of an enzyme (*i.e.,* an organic catalyst) to form an active enzyme. Antivitamins that act by inhibiting the coenzyme function of a vitamin usually are known as antagonists or antimetabolites.

Letters originally were assigned to the vitamins, which were differentiated according to physiological function. As chemical structures of the vitamins became known, however, they were also given names and now are commonly known either by letter or by name. Currently accepted names for the vitamins, established by international agreement, are given in Tables 14 and 15.

BIOLOGICAL SIGNIFICANCE

Regulatory role. The vitamins regulate reactions that occur in metabolism, in contrast to other dietary components known as macronutrients (*e.g.,* fats, carbohydrates, proteins), which are the compounds utilized in the reactions regulated by the vitamins. Absence of a vitamin blocks one or more specific metabolic reactions in a cell and eventually may disrupt the metabolic balance within a cell and in the entire plant or animal as well.

With the exception of vitamin C (ascorbic acid), all of the water-soluble vitamins have a catalytic function; *i.e.,* they act as coenzymes of enzymes that function in energy transfer or in the metabolism of fats, carbohydrates, and proteins. The metabolic importance of the water-soluble vitamins is reflected by their presence in most plant and animal tissues involved in metabolism.

Some of the fat-soluble vitamins form part of the structure of biological membranes or assist in maintaining the integrity (and therefore, indirectly, the function) of membranes. Some fat-soluble vitamins also may function at the genetic level by controlling the synthesis of certain enzymes. The roles of the fat-soluble vitamins have not yet been established with certainty. Unlike the water-soluble ones, fat-soluble vitamins are necessary for specific functions in highly differentiated and specialized tissues; therefore, their distribution in nature tends to be more selective than that of the water-soluble vitamins.

Sources. Vitamins, which are found in all living organisms either because they are synthesized in the organism or are acquired from the environment, are not distributed equally throughout nature. Some are absent from certain tissues or species; for example, beta-carotene (provitamin A) is synthesized in plant tissues but not in animal tissues. On the other hand, vitamins A and D_3 occur only in animal tissues. Both plants and animals are important natural vitamin sources for man. Since vitamins are not distributed equally in foodstuffs, the more restricted the diet of an animal (man, for example), the more likely it is that he will lack adequate amounts of one or more vitamins. Food sources of vitamin D are limited, but it can be synthesized in the skin through ultraviolet radiation (from the Sun) on the provitamin, 7-dehydrocholesterol; therefore, with adequate exposure to sunlight, the dietary intake of vitamin D is of little significance.

All vitamins either can be synthesized or produced commercially from food sources and are available for human consumption in pharmaceutical preparations. Commercial processing of food (*e.g.,* milling of grains) frequently destroys (or removes) considerable amounts of vitamins. In most such instances, however, the vitamins are replaced by chemical methods. Some foods are "fortified" with vitamins not normally present in them (*e.g.,* vitamin D is added to milk). Loss of vitamins may also occur when food is cooked (*e.g.,* heat destroys vitamin A, water-soluble vitamins may be extracted from food to water and lost). Certain vitamins (*e.g.,* B vitamins, vitamin K) can be synthesized by microorganisms normally present in the intestines of some animals; however, the microorganisms usually do not supply the host animal with an adequate quantity of a vitamin.

Requirements of living things. The vitamins required by most organisms are fairly well established. Vitamin requirements vary according to species and the amount of a vitamin required by a specific organism is difficult to determine because of the numerous factors involved (*e.g.,* genetic variation; presence of specific disease states; therapeutic use of certain drugs that also may act as vitamin antagonists; infestation with parasites; relative proportions of other dietary constituents, food additives, or contaminants; environmental stresses; and stimulation of growth rate, as in animal husbandry). There is not uniform agreement concerning the vitamin requirements for man. Differences in opinion arise mainly from different ways by which requirements can be determined and from the scanty data available for the requirements for some of the vitamins. Recommended daily vitamin allowances, however, are sufficiently high to account for individual variation and normal environmental stresses. For detailed information on vitamin requirements in man, see NUTRITION.

Factors influencing vitamin requirements

A number of interrelationships exist among vitamins and between vitamins and other dietary constituents. The interactions may be synergistic (*i.e.,* cooperative) or antagonistic, reflecting, for example, overlapping metabolic roles (of the B vitamins in particular), protective roles (*e.g.,* vitamins A and E), or structural dependency (*e.g.,* cobalt in the vitamin B_{12} molecule).

Results of deficiencies. An inadequate intake of a specific vitamin results in a characteristic deficiency disease (or hypovitaminosis); the severity of the disease depends upon the degree of vitamin deprivation (see Figure 13).

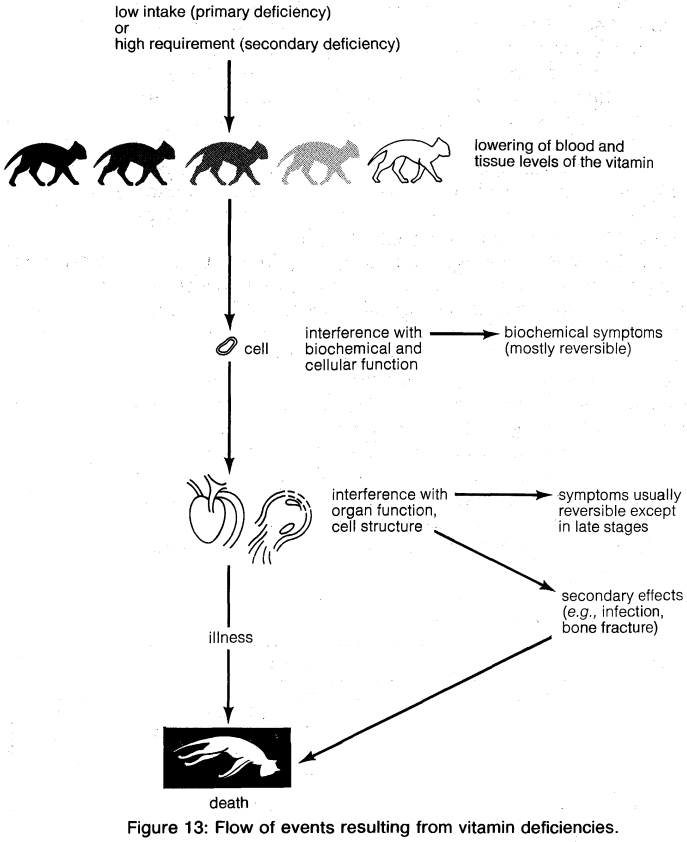


Figure 13: Flow of events resulting from vitamin deficiencies.

Symptoms of deficiencies

Symptoms may be specific (*e.g.,* functional night blindness of vitamin A deficiency) or nonspecific (*e.g.,* loss of appetite, failure to grow). All symptoms for a specific deficiency disease may not appear; in addition, the nature of the symptoms may vary with the species. Some effects of vitamin deficiencies cannot be reversed by adding the vitamin to the diet, especially if damage to nonregenerative tissue (*e.g.,* cornea of the eye, nerve tissue, calcified bone) has occurred.

Primary and secondary deficiencies

A vitamin deficiency may be "primary" (or dietary), in which case the dietary intake is lower than the normal requirement of the vitamin. A "secondary" (or conditioned) deficiency may occur (even though the dietary intake is adequate) if a pre-existing disease or state of stress is present (*e.g.,* malabsorption of food from the intestine, chronic alcoholism, repeated pregnancies and lactation). For more details on vitamin deficiencies in man, see NUTRITION: *Nutritional diseases and disorders.*

Evolution of vitamin-dependent organisms. Evolution of metabolic processes in primitive forms of life required the development of enzyme systems to catalyze the complex sequences of chemical reactions involved in metabolism. In the beginning, the environment presumably could supply all the necessary compounds (including the vitamin coenzymes); eventually, these compounds were synthesized within an organism. As higher forms of life evolved, however, the ability to synthesize certain of these vitamin coenzymes was gradually lost.

Since higher plants show no requirements for vitamins or other growth factors, it is assumed that they retain the ability to synthesize them. Among insects, however, nicotinic acid; vitamins B_1, B_2, B_6, and C; and pantothenic acid are required by a few groups. All vertebrates, including man, require dietary sources of vitamins A, D, B_1, B_2, B_6, and pantothenic acid; some vertebrates, particularly the more highly evolved ones, have additional requirements for other vitamins. The inability of man and a few other primates to synthesize vitamin C, however, may be a genetic disease (*i.e.,* an inborn error of metabolism), rather than the result of a true evolutionary process. For more details on the evolution of primitive forms, see LIFE.

METHODS USED IN VITAMIN RESEARCH

Determination of vitamin requirements. If a specific factor in food is suspected of being essential for the growth of an organism (either by growth failure or some other clinical symptoms that are alleviated by adding a specific food to the diet) a systematic series of procedures (described below) is used to characterize the factor.

The active factor (or principle) is isolated from specific foods and purified; then its chemical structure is determined, and it is synthesized in the laboratory. Structural determination and synthesis, which may be achieved only after long and intensive research, must be completed before the function and the quantitative requirements of the factor can be established accurately. Established organic and analytical chemical procedures are used to determine the structure of the factor and to synthesize it.

Biological studies may be performed to determine functions, effects of deprivation, and quantitative requirements of the factor in various organisms. The development in an organism of a deficiency either by dietary deprivation of the vitamin or by administration of a specific antagonist or antivitamin often is the method used. The obvious effects (*e.g.,* night blindness, anemia, dermatitis) of the deficiency are noted. Less obvious effects may be discovered after microscopic examination of tissue and bone structures. Changes in concentrations of metabolites or in enzymatic activity in tissues, blood, or excretory products are examined by numerous biochemical techniques (including the use of radioisotopes). The response of an animal to a specific vitamin of which it has been deprived usually confirms the deficiency symptoms for that vitamin. Effects of deprivation of a vitamin sometimes indicate its general physiological function, as well as its function at the cellular level. Biochemical function often is studied by observing the response of tissue enzymes (removed from a deficient host animal) after a purified vitamin preparation is added. The functions of most of the known vitamins have been reasonably well defined; however, the mechanism of action has not yet been established for some of them (*e.g.,* vitamins A, E, C.)

The procedure for determining the amount of a vitamin required by an organism is less difficult for microorganisms than for higher forms; in microorganisms, the aim is to establish the smallest amount of a vitamin that produces maximal growth (*i.e.,* maximal rate of multiplication) of the organisms when it is added to the culture medium. Among vertebrates, particularly man, a number of procedures are used together to provide estimates of the vitamin requirement. These procedures include determinations of: the amount of a vitamin required to cure a deficiency that has been developed under controlled, standard conditions; the smallest amount required to prevent the appearance of clinical or biochemical symptoms of the deficiency; the amount required to saturate body tissues (*i.e.,* to cause "spillover" of the vitamin in the urine; valid only with the water-soluble vitamins); the amount necessary to produce maximum blood levels of the vitamin plus some tissue storage (applicable only to the fat-soluble vitamins, particularly vitamin A); the amount required to produce maximum activity of an enzyme system if the vitamin has a coenzyme function; the actual rate of utilization, and hence the requirement, in healthy individuals (as indicated by measuring the excreted breakdown products of radioisotope-labelled vitamins).

Comparisons of intake in human populations

The above procedures are practical only with small groups of animals or human subjects and thus are not entirely representative of larger populations of a particular species. A less precise (but more representative) method

used among human populations involves comparing levels of dietary intake of a vitamin in a population that shows no deficiency symptoms with levels of intake of the vitamin in a population that reveals clinical or biochemical symptoms. The data for dietary intakes and incidence of deficiency symptoms are obtained by surveys of representative segments of a population.

Determination of vitamin sources. A quantitative analysis of the vitamin content of foodstuffs is important in order to identify dietary sources of specific vitamins (and other nutrients as well). Three methods commonly used to determine vitamin content are described below.

Physicochemical methods. The amount of vitamin in a foodstuff can be established by studying the physical or chemical characteristics of the vitamin; *e.g.,* a chemically reactive group on the vitamin molecule, fluorescence (for thiamin, riboflavin, vitamin A), absorption of light at a wavelength characteristic of the vitamin (for most vitamins), radioisotope dilution techniques (for vitamin B_{12}). These methods are accurate and can detect very small amounts of the vitamin. Biologically inactive derivatives of several vitamins have been found, however, and may interfere with such determinations; in addition, these procedures also may not distinguish between bound (*i.e.,* unavailable) and available forms of a vitamin in a food.

Microbiological assay. Microbiological assay is applicable only to the B vitamins. The rate of growth of a species of microorganism that requires a vitamin is measured in growth media that contain various known quantities of a foodstuff preparation containing unknown amounts of the vitamin. The response (measured as rate of growth) to the unknown amounts of vitamin is compared with that obtained from a known quantity of the pure vitamin. Depending on the way in which the food sample was prepared, the procedure may indicate the availability of the vitamin in the food sample to the microorganism.

Animal assay. All of the vitamins, with the exception of vitamin B_{12}, can be estimated by the animal-assay technique. One advantage of this method is that animals respond only to the biologically active forms of the vitamins. On the other hand, many other interfering and complicating factors may arise; therefore, experiments must be rigidly standardized and controlled. Simultaneous estimates usually are made using a pure standard vitamin preparation (as a reference) and the unknown food (whose vitamin content is being sought); each test is repeated using two or more different amounts of both standard and unknown in the assays listed below (*i.e.,* growth, reaction time, graded response, all-or-none).

Types of animal assays

In a growth assay, the rat, chick, dog (used specifically for niacin), and guinea pig (used specifically for vitamin C) usually are used. One criterion used in a vitamin assay is increase in body weight in response to different amounts of a specific vitamin in the diet. There are two types of growth assay. In a prophylactic growth assay, the increase in weight of young animals given different amounts of the vitamin is measured. In a curative growth assay, weight increase is measured in animals first deprived of a vitamin and then given various quantities of it. The curative growth assay tends to provide more consistent results than the prophylactic technique.

In a reaction time assay, an animal is first deprived of a vitamin until a specific deficiency symptom appears; then the animal is given a known amount of a food extract containing the vitamin, and the deficiency symptom disappears within a day or two. The time required for the reappearance of the specific symptoms when the animal again is deprived of the vitamin provides a measure of the amount of vitamin given originally. The graded response assay, which may be prophylactic or curative, depends on a characteristic response that varies in degree with the vitamin dosage. An example of the use of this technique is an assay for vitamin D in which the measured ash content of a leg bone of a rat or chick is used to reflect the amount of bone calcification that occurred as a result of administration of a specific amount of vitamin D. In an all-or-none assay, the degree of response cannot be measured; an arbitrary level is selected to separate positive responses from negative ones. The percent of positively reacting animals provides a measure of response; *i.e.,* vitamin E can be measured by obtaining the percent of fertility in successfully mated female rats.

Biochemistry of the water-soluble vitamins

BASIC PROPERTIES

Although the vitamins included in this classification are all water-soluble, the degree to which they dissolve in water is variable. This property influences the route of absorption, their excretion, and their degree of tissue storage and distinguishes them from fat-soluble vitamins, which are handled and stored differently by the body. The water-soluble vitamins (excluding vitamin C) popularly are termed the B-vitamins; these include B_1 (thiamine), B_2 (riboflavin), B_6 (pyridoxine), niacin (nicotinic acid), B_{12} (cyanocobalamin), folic acid, pantothenic acid, and biotin. These relatively simple molecules contain the elements carbon, hydrogen, and oxygen; some also contain nitrogen, sulfur, or cobalt.

The water-soluble vitamins, inactive in their so-called free states, must be activated to their coenzyme forms (see Table 14); addition of phosphate groups occurs in the activation of vitamins B_1, B_2, and B_6; a shift in structure activates biotin, and formation of a complex between the free vitamin and parts of other molecules is involved in the activation of nicotinic acid, pantothenic acid, folic acid, and vitamin B_{12}. After an active coenzyme is formed, it must combine with the proper protein component (called an apoenzyme) before enzyme-catalyzed reactions can occur. Vitamin antagonists (or antimetabolites) function by: (1) preventing coenzyme formation, (2) competing with the vitamin coenzyme for a site on the protein portion of the enzyme, (3) competing with a compound (whose transformation is catalyzed by the enzyme) for the active centre on the enzyme, or (4) inactivating an already formed coenzyme.

Activation

Table 14: The Water-Soluble Vitamins

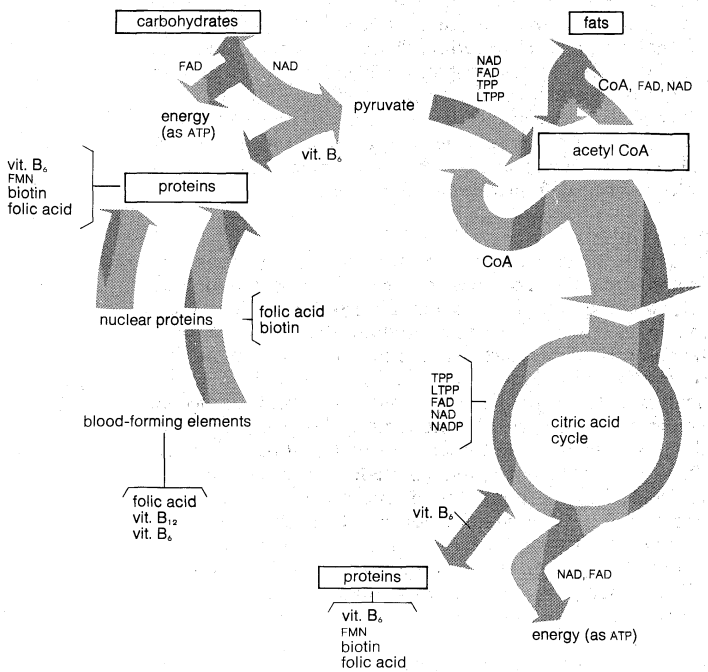
free form*	coenzyme form
Thiamine (vitamin B_1)	thiamine pyrophosphate (TPP) or cocarboxylase
	lipothiamide pyrophosphate (LTPP)
Riboflavin (vitamin B_2)	flavin mononucleotide (FMN)
	flavin–adenine dinucleotide (FAD)
Pyridoxine or pyridoxol† (vitamin B_6)	pyridoxal phosphate or codecarboxylase
also: Pyridoxal	
Pyridoxamine	
Nicotinic acid† (niacin, vitamin PP)	nicotinamide–adenine dinucleotide (NAD)
also: Nicotinamide (niacinamide)	nicotinamide–adenine dinucleotide phosphate (NADP)
Cyanocobalamin† (vitamin B_{12})	several cobamide coenzymes
also: Aquocobalamin (vitamin B_{12}^a)	
Hydroxocobalamin (vitamin B_{12}^b)	
Nitrotocobalamin (vitamin B_{12}^c)	
Pteroylglutamic acid (folic acid) (folic acid is now a general term for the family of pteroic acids and their salts)	tetrahydropteroylglutamic acid
Pantothenic acid	coenzyme A
Biotin	1′–N–carboxybiotin
Ascorbic acid (vitamin C)	none yet established

*The names of the free forms of the first six vitamins are those approved thus far by the IUPAC-IUB Commission of Nomenclature of Biological Compounds. Terms given in parentheses after the approved forms are older or "popular" terms. †Considered to be the basic or standard vitamin form. Those following also have vitamin activity.

FUNCTIONS

The B-vitamin coenzymes function in enzyme systems that transfer certain groups between molecules (see Figure 14); as a result, specific proteins, fats, and carbohydrates are formed and may be utilized to produce body tissues (including blood cells) or to store or release energy. The pantothenic acid coenzyme functions in the Krebs tricarboxylic acid (or citric acid) cycle, which interconnects carbohydrate, fat, and protein metabolism; this coenzyme (coenzyme A) acts at the hub of these reactions and thus is an important molecule in controlling the interconversion of fats, proteins, and carbohydrates and their conversion into metabolic energy. Thiamine (B_1) and pyridoxine (B_6) coenzymes control the conversion of carbohydrates and

proteins respectively into metabolic energy during the citric acid cycle. Nicotinic acid (niacin) and riboflavin (B_2) coenzymes facilitate the transfer of hydrogen ions or electrons (negatively charged particles), which occurs during the reactions of the citric acid cycle. All of these coenzymes also function in transfer reactions that are involved in the synthesis of structural compounds; these reactions are not part of the citric acid cycle (see Figure 14).



CoA—coenzyme A
FAD—flavine adenine dinucleotide
FMN—flavine mononucleotide
LTPP—lipothiamide pyrophosphate
NAD—nicotinamide-adenine dinucleotide
NADP—nicotinamide-adenine dinucleotide phosphate
TPP—thiamine pyrophosphate

Figure 14: Functions of B vitamin coenzymes in metabolism.

Although vitamin C (ascorbic acid) participates in some enzyme-catalyzed reactions, it has not yet been established that the vitamin is a coenzyme. Its function probably is related to its properties as a strong reducing agent (*i.e.*, it readily gives electrons to other molecules).

METABOLISM

The water-soluble vitamins are absorbed in the animal intestine, pass directly to the blood, and are carried to the tissues in which they will be utilized. Vitamin B_{12} requires a substance known as "intrinsic factor for absorption.

Utilization

Some of the B vitamins can occur in forms that cannot be used by an animal. Most of the nicotinic acid in some cereal grains (wheat, corn, rice, barley, bran), for example, is bound to another substance, and the complex is called niacytin; this compound is not absorbed in the animal intestine. Biotin can be bound by avidin, which is found in raw egg white; this complex cannot be absorbed or broken down by digestive-tract enzymes, and thus the biotin cannot be utilized. In animal products (*e.g.*, meat), biotin, vitamin B_6, and folic acid are bound to other molecules to form complexes or conjugated molecules; although none is active in the complex form, the three vitamins normally are released from the bound forms by the enzymes of the intestinal tract (for biotin and vitamin B_6) or in the tissues (for folic acid) and thus can be utilized. The B vitamins are distributed in most metabolizing tissues of plants and animals.

Water-soluble vitamins usually are excreted in the urine of man. Thiamine (B_1), riboflavin (B_2), pyridoxine (B_6), ascorbic acid (C), pantothenic acid, and biotin appear in urine as free vitamins (rather than as coenzymes); however, little free nicotinic acid is excreted in the urine. Products (also called metabolites) that are formed during the metabolism of thiamine, nicotinic acid, and pyridoxine also appear in the urine. Urinary metabolites of biotin, riboflavin, and pantothenic acid also are formed, although they have not yet been characterized. Excretion of these vitamins (or their metabolites) is low when intake is just sufficient for proper body function. If intake begins to exceed minimal requirements, excess vitamins are stored in the tissues. Tissue storage capacity is limited, however, and, as the tissues become saturated, the rate of excretion increases sharply. This knowledge is valuable in establishing the nutritional status of an individual with respect to the vitamins. Unlike the other water-soluble vitamins, however, vitamin B_{12} is excreted solely in the feces. Some folic acid and biotin also are normally excreted in this way. Although fecal excretion of water-soluble vitamins (other than vitamin B_{12}, folic acid, and biotin) occurs, their source probably is the intestinal bacteria (which synthesize the vitamins), rather than vitamins that have been eaten and utilized by an animal.

Toxicity

The water-soluble vitamins generally are not considered toxic if taken in excessive amounts. There is, however, one exception in man; large amounts (50–100 milligrams) of nicotinic acid produce dilation of blood vessels; in larger amounts, the effects are more serious and may result in impaired liver function. Thiamine given to animals in amounts 100 times the requirement (*i.e.*, about 100 milligrams) can cause death from respiratory failure. Therapeutic doses (100–500 milligrams) of thiamine have no known toxic effects in man (except rare instances of anaphylactic shock in sensitive individuals). There is no known toxicity for any other B vitamins.

Biochemistry of fat-soluble vitamin groups

The four fat (lipid) soluble vitamin groups are A, D, E, and K; they are related structurally in that all have as a basic structural unit of the molecule a five-carbon isoprene segment, which is

$$-CH=CH-\overset{\overset{CH_3}{|}}{C}=CH-$$

Each of the lipid-soluble vitamin groups contains several related compounds that have biological activity. The active forms and the accepted nomenclature of individual vitamins in each vitamin group are given in Table 15. The potency of the active forms in each vitamin group varies, and not all of the active forms now known are available from dietary sources; *i.e.*, some are produced synthetically.

Table 15: Lipid-Soluble Vitamins: Active Forms of Biological Importance

vitamin*	provitamin
Vitamin A group	
Retinol (vitamin A alcohol)	Alpha- and beta-carotene, cryptoxanthin
Retinal (vitamin A aldehyde, retinene)	
Retinyl ester (vitamin A ester)	
Retinoic acid (vitamin A acid)	
Vitamin D group	
Ergocalciferol (vitamin D_2, calciferol)	Ergosterol
Cholecalciferol (vitamin D_3)	7-dehydrocholesterol
Vitamin E group	
Alpha-tocopherol (vitamin E)	
Beta-tocopherol	
Gamma-tocopherol	
Delta-tocopherol	
Beta-tocotrienol (epsilon-tocopherol)	
Tocochromanol-3 (zeta-tocopherol)	
Plastochromanol-3 (eta-tocopherol)	
Vitamin K group	
Phylloquinone or Phylloquinone K (vitamin K_1 [20])†	
Menaquinone-4‡ or Menaquinone K_4 (vitamin K_2 [20])	
Menaquinone-6 or Menaquinone K_6 (vitamin K_2 [30])	
Menaquinone (vitamin K_3, menadione)	
also: vitamins K_4, K_5, K_6, K_7	

*Names of specific vitamins given are approved by the IUPAC-IUB Commission of Nomenclature of Biological Compounds. Terms given in parentheses after the approved forms are older or "popular" terms. Since there is disagreement as to which name is more appropriate; alternate names are included. †Number in brackets indicates total number of carbon atoms in side chain. ‡Number refers to number of isoprenoid units in side chain.

The characteristics of each lipid-soluble vitamin group are discussed below.

PRINCIPAL CHARACTERISTICS

Active forms of biological importance. *Vitamin A group.* Ten carotenes, coloured molecules synthesized only in plants, show vitamin A activity (*i.e.*, function like vitamin A); however, only the alpha- and beta-carotenes and cryptoxanthin are important to man, and beta-carotene is the most active. Retinol (vitamin A alcohol) is considered the primary active form of the vitamin, although retinal, or vitamin A aldehyde, is the form involved in the visual process in the retina of the eye. A metabolite of retinol with high biological activity may be an even more direct active form than retinol. The ester form of retinol is the storage form of vitamin A; presumably, it must be converted to retinol before it is utilized. Retinoic acid is a short-lived product of retinol; only retinoic acid of the vitamin A group is not supplied by the diet.

Retinol

Vitamin D group. Although about 10 compounds have vitamin D activity, the two most important ones are vitamins D_2 (or ergocalciferol) and D_3 (or cholecalciferol). Vitamin D_3 represents the dietary source, while vitamin D_2 occurs in yeasts and fungi. Both can be formed from their respective provitamins by ultraviolet irradiation; in man and other animals the provitamin (7-dehydrocholesterol), which is found in skin, can be converted by sunlight to vitamin D_3 and thus is an important source of the vitamin. Both vitamins D_2 and D_3 can be utilized by the rat and man; however, chicks cannot use vitamin D_2 effectively. The form of the vitamin probably active in man is 1, 25-di-hydroxycholecalciferol.

Vitamin E group. The tocopherols are a closely related group of biologically active compounds that vary only in number and position of methyl (*i.e.*, $-CH_3$) groups in the molecule; however, these structural differences influence the biological activity of the various molecules. The active tocopherols are named in order of their potency; *i.e.*, alpha-tocopherol is the most active. Some metabolites of alpha-tocopherol (*e.g.*, alpha-tocopherolquinone, alphatocopheronolactone) have activity in some mammals (*e.g.*, rats, rabbits); however, these metabolites do not support all the functions attributed to vitamin E.

Vitamin K group. Vitamin K_1 (20) is synthesized by plants; the members of the vitamin K_2 (30) series are of microbial origin. Vitamin K_2 (20) is the important form in mammalian tissue; all other forms are converted to K_2 (20) from vitamin K_3. Since vitamin K_3 does not accumulate in tissue, it does not furnish any dietary vitamin K. Vitamins K_4, K_5, K_6, and K_7 have been synthesized and have vitamin K activity, but they are not found in nature.

International Unit

Units of activity. For many years, the unit of measurement for the activity of vitamins A, D, and E has been the International Unit (IU); this is analogous with the United States Pharmocopeia unit (USP), which is based on a measured biological activity. The weight equivalents of these vitamins are found in Table 16.

Table 16: Weight Equivalents of International Units of Vitamins A, D, and E

vitamin	IU	weight equivalents
Vitamin A	1 IU of all-trans retinol	0.300 micrograms
	1 IU of all-trans retinyl acetate	0.344 micrograms
	1 IU of beta-carotene	0.600 micrograms
Vitamin D	1 IU of pure crystalline vitamin D_3	0.025 micrograms
Vitamin E	1 IU of synthetic *dl*-alpha-tocopherol acetate	1 milligram
	1.1 IU of *dl*-alpha-tocopherol	1 milligram
	1.36 IU of the natural form *d*-alpha-tocopherol acetate	1 milligram

Vitamin K activity is expressed only in micrograms. The activity of fat-soluble vitamins now is given on a weight basis (as is that of water-soluble vitamins) because pure vitamin preparations are available for use as references.

FUNCTIONS

The vitamin A group has at least one known function. In the retina of the eye, retinal is combined with a protein called opsin; the complex molecules formed as a result of this combination and known as rhodopsin (or visual purple) are involved in dark vision. Other metabolic functions of vitamin A have not yet been defined with certainty. The vitamin D group is required for growth (especially bone growth or calcification). The vitamin E group also is necessary for normal animal growth; without these tocopherols, animals are not fertile and develop abnormalities of the central nervous system, muscles, and organs (especially the liver). The vitamin K group is required for normal metabolism, including the conversion of food into cellular energy in certain biological membranes; vitamin K also is necessary for the proper clotting of blood.

Knowledge concerning the functions of the fat-soluble vitamins is not yet so extensive as that for the water-soluble ones. It appears, however, that most vitamins may be concerned with proper enzyme function.

METABOLISM

The fat-soluble vitamins are transported primarily by lymph from the intestines to the circulating blood. Bile salts are required for efficient absorption of fat-soluble metabolites in the intestine; anything that interferes with fat absorption, therefore, also inhibits absorption of the fat-soluble vitamins. Since a fatty acid (preferentially palmitic acid) is added to the vitamin A alcohol (retinol) molecule before it is transported by the lymph, this ester form predominates in the bloodstream during digestion. Vitamins D, E, and K do not require the addition of a fatty acid molecule for absorption. Small amounts of vitamin A alcohol (and possibly vitamin K) may be absorbed directly into the bloodstream; however, both vitamins A and D are bound to a protein during transport in the bloodstream.

Larger quantities of the fat-soluble vitamins than of water-soluble ones can be stored in the body. Vitamins A, D, and K are stored chiefly in the liver, with smaller amounts stored in other soft body tissues; however, most of the stored vitamin E is found in body fat, although large amounts also occur in the uterus of females and testis of males. The various forms of vitamin E are stored in tissues in different amounts; alpha-tocopherol is stored in higher concentrations than are the other forms. More vitamin A is stored (in ester form) than any other fat-soluble vitamin. Before its release in the bloodstream, the ester form of vitamin A is converted to retinol. In animals in the postabsorptive stage, retinol is the predominant form in the blood.

Excessive intakes of both vitamins A and D may produce toxicity (or hypervitaminosis A or D). Among natural foods, only seal liver and polar bear liver are known to contain sufficiently high concentrations of vitamin A to produce toxic effects. Toxicity of both vitamin A and vitamin D can easily occur, however, if pharmaceutical vitamin preparations are used in excess.

Toxic levels of vitamin A exceed the normal requirement by 100 times; *i.e.*, about 500,000 IU each day for a period of several months. Toxicity in infants may occur with doses of 18,000 to 60,000 IU per day. Excessive doses of the natural vitamins K_1 and K_2 have no obvious effects except that resistance may develop to therapy with dicumarol compounds (anticoagulants); however, vitamin K_3 is toxic to newborn infants if given in large doses. Vitamin E, even if given in large excess of the normal requirement, has no apparent obvious adverse effects.

Vitamin groups E and K belong to a class of organic compounds called quinones. These substances are changed to sugar-like substances known as alpha-lactones, which are excreted in the urine (probably as carbohydrates called glucuronides). Some vitamin K_1 also is excreted in the bile and thus appears in the feces. Vitamin A is broken down and excreted as a glucuronide and a number of compounds in bile (and, therefore, feces) and urine. Vitamin D and its breakdown products are excreted only in the feces.

Vitamin-like substances

There are a number of organic compounds that, although related to the vitamins in activity, cannot be defined as

true vitamins; normally they can be synthesized by man in adequate amounts and therefore are not required in the diet. These substances usually are classified with the B vitamins, however, because of similarities in biological function or distribution in foods.

Choline. Choline appears to be an essential nutrient for a number of animals (including some birds) and microorganisms that cannot synthesize adequate quantities to satisfy their requirements.

Importance of choline

Choline is a constituent of an important class of lipids called phospholipids, which form structural elements of cell membranes; choline is a component of the acetylcholine molecule, which is important in nerve function. Choline also serves as a source of methyl groups ($-CH_3$ groups) that are required in various metabolic processes. The effects of a dietary deficiency of choline itself can be alleviated by other dietary compounds that can be changed into choline. Choline also functions in the transport of fats from the liver; for this reason, choline may be called a lipotropic factor. A deficiency of choline in the rat results in an accumulation of fat in the liver. Choline deficiency symptoms vary among species; it is not yet known if choline is an essential nutrient for man since a dietary deficiency has not been demonstrated.

Myoinositol. The biological significance of myoinositol (also previously known as meso-inositol) has not yet been established with certainty. It is present in large amounts (principally as a constituent of phospholipids) in man. Inositol is a carbohydrate that closely resembles glucose in structure; inositol can be converted to phytic acid, which is found in grains and forms an insoluble (and thus unabsorbable) calcium salt in the intestines of mammals. Inositol has not yet been established as an essential nutrient for man; however, it is a required factor for the growth of some yeasts and fungi.

Para-aminobenzoic acid. Para-aminobenzoic acid (abbreviated to PABA) is required for the growth of several types of microorganisms; however, a dietary requirement by vertebrates has not yet been shown. The sulfa drug (sulfanilamide) can kill bacteria because it competes with PABA for a position in a coenzyme that is necessary for bacterial reproduction. Although a structural unit of folic acid, PABA is not considered a vitamin.

Carnitine. Carnitine (originally called vitamin B_T) is essential for the growth of mealworms. The role of carnitine in all organisms is associated with the transfer of fatty substances (*e.g.,* fatty acids) from the bloodstream to active sites of fatty acid oxidation within muscle cells. Carnitine, therefore, regulates the rate of oxidation of these acids; this function may afford means by which a cell can rapidly shift its metabolic patterns (*e.g.,* from fat synthesis to fat breakdown). Synthesis of carnitine occurs in insects and in higher animals; therefore, it probably should not be considered a true vitamin.

Role of carnitine

Lipoic acid. Lipoic acid has a coenzyme function similar to that of thiamine. Although it is apparently an essential nutrient for some microorganisms, no deficiency in mammals has yet been observed; therefore, lipoic acid is not considered a true vitamin.

Bioflavinoids. The bioflavinoids once were thought to prevent scurvy and were designated as vitamin Pc, but additional evidence refuted this claim. (M.J.B./Ed.)

HORMONES

Hormones are organic substances that are secreted by plants and animals and that function in the regulation of physiological activities and in maintaining homeostasis. They carry out their functions by evoking responses from specific organs or tissues that are adapted to react to minute quantities of them. The classical view of hormones is that they are transmitted to their targets in the bloodstream after discharge from the glands that secrete them. This mode of discharge (directly into the bloodstream) is called endocrine secretion. The meaning of the term hormone has been extended beyond the original definition of a blood-borne secretion, however, to include similar regulatory substances that are distributed by diffusion across cell membranes instead of by a blood system.

General features

RELATIONSHIPS BETWEEN ENDOCRINE AND NEURAL REGULATION

Hormonal regulation is closely related to that exerted by the nervous system, and the two processes have generally been distinguished by the rate at which each causes effects, the duration of these effects, and their extent; *i.e.*, the effects of endocrine regulation may be slow to develop but prolonged in influence and widely distributed through the body, whereas nervous regulation is typically concerned with quick responses that are of brief duration and localized in their effects. Advances in knowledge, however, have modified these distinctions.

Neurohormones

Nerve cells are secretory, for responses to the nerve impulses that they propagate depend upon the production of chemical transmitter substances, or neurohumors, such as acetylcholine and noradrenaline (norepinephrine), which are liberated at nerve endings in minute amounts and have only a momentary action. It now has been established, however, that certain specialized nerve cells, called neurosecretory cells, can translate neural signals into chemical stimuli by producing secretions called neurohormones. These secretions, which are often polypeptides (compounds similar to proteins but composed of fewer amino acids), pass along nerve-cell extensions, or axons, and are typically released into the bloodstream at special regions called neurohemal organs, where the axon endings

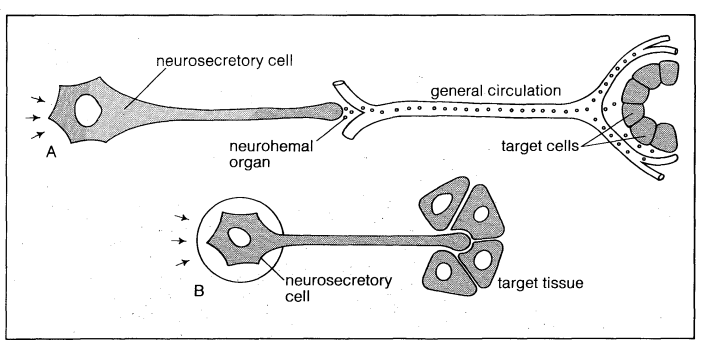


Figure 15: *The release of neurohormones from neurosecretory nerve cells.*
(A) Nerve signals (arrows) are translated into neurohormones, which enter the circulation at the neurohemal organ, reaching target cells via the bloodstream. (B) Release of neurohormone close to target cells without intervention of bloodstream.

are in close contact with blood capillaries (Figure 15A). Once released in this way, neurohormones function in principle similar to hormones that are transmitted in the bloodstream and are synthesized in the endocrine glands.

The distinctions between neural and endocrine regulation, no longer as clear-cut as they once seemed to be, are further weakened by the fact that neurosecretory nerve endings are sometimes so close to their target cells that vascular transmission is not necessary (Figure 15B). There is good evidence that hormonal regulation occurs by diffusion in plants and (although here the evidence is largely indirect) in lower animals (*e.g.,* coelenterates), which lack a vascular system.

THE EVOLUTION OF HORMONES

Hormones have a long evolutionary history, knowledge of which is important if their properties and functions are to be understood. Many important features of the vertebrate endocrine system, for example, are present in the lampreys and hagfishes, modern representatives of the primitively jawless vertebrates (Agnatha), and these features were presumably present in fossil ancestors that lived more than 500,000,000 years ago. The evolution of the

endocrine system in the more advanced vertebrates with jaws (Gnathostomata) has involved both the appearance of new hormones and the further evolution of some of those already present in agnathans; in addition, extensive specialization of target organs has occurred to permit new patterns of response.

ole in
hemical
egulation

The factors involved in the first appearance of the various hormones is largely a matter for conjecture, although hormones clearly are only one mechanism for chemical regulation, diverse forms of which are found in living things at all stages of development. Other mechanisms for chemical regulation include chemical substances (so-called organizer substances) that regulate early embryonic development and the pheromones that are released by social insects as sex attractants and regulators of the social organization. Perhaps, in some instances, chemical regulators including hormones appeared first as metabolic by-products. A few such substances are known in physiological regulation: carbon dioxide, for example, is involved in the regulation of the respiratory activity of which it is a product, in insects as well as in vertebrates. Substances such as carbon dioxide are called parahormones to distinguish them from true hormones, which are specialized secretions.

The hormones of vertebrates

HORMONES OF THE PITUITARY GLAND

The pituitary gland, or hypophysis (Figure 16), which dominates the vertebrate endocrine system, is formed of two distinct components. One is the neurohypophysis,

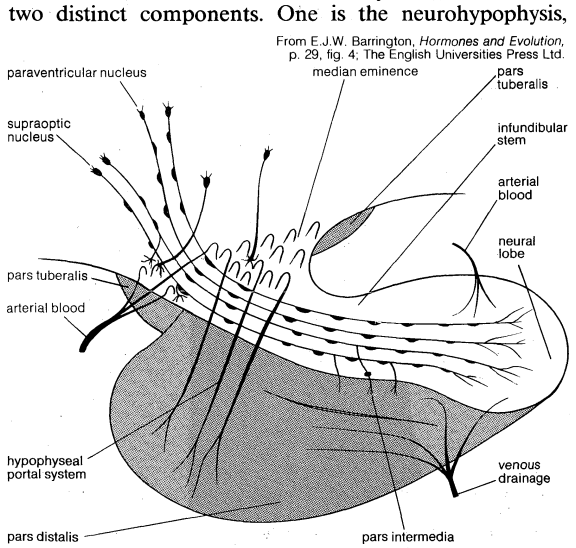


Figure 16: Elements in a generalized mammalian pituitary gland.

which forms as a downgrowth of the floor of the brain and gives rise to the median eminence and the neural lobe; these structures are neurohemal organs. The other is the adenohypophysis, which develops as an upgrowth from the buccal cavity (mouth region) and usually includes two glandular portions, the pars distalis and the pars intermedia, which secrete a number of hormones. The hormones secreted by the adenohypophysis are protein or polypeptide in nature and vary in complexity; as a result, their chemical constitution has not always been as fully characterized as has that of structurally simpler molecules of some other endocrine secretions. Functional analysis of these hormones also is difficult, for the targets of certain hormones of the adenohypophysis, called tropic, or trophic, hormones, are other endocrine glands. The action of such tropic hormones can be understood only in the light of the mode of function of the endocrine glands they regulate.

Adenohypophysis. *Growth hormone (somatotropin; STH).* Growth hormone is a protein, the primary structure of which has been fully established for the human and bovine forms of the hormone. It is probably universally distributed in gnathostomes (vertebrates with jaws), in which it is essential for the maintenance of growth, but its presence in agnathans (jawless vertebrates) has not yet been established with certainty. The physical and chemical properties of growth hormone (Table 17), which differ from species to species, are associated with marked differences in biological activity. Only part of the molecule, however, is actually responsible for its biological activity, for up to 25 percent of it can be lost without causing any decline in potency. Man responds to growth hormones obtained from other primates, but the rat responds to those from a wide range of species. Even more striking, growth of teleost (bony) fishes, which stops if the pituitary gland is removed, can be restarted by treatment with mammalian growth hormone; on the other hand, preparations of pituitary glands from these fishes have no effect on the growth of mammals. The growth hormones of lungfishes, which are closely related to the terrestrial vertebrates, and of sturgeons, which are primitive members of the evolutionary line that led to bony fishes, affect mammalian growth, perhaps because these hormones have a more generalized molecular structure.

Table 17: Some Physicochemical Properties of Growth Hormones of Mammals

	human	monkey	bovine	sheep	pig	whale
Molecular weight	21,500	23,000	45,000	47,800	41,600	39,900
Isoelectric point (pH)	4.9	5.5	6.8	6.8	6.3	6.2
Sedimentation coefficient	2.18	1.88	3.19	2.76	3.02	2.84
Diffusion coefficient	8.88	7.20	7.23	5.25	6.54	6.56
Disulfide linkages	2	4	4	5	3	3

Source: G.H. Li, *Perspectives in Biology and Medicine*, 11, 1968.

Effects of growth hormone

Growth is such a complex process that definition of the growth hormone's mode of action is difficult. One of its known effects is an increase in the rate of protein synthesis, which is to be expected, since growth involves the deposition of new protein material. In addition, growth hormone affects the metabolism of certain ions (including sodium, potassium, and calcium), promotes the release of fats from fat stores, and influences carbohydrate metabolism in ways that tend to cause an increase in the level of glucose in the bloodstream. The last action creates a demand for an increased output of insulin (a hormone secreted by the pancreas), which acts to return the blood-glucose level to normal. Prolonged treatment of dogs with growth hormone can overstrain the pancreatic tissue in which insulin is synthesized and bring about a diabetic condition, in which insulin is formed in inadequate quantities. It is unlikely, however, that this is a factor in establishing diabetes mellitus in man. Excess secretion of growth hormone does, however, have damaging effects in man, for it produces overgrowth of the skeleton. If this occurs in youth, before the closure of the epiphyses (ends) of the long bones, it results in gigantism. If it occurs afterward, it causes acromegaly, in which the disturbance is more serious, with enlargement of the bones and soft tissues, and consequent distortion of the skull.

Prolactin. Prolactin is a protein hormone that in sheep has a molecular weight of about 23,000 (based on a molecular weight for hydrogen of one). It is doubtful that it is present in agnathans, but it is widely distributed in jawed vertebrates. In female mammals, prolactin initiates and maintains the secretion of milk, the mammary glands having been previously prepared for this function by the action of other hormones. In the female rat prolactin also maintains the secretion of the hormone progesterone, which is formed by the corpus luteum, an endocrine gland of the ovary; *i.e.*, prolactin, termed luteotropin in the rat, is a gonadotropin (see below *Hormones of the reproductive system*) in this animal, because its target is an endocrine gland. Evidence is accumulating that the molecular structures of prolactin and growth hormone are similar. This explains why they show some overlap in biological properties; in particular, administration of prolactin promotes some growth in many terrestrial vertebrates. Human growth hormone has prolactin-like luteotropic properties, and it is not yet certain that man actually has a distinct prolactin hormone.

Prolactin itself shows remarkable variety in biological action from one vertebrate group to another. It promotes the production of so-called crop-milk with which pigeons feed their young, and the associated changes in structure and arrangement of the wall of the crop provide a convenient means to assay the hormone. In certain newts (*Triturus* species) prolactin induces the change of behaviour that drives young animals into the water (water-drive action). In bony fishes, prolactin is concerned with the regulation of the level of sodium in blood plasma; it therefore is essential in some teleost species (*e.g., Poecilia lattipinna*) for the maintenance of life in fresh water. Although other teleosts (*e.g.,* eels) can survive in fresh water after hypophysectomy, this means only that prolactin is but one factor in a complex regulatory mechanism involving several factors. Mammalian prolactin can regulate sodium metabolism when given to eels and can maintain the life of hypophysectomized *Poecilia.* Yet, although other convincing evidence suggests that the hormone must be present in the pituitaries of these teleosts, preparations of their glands tested on pigeons do not have a typical crop-stimulating action. This evidence is best accounted for by supposing that the prolactin molecule has undergone evolutionary changes in its molecular structure and biological properties and has also established specific adaptive relationships with target organs such as the crop and mammary glands.

Adrenocorticotropic hormone (ACTH; corticotropin). ACTH is present in all jawed vertebrates but has not yet been decisively demonstrated in agnathans. It regulates the activity of part of the outer region (cortex) of the adrenal glands (considered below under *Hormones of the adrenal glands*). In mammals its action on the adrenal cortex is limited to areas called the zona reticularis and zona fasciculata, in which important steroid hormones (*e.g.,* cortisol and corticosterone, known as glucocorticoids) are formed; ACTH does not affect the synthesis of the mineralocorticoid hormone aldosterone, which takes place chiefly in the outer cortical region (zona glomerulosa). Evidence strongly suggests that the action of ACTH is mediated by a substance known as CAMP (cyclic 3′,5′-adenosine monophosphate), the rate of synthesis of which increases in adrenal tissue in the presence of ACTH; CAMP in turn promotes synthesis of enzymes necessary for the formation of cortisol and corticosterone. The relationship between ACTH and the adrenal cortex is an example of the negative feedback characteristic of endocrine systems; *i.e.,* a decrease in the level of glucocorticoids circulating in the bloodstream evokes an increase in the secretion of ACTH, which, by stimulating the secretory activity of its target gland (the adrenal cortex), tends to restore to normal the level of glucocorticoids in the bloodstream. The release of ACTH can also be influenced by the level of circulating adrenaline, which is not surprising in view of the close functional relationship between the hormones of the adrenal cortex and medulla.

Negative feedback

The ACTH of mammals is a polypeptide molecule consisting of 39 amino acids, only the first 20 of which are required for full activity. This region, often referred to as the active centre, is constant in composition in all mammals studied thus far; the remainder of the molecule varies slightly in amino-acid composition among different species. Since, however, the mammalian hormone is active in all vertebrates, ACTH structure probably varies little from one class to another. The concept that biological activity is localized in an active centre of a complex molecule is applicable to other polypeptide and protein hormones, including growth hormone, whose structure, as noted previously, can be partly lost without causing loss of activity. The concept of an active centre, however, raises the question of the function of the rest of the molecule. It may serve as the site of antigenic properties or of structural features important in establishing relations with specialized receptors in target cells.

Thyrotropin (thyroid-stimulating hormone; TSH). Thyrotropin regulates the thyroid gland through a feedback relationship similar to that for ACTH; thyrotropin increases the secretion of the hormones from the thyroid gland and, if its action is prolonged, evokes increase in cell number (hyperplasia) and increase in size of the gland. One consequence of an overactive thyroid in man is a bulging of the eyes (exophthalmos). The cause of this is obscure, although it has been thought to result from the action of a distinct exophthalmos-producing substance that, while closely associated with thyrotropin, can be chemically separated from it. Thyrotropin, which is probably absent from agnathans, is a glycoprotein; *i.e.,* a protein combined with carbohydrate. Its molecular weight is estimated to be about 26,000 to 30,000 in mammals. Some variability occurs in the degree of response obtained when a hormonal preparation from one species is tested on other species. This suggests, as with prolactin, that it has undergone molecular evolution.

Follicle-stimulating hormone (FSH). FSH is termed a gonadotropin because it is concerned with the regulation of the activity of the gonads, or sex organs, which are endocrine glands as well as the sources of eggs and sperm. FSH stimulates development of the graafian follicle, a small vesicle containing an egg, in the ovary of the female mammal; in the male, it promotes the development of the tubules of the testes and the differentiation of sperm. FSH, like thyrotropin, is a glycoprotein, with an estimated molecular weight (in man) of 41,000 to 43,000. The effects of FSH are discussed further in *Hormones of the reproductive system,* below.

Function of FSH

Luteinizing hormone (LH; interstitial-cell-stimulating hormone; ICSH). Luteinizing hormone is another gonadotropin, a glycoprotein with a molecular weight of 26,000 in man. In the female mammal it promotes the transformation, following release of the egg (ovulation), of the graafian follicle into the corpus luteum, an endocrine gland; its complex functional interrelationship with FSH is dealt with below in *Hormones of the reproductive system.* In the male, luteinizing hormone promotes the development of the interstitial tissue (Leydig cells) of the testes and hence promotes the secretion of the male sex hormone, testosterone. It may be associated with FSH in this function. The interrelationship of LH and FSH has made it difficult to establish with certainty that two separate hormones exist, particularly since both are glycoproteins. Although the existence of two hormones has been established in mammals, the situation in lower vertebrates is not yet certain. All vertebrates undoubtedly have gonadotropic activity in their pituitary glands; but, although FSH-like and LH-like effects are detectable, it is not yet clear that two distinct hormones always exist.

Effects of MSH

An unexpected property of mammalian FSH and LH is that both have a thyrotropic action (*i.e.,* stimulate secretion of thyroid hormones) in lower vertebrates. This so-called heterothyrotropic effect has led to the supposition that FSH, LH, and thyrotropin may have evolved by modification of a common ancestral glycoprotein molecule, resulting in an overlap of properties. Similar examples are pointed out in later sections.

Melanocyte-stimulating hormone (MSH; intermedin). This hormone, secreted by the pars intermedia region of the pituitary gland, regulates colour changes in animals by promoting the concentration of pigment granules in pigment-containing cells (melanocytes, chromatophores) in the skin of lower vertebrates; MSH acts in conjunction with the nervous system in bony fishes and reptiles. No response involving physiological colour change is found in birds and mammals, although the hormone is secreted by them, even in species in which a pars intermedia region is no longer distinguishable in the adenohypophysis. The reason for the presence of MSH in birds and mammals is not clear since the function of the hormone in these animals has not yet been established. MSH is known to influence the behaviour of mammals and the total amount of pigment in their skin, which darkens in man after administration of large doses of the hormone. This type of change, however, which results from a change in the total amount of pigment present, is called a morphological colour change, in contrast to the physiological one that occurs in the skin of lower vertebrates.

As noted above, MSH exists in two forms. α-MSH contains 13 amino acids, which are found in the same sequence in all species studied thus far; β-MSH has 18 amino acids, in sequences that differ in different species. Remarkable are the

ormone									amino acid sequence																		
CTH:									Ser	Tyr	Ser	Met	Glu	His	Phe	Arg	Try	Gly	Lys	Pro	Val	Gly	Lys	Lys	Arg	Arg	Pro
(pig, sheep, beef)									1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
lpha-MSH:							CH_3CO		Ser	Tyr	Ser	Met	Glu	His	Phe	Arg	Try	Gly	Lys	Pro	Val	NH_2					
(pig, beef, horse)									1	2	3	4	5	6	7	8	9	10	11	12	13						
eta-MSH:					Asp	Glu	Gly	Pro	Tyr	Lys	Met	Glu	His	Phe	Arg	Try	Gly	Ser	Pro	Pro	Lys	Asp						
(pig)					1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18						
eta-MSH:					Asp	Ser	Gly	Pro	Tyr	Lys	Met	Glu	His	Phe	Arg	Try	Gly	Ser	Pro	Pro	Lys	Asp						
(beef)					1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18						
eta-MSH:					Asp	Glu	Gly	Pro	Tyr	Lys	Met	Glu	His	Phe	Arg	Try	Gly	Ser	Pro	Arg	Lys	Asp						
(horse)					1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18						
eta-MSH:	Ala	Glu	Lys	Lys	Asp	Glu	Gly	Pro	Tyr	Arg	Met	Glu	His	Phe	Arg	Try	Gly	Ser	Pro	Pro	Lys	Asp						
(human)	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22						

Figure 17: Relationships between the structure of melanocyte-stimulating hormones (MSH) and the first 19 amino acids of adrenocorticotropin (ACTH).

Adapted from R.S. Harris, ed., *Vitamins and Hormones* (1961); Academic Press

facts that the 13 amino acids of α-MSH are identical with the first 13 amino acids of ACTH and that both α and β forms of MSH have a heptapeptide (seven-amino-acid) sequence that has some melanocyte-stimulating activity, and that is identical with an amino-acid sequence of ACTH (see Figure 17). This close correspondence in sequence can hardly be coincidental and suggests, as has been postulated above for FSH, LH, and thyrotropin, that ACTH and α- and β-MSH may have differentiated within the adenohypophysis by evolutionary modification of a common ancestral molecule. A change in biological activity results from modifications in the amino-acid composition; β-MSH preparations from the pig and the horse, for example, are five times more effective than those of the ox in evoking pigment dispersion in frogs. MSH molecules do not show ACTH activity, which is dependent on the presence of amino acids that occur in the region of the molecule not found in MSH. On the other hand, ACTH does have a slight effect on pigment dispersion, presumably because its structure contains the heptapeptide sequence mentioned above.

Evidence shows that each of the adenohypophysial hormones is secreted by a specific cell type. The cell types can be differentiated by staining sections of the pituitary gland, and known changes in the output of an individual hormone, induced experimentally or correlated with phases in the life cycle, can be shown to correspond with changes in the appearance of the corresponding cell type.

Regulation of adenohypophysial activity

The regulation of the activity of the secretory cells of the adenohypophysis depends upon its association with the floor of the brain and results from the existence of a neurosecretory system located mainly, perhaps entirely, in the hypothalamic region there. Much remains to be learned about this system, which involves the passage into the adenohypophysis of neurosecretions from the hypothalamus called hypothalamic releasing factors. Chemical characterization of these factors shows them to be simple polypeptides, in which respect they resemble the hypothalamic polypeptide hormones (discussed in the next section). This neurosecretory system is best understood in mammals, in which good evidence has been found for the existence of a separate releasing factor for each hormone secreted by the pars distalis region of the adenohypophysis; a similar arrangement probably exists in other gnathostomes. The situation in agnathans is obscure, but the anatomical organization of the pituitary glands of these animals implies at least some form of chemical communication between the hypothalamus and the pituitary gland.

Chemical communication is achieved by two routes. One route is by the entry of neurosecretory-cell fibres from the hypothalamus into the adenohypophysis, so that the hypothalamic factors, when released, are either in immediate contact with the secretory cells (Figure 15B) or in blood capillaries very closely related to them. This route is characteristic of the pars intermedia region, in which neurosecretory fibres from the hypothalamus control the functioning of the secretory cells. If the pars intermedia is separated from its direct connection with the floor of the brain, for example, MSH secretion in amphibians increases, and prolonged darkening of the skin results. Secretory activity of the pars intermedia cannot then be regulated again until the nerve fibres have regenerated.

Direct innervation similar to that of the pars intermedia is also found in the pars distalis of bony fishes. Here neurosecretory fibres arise from a localized region of the hypothalamus, called the nucleus lateralis tuberis, and end in contact either with the various types of secretory cells or with blood capillaries related to them. The other route of chemical communication to the pars distalis is found in many fishes and in all terrestrial vertebrates; it is a vascular route that depends upon the median eminence, which lies at the front end of the neurohypophysis (see Figure 16). The median eminence is a neurohemal organ containing a capillary bed into which hypothalamic neurosecretory fibres discharge their releasing factors. These are then transmitted through blood vessels known as the hypophysial portal system, into the capillaries of the pars distalis, where each factor influences its specific target cells (compare Figure 15A).

Both hypothalamic neurosecretory routes have the same physiological significance; *i.e.*, they provide chemical communication between the adenohypophysis and the central nervous system, thus making it possible for the latter to regulate the activity of the gland (and also of the endocrine glands its tropic hormones influence) in response to the demands of both the internal and external environments. The hypothalamic neurosecretory system is also involved in the function of the negative-feedback mechanisms that regulate the secretion of the tropic hormones. As already mentioned for ACTH, the secretions of tropic hormones from the adenohypophysis are controlled by bloodstream levels of the hormones secreted by their target glands; the hormones of the target glands may act directly on specific adenohypophysial cells or indirectly by influencing the output of releasing factors from the hypothalamus.

Neurohypophysis and the polypeptide hormones of the hypothalamus. Another neurosecretory system, which involves the hypothalamic region of the brain and the neurohypophysis of the pituitary gland, originates in groups of neurosecretory cells in the hypothalamus called, in mammals, the nucleus supraopticus and the nucleus paraventricularis and, in lower vertebrates, the nucleus preopticus. Neurohormones from these regions pass along the axons of the neurosecretory cells to the neural lobe (see Figure 16) bound to a protein called neurophysin (molecular weight of 20,000 to 25,000). In the neural lobe, which is the neurohemal organ of this neurosecretory system, the hormones separate from neurophysin and are released into the bloodstream.

Effects of hypothalamic polypeptide hormones

In most mammals, the neurohormones are oxytocin and arginine vasopressin. Both have relatively simple and very similar molecular structures; each is composed of nine amino acids arranged as a ring, which is formed by the linkage of two molecules of the amino acid cysteine (a disulfide linkage —S—S—), and a short side chain (Table 18). The two hormones differ in structure only at amino acids numbered 3 and 8. In some species of the family Suidae (pig, peccary, hippopotamus) arginine vasopressin is replaced by lysine vasopressin; in others, both may be present. The difference between the two vasopressin hormones is that one has the amino acid lysine (Lys) at position 8; the other has arginine (Arg). Both the vasopressins and oxytocin show some overlap of activity, which is a consequence of the similarities in their molecular structures. Preparations of the three hormones evoke responses from the mammalian kidney, from the epithelial-cell layer of the frog bladder, and from the smooth muscle in blood vessels, uterus, and milk glands. The slight variation in amino-acid composi-

tion, however, affects the levels of the responses; *i.e.*, the vasopressins differ slightly from each other in response, and oxytocin differs markedly from both. Each, therefore, is said to have a characteristic pharmacological spectrum, and all have some medical use.

The primary actions of oxytocin are the promotion of uterine contraction (of value in obstetrical medicine) and the release of milk during suckling. The stimulation exerted upon the nipples during suckling leads to the transmission of nerve impulses to the hypothalamus. These bring about the discharge of oxytocin, which causes contraction of the smooth muscle of the small ducts of the mammary glands and the release of milk. Although the vasopressins cause an increase in blood pressure in mammals through vasoconstriction (*i.e.*, contraction of blood vessels), this action requires a high concentration of hormone and is probably not a normal physiological effect. The primary action of the vasopressins is on the kidneys; it brings about a reduction in the output of urine. As a result arginine vasopressin is commonly called the antidiuretic hormone (ADH). A lack of this hormone in man results in a copious flow of urine, a condition called diabetes insipidus, which is readily alleviated by preparations containing arginine vasopressin from bovine sources.

The antidiuretic action of vasopressin is thought to depend upon its binding to the outer surface of the kidney tubule, resulting in an increase in the uptake of sodium from the urine into the tubule cells and, concurrently, an increase in the uptake of water. The amount of water, however, is greater than can be accounted for merely by increased diffusion of sodium into tubule cells, suggesting that ADH increases either the number of or the size of pores on the surfaces of the cells. One stimulus that increases the release of vasopressin is a rise in the concentration of certain substances—chloride, for example—in blood plasma. These substances act directly upon the neurosecretory cells, although other receptors may also be involved. Another stimulus is a lowering of plasma volume, which probably acts chiefly through receptors in the vascular system, particularly in the heart and in the carotid blood sinuses. Both conditions necessitate increased retention of fluid; as soon as normal conditions are restored in the bloodstream, the secretion of ADH is reduced by negative feedback.

Origin of hypothalamic hormones

Oxytocin and the vasopressins are members of a series of hormones of which seven members have thus far been fully characterized. The existence of others is suspected (see Table 18). All show the same molecular structure but differ with respect to individual amino acids. The hormones are believed to have been derived from each other by mutations that resulted in one amino-acid substitution at a time; the starting point in the series is arginine vasotocin, which is the only one of the series found in agnathans. Two types of molecule are found in gnathostomes—a result, presumably, of a genetic duplication that established two lines of evolution. One line (basic vasopressor principles) is constituted mainly of arginine vasotocin, which is present in all gnathostomes except mammals; amino-acid substitution in the molecule gave rise to the vasopressins of mammals. The second line (neutral oxytocin-like principles) is represented by oxytocin, isotocin, glumitocin, and mesotocin. Each evolutionary line tends to have characteristic molecules, but the molecular history in the second line is not clear. Oxytocin is thought to exist in some lower gnathostomes, and it is not yet certain whether it or mesotocin is phylogenetically the older molecule.

The functions of the hypothalamic polypeptide hormones in lower vertebrates are not yet clear, except to some extent in amphibians, in which arginine vasotocin evokes the so-called Brunn (water-balance) response; that is, water accumulates within the body as a result of a combination of increased water uptake through the skin and the wall of the bladder and decreased urinary output. This response, which also involves the uptake of sodium by the skin, is found only in the more terrestrial members of the Amphibia, in which it is an adaptation that enables them to conserve water. It is not yet known whether or not comparable adaptive specializations are associated with the molecules characteristic of the other groups of lower vertebrates. There is some evidence that hypothalamic polypeptides may be involved in the movements of water and ions (charged particles) in fishes. Changes in the functions of the polypeptide hypothalamic hormones during vertebrate evolution have occurred, partly as a result of evolution of their targets; *e.g.*, water balance in amphibians is mediated by a hormonal molecule that was already present in agnathans and was thus a part of the earliest hormonal endowment of vertebrates.

Table 2: The Structure and Distribution of the Hypothalamic Polypeptide Hormones of Vertebrates*

Basic Vasopressor Principles

1. Arginine vasotocin

Cys	Tyr	Ile	Gln	Asn	Cys	Pro	Arg	Gly(NH_2)
1	2	3	4	5	6	7	8	9

Exists in all fishes; possibly in all vertebrates, at some stage of development

2. Arginine vasopressin

Cys—Phe—Cys—Arg—

Exists in most mammals except the pig family

3. Lysine vasopressin

Cys—Phe—Cys—Lys—

Exists in mammals of the pig family

Neutral Oxytocinlike Principles

4. Oxytocin

Cys—Ile—Cys—Leu—

Exists in mammals, birds, reptiles, amphibia (?), Dipnoi (?), and holocephalians (?)

5. 8 Ile oxytocin (Mesotocin)

Cys—Ile—Cys—Ile—

Exists in reptiles, amphibia, and Dipnoi

6. 4 Ser, 8 Ile oxytocin (ichthyotocin, isotocin)

Cys—Ile • Ser—Cys—Ile—

Exists in teleost, holostean, and brachiopterygian fishes

7. 4 Ser, 8 Glu(NH_2) oxytocin (glumitocin)

Cys—Ile • Ser—Cys—Gln—

Exists in elasmobranch fishes, particularly the skate

8. Unknown elasmobranch oxytocic principle (EOPI) (Val, Ser peptide[s]?).
Exists in *Squalus acanthias* and perhaps other sharks

*Amino acids indicated only by lines are the same as those present in arginine vasotocin.

Source: Hoar and Randall (eds.), *Fish Physiology*, 1969.

HORMONES OF THE THYROID GLAND

Biosynthesis. The two thyroid hormones, thyroxine (3,5,3′,5′-tetraiodothyronine) and 3,5,3′-triiodothyronine, are formed by the addition of iodine to an amino-acid (tyrosine) component of a glycoprotein called thyroglobulin. Thyroglobulin is stored within the gland in follicles as the main component of a substance called the thyroid colloid. This arrangement, which provides a reserve of thyroid hormones, perhaps reflects the frequent scarcity of environmental iodine, particularly on land and in fresh water. Iodine is most abundant in the sea, where thyroidal biosynthesis probably first evolved. Although the possibility that the thyroid hormones originated as metabolic by-products is suggested by the widespread occurrence in animals of the binding of iodine to tyrosine, the binding commonly results only in the formation of iodotyrosines, not the thyroid hormones. On present evidence, only the vertebrates and the closely related protochordates have a mechanism to synthesize significant amounts of biologically active thyroid hormones.

The synthesis of thyroid hormones in vertebrates begins with the active uptake by thyroid-gland cells of inorganic iodide circulating in the bloodstream; the inorganic iodide is oxidized (combined with oxygen) during a reaction catalyzed by an enzyme (iodide peroxidase). The product of this reaction (active iodine) combines with tyrosine components of the thyroglobulin molecule to form two compounds (3-monoiodotyrosine and 3,5-diiodotyrosine), which then join to form the active hormones. The synthesis of the thyroid hormones is inhibited by certain chemical agents called goitrogens, which reduce the output of thyroid hormones, thereby causing, through negative feedback, an increased output of thyrotropin and hence an enlargement of the thyroid gland (see below). Some goitrogens (*e.g.,* thiocyanates) reduce or inhibit the uptake of iodide; others (*e.g.,* thiourea, thiouracil) inhibit the peroxidase system and thus prevent the binding of iodine to thyroglobulin.

Release of the thyroid hormones into the bloodstream begins when the thyroid cells take up droplets of the stored thyroid colloid. The thyroglobulin in these droplets is then hydrolyzed (broken down in a reaction involving the elements of water) by an enzyme to form both iodotyrosines and the hormones. Normally, only the latter pass out of the cells in significant quantities. The iodine is removed from the iodotyrosines, which are not hormonally active, by an enzyme (deiodinase), and the iodine thus is conserved and used again. The hormones, usually bound to proteins (globulin and albumin) in the bloodstream, where they constitute the protein-bound iodine of the plasma, must be unbound from the proteins before they can function. The iodine is removed from the hormones largely in the liver and in the kidneys, and most of it returns to the thyroid gland, an economy that again emphasizes the need for conservation; some iodine, however, is lost in the alimentary tract.

Conservation of iodine

Synthesis of the thyroid hormones is regulated by the level of circulating hormones (*i.e.,* a negative-feedback mechanism) operating, as indicated earlier, partly by direct action on the thyrotropin-secreting cells of the pituitary gland and partly by indirect action on the hypothalamus and its thyrotropin-releasing factor. Thyrotropin attaches to the cells of the thyroid gland and may exert its effect by stimulating CAMP synthesis. It causes resorption of thyroid colloid and increases the rates of both glucose metabolism and protein synthesis as secretion of thyroid hormones increases in response to it. After the thyroid gland of the rat has been under thyrotropin stimulation for two or three hours, an increase in the size of the cells of the gland occurs, along with an increase in iodide uptake into them; prolonged thyrotropin action causes a marked enlargement of the gland (goitre), which in man may become externally apparent as a swelling. Goitres, which are of various types, result from a negative-feedback reaction that attempts to maintain output from the thyroid gland.

Effects. One established effect of the thyroid hormones in mammals is an increase in metabolic rate and in oxygen consumption, but the effects of the hormones undoubtedly are more wide-ranging than this. On the one hand, impairment of the thyroid function in mammals results in disturbances in the processes of growth and maturation. Both growth and maturation disturbances occur in the cretinous dwarfism resulting from thyroid deficiency in newborn infants; on the other hand, the metabolic effect is not apparent in lower vertebrates (*e.g.,* fish), even though treatment of these animals with thyroid hormones promotes an increase in the growth rate, provided pituitary growth hormone is also secreted. In addition, evidence suggests that, in lower vertebrates, the thyroid hormones are active during moments of stress in the life cycle (*e.g.,* migration and reproduction) and affect the activity of the central nervous system. Disturbance of thyroid output also affects reproduction in mammals, impairing the functioning of the ovary, for example, and causing irregularities of the ovarian cycle.

Effects of thyroid hormones on frog metamorphosis

The complex effects of thyroid hormones are well documented in the metamorphosis, or change in body form, of the amphibian tadpole into a frog. Metamorphosis, which involves a diversity of integrated morphological and biochemical changes, requires the presence of the thyroid gland and depends upon a delicate balance between the changing output of its hormones and changing sensitivities of the target tissues. Studies involving the tail of the frog tadpole show that the thyroid hormones directly promote the formation of the enzymes needed for reduction of the tail and suggest that the diverse effects produced in vertebrates by the thyroid hormones might depend upon their capacity to regulate protein metabolism, in which case the target cells would have to be adapted to respond by appropriate patterns of enzyme synthesis.

Ultimobranchial tissue and calcitonin. The discovery of calcitonin (thyrocalcitonin) in 1961 demonstrated the importance of comparative studies in endocrinology. It originally had been thought that this hormone, which is present in preparations made from mammalian thyroid glands, was secreted by the parathyroid glands, which in some species are combined with the thyroid gland. Later, the hormone was concluded to be a secretion of the thyroid gland itself. In fact, calcitonin is not a product of either of them. Its actual source is the ultimobranchial tissue, represented in vertebrates from fishes upward by the ultimobranchial gland, which develops from the hinder part of the pharynx. Ultimobranchial tissue is the source of distinctive cells (called light, C, or parafollicular cells), which are found in the thyroid gland of mammals; in birds, however, the ultimobranchial gland is separate, thus making it possible to remove the gland and to show that it is the source of the hormone. The molecular structure of hog calcitonin is that of a polypeptide, containing 32 amino acids and having a molecular weight of about 3,600. The calcitonin of the salmon, which is more potent than that of the pig, has the same number (but some different types) of amino acids, and the molecular weight is 3,427.

Effects of calcitonin

Calcitonin lowers the level of calcium in the blood (hypocalcemic action) when it rises above the normal level. Its secretion probably is regulated by a negative-feedback relationship between the gland and the blood plasma. The hormone affects bone, which is an active tissue. It undergoes not only growth but also remodelling as it adapts to the changing patterns of stress to which it is subjected; its calcium exchanges continuously with that of the plasma. The effect of calcitonin is to decrease the mobilization (resorption) of calcium from the skeleton into the blood plasma. In this respect, as is discussed in the next section, it is opposite in direction to the effect of parathormone of the parathyroid glands. Little is known of the action of calcitonin in the lower vertebrates, but its presence in fish raises interesting functional problems. Elasmobranch fishes (*e.g.,* sharks) lack bone, and many bony fishes have a type of bone that cannot be remodelled; the hormone, therefore, cannot act in these vertebrates as it does in higher ones. It is possible that in these fishes the hormone may control the level of plasma calcium by regulating its movement across cell membranes.

PARATHORMONE OF THE PARATHYROID GLAND

The parathyroid glands, which are found only in terrestrial vertebrates (amphibians, birds, reptiles, and mam-

mals), develop from certain pharyngeal pouches, which are embryonic remnants of the gill slits of fish. The parathyroid glands secrete a hormone called parathormone (PTH), which is a polypeptide of variable amino-acid composition. PTH, which consists of 83 to 85 amino acids in the human, regulates calcium metabolism in conjunction with calcitonin; its evolution in terrestrial vertebrates may have been an adaptation to the increased demand for continuous skeletal adjustments imposed by the evolution of terrestrial locomotion. Skeletal adjustments must be made without disturbing the delicate calcium balance of the rest of the body, for calcium is involved in maintaining the transport of substances through cell membranes; hence, it has an important role in muscle contractility, excitability of motor end plates in the nervous system, and coagulation of blood.

Removal of the parathyroid glands in mammals causes a fall in the level of calcium in the blood plasma, which, if sufficiently severe, is accompanied by convulsions and other symptoms resulting from increased excitability of the motor nerves. These symptoms can be corrected by injection of appropriate preparations of parathyroid glands. The activity of the glands, like that of the ultimobranchial tissue, is regulated by negative feedback; *i.e.,* lowering of the plasma calcium level increases the output of parathormone (but decreases the output of calcitonin). The hypercalcemic effect (*i.e.,* increase in level of blood calcium) of the hormone depends largely upon its action on bone, since it promotes the transfer of calcium from this tissue into the plasma, probably by a direct action on the active bone-forming cells (osteocytes). In addition, however, parathormone promotes the formation of new bone tissue, and thus also increases its metabolic activity and the turnover of its structural material. Other effects of parathormone, at least in part, contribute to the elevation of plasma calcium; *i.e.,* PTH increases both the absorption of calcium by the intestine and its resorption by the kidney tubule. Since, however, the hypercalcemia induced by the hormone results in more of it passing into the kidney tubule, the net result may be increased excretion of calcium despite the increased resorption. Other actions of the hormone, less easy to relate to its well-defined influence upon calcium metabolism, include a regulatory influence upon the level of magnesium in blood plasma and upon the rate of removal of phosphate from urine.

In general, therefore, the action of parathormone is opposite in direction to that of calcitonin. Parathormone keeps the level of blood calcium up to its normal value; on the other hand, calcitonin ensures, through its hypocalcemic action, that the level does not rise far above this critical point. The combined actions of the two hormones serve to illustrate the importance of endocrine regulation in homeostasis. Vitamin D is a third factor in calcium regulation; its absence in young children results in skeletal malformations (rickets). Parathormone is unable to regulate the absorption and mobilization of calcium in the absence of vitamin D, which is also associated with the hormone in promoting mobilization of magnesium from bone and perhaps in the movement of phosphate within the kidney tubule.

Role of vitamin D in calcium regulation

HORMONES OF THE PANCREAS

Insulin. The vertebrate pancreas contains, in addition to the zymogen cells that secrete digestive enzymes, groups of endocrine cells called the islets of Langerhans. Certain of these cells (the B, or beta, cells) secrete the hormone insulin, inadequate production of which is responsible for the condition called diabetes mellitus. Insulin and the characteristic B cells are present in gnathostomes and in agnathans; in the latter, however, the islet cells are not associated with zymogen cells to form a typical pancreas. Insulin is, as mentioned earlier, a polypeptide molecule composed of two chains of amino acids, an A chain of 21 amino acids containing an intrachain disulfide linkage (—S—S—) and a B chain of 30 amino acids. The two chains are linked by two other disulfide linkages, the destruction of which destroys the activity of the molecule. It is thought that the molecule first appears in the B cell as the single-chain compound proinsulin, which is disrupted by an enzyme-catalyzed reaction to form the two chains of the active hormone. As with other polypeptide hormones, extensive variation in amino-acid composition of the molecule occurs among different species, with the differences tending to be greater between the more widely separated species—*e.g.,* between fish and mammal. The variations in amino-acid composition have little effect on the biological activity of the molecules, but certainly influence their immunological reactions; this suggests that the two properties depend on the amino-acid sequences at different parts of the molecule.

Injection of insulin lowers blood-sugar (glucose) levels, but this so-called hypoglycemic effect is only one expression of the wide-ranging influence of insulin on storage and mobilization of energy, in which the target tissues of primary importance are muscle, adipose (fat) tissue, and liver. The actions of insulin on these tissues are varied. First, it promotes the use of the sugar glucose as an energy source; at the same time, it encourages the storage of excess carbohydrate as glycogen, the storage carbohydrate of animals. Second, insulin reduces the use of fat as an energy source and promotes its storage. Third, it reduces the use of protein as an energy source and promotes the formation of proteins from amino acids.

Insulin probably acts on carbohydrate metabolism in muscle by increasing the ability of glucose to pass through the muscle-cell membranes. This effect depends on a specific interaction between the cell membrane and the hormone; although the same effect occurs in adipose (fat) tissue, it does not occur either in the liver or in the central nervous system, despite the latter's complete dependence upon glucose for its energy supply. After the entry of glucose into a muscle cell, phosphate is added to the molecule, and two compounds form in succession, first glucose-6-phosphate, then glucose-1-phosphate; after these reactions, the metabolism of glucose is probably aided by two secondary actions of insulin (see also METABOLISM). The hormone stimulates the synthesis of an enzyme (glycogen synthetase), thus promoting the transformation of glucose-1-phosphate into glycogen; it also aids in the breakdown of glucose, thus providing energy to the cell. All of these effects contribute to the hypoglycemic (blood-glucose-lowering) action of the hormone. Insulin has other effects on muscle cells—it slows the breakdown of fat and increases the formation of proteins from amino acids. Insulin affects carbohydrate and protein metabolism in adipose tissue much as it does in muscle and also promotes storage of fat.

The action of insulin in liver differs from that in muscle in that it has no direct influence upon the transport of glucose into liver cells; probably, however, insulin promotes the metabolism of glucose within liver cells in much the same way that it does in those of muscle, resulting in increased uptake of glucose from the bloodstream. In addition, insulin decreases gluconeogenesis (the formation of glucose in the liver from amino acids and other noncarbohydrate sources). These various effects cause a decrease in the level of blood glucose. Other actions of the hormone upon the liver include, as in adipose tissue, increases in fat deposition and protein synthesis.

Effects of insulin in liver

The diverse effects of insulin apparently are adaptively linked to regulating the storage and release of energy, but it is difficult to judge whether or not all of the effects result from a single mode of action of the hormone. The interaction of insulin with the muscle-cell membrane suggests that all of its effects might be produced by similar interactions between it and membranes within cells. The mechanism, however, has not yet been established with certainty.

The B cells of the islets of Langerhans respond directly through negative feedback to the level of glucose in the blood that reaches them; *i.e.,* an increase in blood glucose above the normal level (80 to 100 milligrams per 100 millilitres in man) brings about increased synthesis and release of insulin with the result that the level of blood glucose falls. As a consequence, the rate of insulin output then decreases. This, however, is only part of the complex hormonal mechanism that regulates carbohydrate metabolism. Another factor is the hormone glucagon, which is

secreted in the islets of Langerhans by a second cell type, the A (alpha, or A_2) cells. (The function of a third type, the D (gamma, or A_1) cells, has not yet been established.)

Glucagon. Glucagon, which is present in gnathostomes but absent from agnathans, is a polypeptide molecule consisting of 29 amino acids. It strongly opposes the action of insulin, primarily through a hyperglycemic (blood-glucose-raising) effect that results from its promotion of the breakdown of glycogen (glycogenolysis) in the liver, a process that results in the formation of glucose. Glucagon exerts its action by increasing the availability of the enzyme required for the reaction by which glucose units are released from the glycogen molecule. It also reduces the rate of synthesis of glycogen, promotes the breakdown of protein, promotes the use of fat as an energy source, and evokes increased glucose uptake by muscle cells. The last effect, however, may be a consequence of hyperglycemia induced by the increased secretion of insulin.

Another form of glucagon, called gastronintestinal glucagon, is secreted into the blood when glucose is ingested. Its only action appears to be to stimulate insulin secretion, an effect that may provide information to the islet cells of the pancreas about the entry of glucose into the bloodstream. It is also possible that pancreatic glucagon, which is secreted in the islets by the A cells, may directly stimulate the release of insulin from the adjacent B cells without actually entering the bloodstream.

Other hormones that affect insulin release

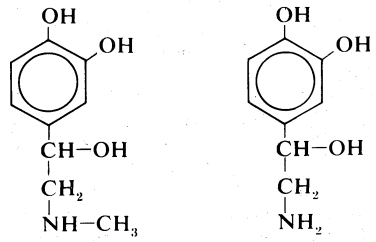
A number of other hormones also influence the release of insulin, mainly through their own actions upon blood-sugar levels. Growth hormone, thyroxine, adrenaline, and cortisol, *e.g.*, may increase insulin release because they can promote a rise in blood sugar through effects upon carbohydrate metabolism. Growth hormone and cortisol can also probably act directly upon the B cells.

The complexity and delicacy of the control of metabolism by insulin and other hormones in mammals illustrate again the importance of homeostasis, the control of which may not be as well organized in the lower vertebrates. Some of the responses in mammals, however, do occur in lower forms; for example, removal of pancreatic islet tissue from fishes produces hyperglycemia. Thyroxine induces hyperglycemia in amphibians, and corticosteroids promote gluconeogenesis in them. Far more information is needed, however, before the evolution of these remarkable regulating mechanisms can be determined.

HORMONES OF THE ADRENAL GLANDS

Chromaffin tissue of the medulla. The adrenal gland of mammals is composed of an outer region, the cortex, which consists of adrenocortical tissue that secretes steroid hormones (steroids are fat-soluble organic compounds), and an inner region, the medulla, which is composed of chromaffin tissue, so called because its cells contain granules that can be characteristically coloured by certain reagents. Chromaffin tissue secretes two hormones, adrenaline (epinephrine) and noradrenaline (norepinephrine), which are members of a class of compounds called catecholamines. Both chromaffin and adrenocortical tissues are present in gnathostomes and probably in agnathans (although the evidence on the latter point is not yet decisive), but the tissues vary in the degree to which they are associated, being completely separated in elasmobranch fishes.

Noradrenaline and adrenaline are each composed of a benzene ring containing two hydroxyl (— OH) groups and an amine (NH_2-containing) side chain as shown below:



adrenaline (epinephrine) noradrenaline (norepinephrine)

Figure 18: The structures of noradrenaline and adrenaline.

During the synthesis of these hormones, a sequence of enzyme-catalyzed reactions in the chromaffin granules of the secretory tissue transforms tyrosine into a compound commonly called dopa (dihydroxyphenylalanine), which then forms dopamine; dopamine then is hydroxylated (*i.e.*, an —OH group is added) to form noradrenaline. Adrenaline is formed from noradrenaline by methylation (the addition of a methyl, or $-CH_3$, group), a reaction that occurs outside the granules of the chromaffin cells. Noradrenaline (but not adrenaline) is also formed in certain neurons (nerve cells), where it functions as one of the chemical transmitter substances.

After their release, both hormones are so rapidly metabolized that they probably remain in the bloodstream only for a few seconds. The first step in the breakdown, which usually occurs in the liver and kidneys, is methylation of one of the hydroxyl groups of the benzene ring; the products (metanephrine or normetanephrine), or compounds derived from them, are excreted in the urine. Small quantities (about 2 to 5 percent of the daily secretion of the gland in man) of nonmetabolized hormones are also found in the urine.

Table 19: Effects of Adrenaline and Noradrenaline in Man

	adrenaline	noradrenaline
Heart rate	increase	decrease
Cardiac output	increase	variable
Total peripheral resistance	decrease	increase
Blood pressure	rise	greater rise
Respiration	stimulates	stimulates
Skin vessels	constriction	less constriction
Muscle vessels	dilation	constriction
Bronchus	dilation	less dilation
Eosinophil count	increase	no effect
Metabolism	increase	slight increase
Oxygen consumption	increase	no effect
Blood sugar	increase	slight increase
Central nervous system	anxiety	no effect
Uterus in late pregnancy	inhibits	stimulates
Kidney	vasoconstriction	vasoconstriction

Source: G.H. Bell, J.N. Davidson, and H. Scarborough, *Textbook of Physiology and Biochemistry*, 7th ed., 1968.

Effects of noradrenaline and adrenaline

Adrenaline and noradrenaline evoke diverse and widespread responses but differ from each other in certain of their effects (see Table 19 for their effects on man). Both influence the heart and blood vessels in ways which, although opposed to each other in a few respects, generally result in an increase in blood pressure and in output of blood from the heart. Both hormones also have metabolic actions. Adrenaline, for example, like glucagon, stimulates glycogenolysis (breakdown of glycogen to glucose) in the liver, which results in the raising of the level of blood sugar; in addition, it increases oxygen consumption and the output of blood from the heart, probably contributing thereby to the regulation of body temperature in mammals. Adrenaline has effects upon the nervous system, which are recognizable subjectively in man by feelings of anxiety and of increased mental alertness.

The chromaffin tissue is closely related to the sympathetic nerves of the autonomic nervous system, which innervates the components of circulation and digestion and controls their involuntary functions; in fact, the two may be said to form a sympatheticochromaffin complex. It is generally assumed that this complex acts to increase the capacity of the animal for effective action in emergencies. At such times, cardiac output increases, blood is distributed with maximum effectiveness, respiration is enhanced, and the nervous system is stimulated. The sympathetic nerves initiate these reactions and directly promote the release of adrenaline and noradrenaline because these nerves directly innervate the chromaffin cells. The hormones are thus able to develop and prolong an integrated set of responses; noradrenaline functions both as a neurohumor chemical transmitter of the sympathetic nervous system and also as a hormone of the chromaffin tissue.

The fact that adrenaline and noradrenaline, which have very similar molecular structures (see above), can exert different actions is probably in part a consequence of the

specialization of their target tissues. It has been suggested by some researchers that the target tissues possess two different kinds of receptors, the alpha type, which responds to noradrenaline, and the beta type, which responds to adrenaline. Evidence for this theory is that adrenaline has a vasodilator effect (*i.e.*, it expands blood vessels), which can be blocked by certain drugs, and noradrenaline has an opposing vasoconstrictor effect, which can be blocked by other drugs. The actions of both hormones are thought to be mediated by CAMP; alpha responses are associated with reduced synthesis of this mediator and beta ones with increased synthesis.

The interpretation of the function of these hormones in mammals has not yet been established as applicable to lower vertebrates in which the hormones are present, but they are known to influence metabolism and heartbeat in some genera. It is possible that in early stages of vertebrate evolution, the sympatheticochromaffin complex evoked more generalized physiological responses than it now does and that more precise action developed in mammals as part of their high level of homeostatic organization. Laboratory studies show that even in mammals the complex is not essential for life; animals from which it has been removed, however, are much less able to resist environmental stresses than are those whose complex is functional.

Adrenocortical tissue of the cortex. The adrenocortical tissue develops from coelomic epithelium (a cell layer surrounding the body cavity, or coelom). In this respect it resembles the endocrine tissue of the gonads, a resemblance emphasized by the fact that both the adrenocortical hormones (corticoids) and the sex hormones are steroids produced by similar metabolic pathways (the structures of some steroid hormones are illustrated below).

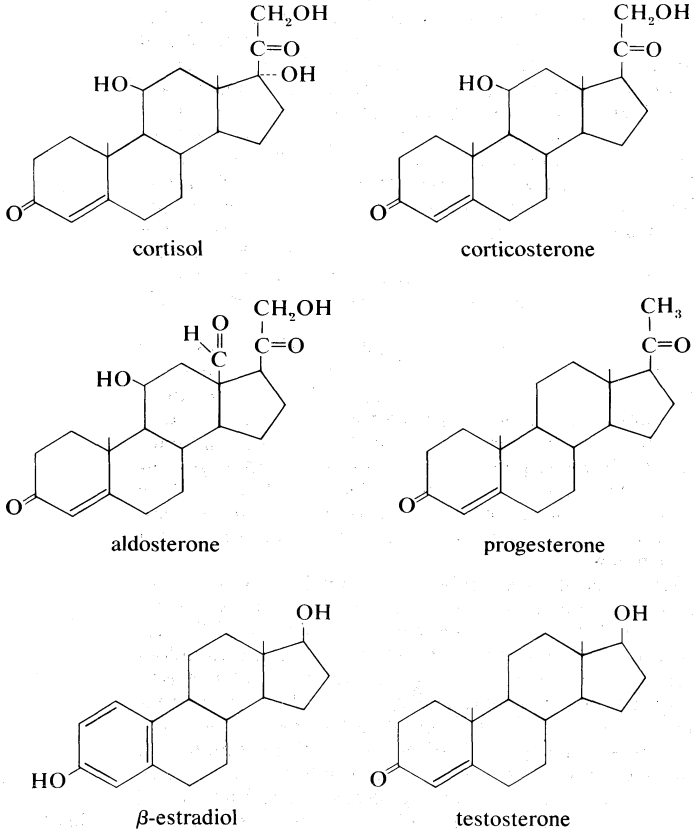


Figure 19: Some steroid hormones of vertebrates.

Active adrenocortical hormones

Many steroids have been isolated from the adrenal cortex, but in most vertebrate groups only three of them are active as hormones; they are cortisol (hydrocortisone; compound F), corticosterone (compound B), and aldosterone. Their biosynthesis is outlined in Figure 20.

The principal sterol of animals is cholesterol, which is formed by a complex series of reactions from a two-carbon compound (acetate). Progesterone, which is derived from

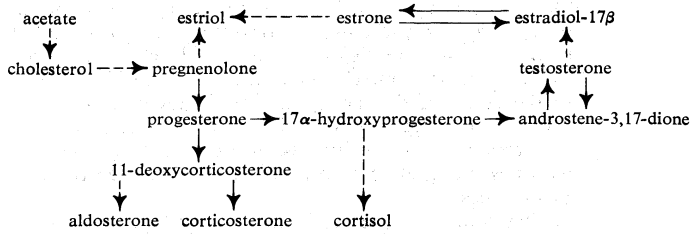


Figure 20: Metabolic pathways involved in the biosynthesis of steroid hormones. Broken lines indicate that more than one step is involved.

Adapted from D.M. Greenberg, ed., *Metabolic Pathways* (1960); Academic Press

cholesterol, can be used to form either corticosterone and aldosterone or cortisol. All three corticoids are bound to proteins during transfer in the bloodstream to their targets; cortisol, for example, is bound to a glycoprotein called transcortin. Some inactivation of the corticoids takes place in the kidney and in the alimentary tract, but most of it occurs in the liver. The metabolic products, which eventually appear in the urine, provide a basis for determination of the output of adrenocortical hormones in man.

The normal secretion of the hormones is best determined by direct measurement of the contents of the venous blood leaving the adrenal gland. In man, the daily secretion rates of the hormones, as determined by this procedure, are cortisol, 20 milligrams; corticosterone, two to five milligrams; aldosterone, 75 to 150 micrograms (one microgram = 1,000,000th of a gram). Very small amounts of aldosterone are secreted, because the molecule has a high level of activity. Animal tissues maintained in culture fluid together with compounds from which the hormones are formed (*e.g.*, acetate, cholesterol, or progesterone containing radioactive isotopes of carbon or hydrogen) show that cortisol and corticosterone are produced in all vertebrates, including the agnathans, although the proportion of each is species-dependent; elasmobranch fishes are unique, however, in having 1-α-hydroxycorticosterone as the principal hormone. Aldosterone is produced by all terrestrial vertebrates. It has also been found in bony fishes, although its function in them has not yet been established as a hormonal one. The presence of aldosterone has not yet been established in elasmobranchs and agnathans, but whether or not this particular molecule occurs in them, the ability to synthesize corticoids must have evolved very early in vertebrate history.

In contrast to the chromaffin tissue of the adrenal medulla, the adrenocortical tissue is essential for life. Two primary functions of the corticosteroids are distinguishable in mammals. One, which contributes to the regulation of carbohydrate metabolism, is an action of cortisol and corticosterone, which are therefore called glucocorticoids. These hormones promote gluconeogenesis (formation of glucose) in the liver and are thus important in maintaining normal blood-sugar levels, particularly during glucose shortage; lack of them results in low levels of blood sugar and an increase in the sensitivity of the liver to insulin (whose effect there is to decrease gluconeogenesis). In addition, lack of the glucocorticoids is associated with a decrease in the entry of amino acids into muscles and an increase in their uptake by the liver, where enzymes required to convert amino acids to glucose must be synthesized.

In contrast to glucocorticoid action is the so-called mineralocorticoid action of aldosterone, which is manifested in mammals in the regulation of sodium metabolism. In the absence of aldosterone, sodium is lost from the body by excretion in urine; secondary consequences include a decrease in blood volume and in the filtration rate of substances through kidney structures called glomeruli. Cortisol and corticosterone also play a minor part in mineral regulation, so that slight overlap in function occurs between the two corticoid types.

The action of aldosterone is exerted mainly upon the distal segment of the nephron (kidney tubule), where it promotes an increase in the permeability of the tubule membrane to the passage of sodium, and also an increase in the quantity of sodium removed into the blood from

the fluid passing through the kidney tubule. At the same time, potassium and hydrogen pass into the fluid from the blood. Aldosterone also exerts other effects. It promotes sodium retention in salivary glands, in sweat glands, and in the colon of the large intestine; it also promotes the excretion of magnesium in the urine. The effects of aldosterone result in an increase in the rate of synthesis of enzymes required to transport these substances through membranes.

Other actions of the corticoids are apparent in patients suffering from Addison's disease, which is caused by a general deficiency in corticoid production. A deficiency of corticoids causes disturbances in urinary output and fat metabolism, diminished resistance to stress, muscular weakness, and nervous disturbances manifested by depression and a general lack of mental alertness. The adrenocortical hormones, then, like the hormones of the chromaffin tissue of the medulla, are involved in resistance to stress. It has been postulated that the response to alarm stimuli initially involves both the sympatheticochromaffin complex and the adrenocortical secretion; then a stage of full resistance occurs that may be followed by mental exhaustion if the alarm stimuli are prolonged. Although a close functional relationship is known to exist between the adrenocortical and chromaffin tissues in mammals, the function of the corticoids in the lower vertebrates has not yet been established. Indications are, however, that the general pattern of action may be similar; for example, the cortisol type of corticoid promotes gluconeogenesis in fish and removal of adrenocortical tissue impairs the metabolism of water and ions in the eel. Any interpretation of corticoid action in teleost, or bony, fishes has to incorporate prolactin, for, as has been noted previously, this hormone also influences the movement of ions.

Resistance to stress

In contrast to the chromaffin cells, the adrenocortical cells are not innervated. Both cortisol and corticosterone production are regulated by the action of ACTH from the pituitary gland on the zona reticularis and the zona fasciculata. The regulation of aldosterone secretion in the zona glomerulosa, however, is associated with the so-called renin–angiotensin system, which is best characterized in mammals. Renin, an enzyme with a molecular weight of about 40,000, is formed in the kidney and is released into the bloodstream, where it catalyzes the formation of angiotensin, a polypeptide molecule. Angiotensin acts upon smooth muscle and raises blood pressure. In man it reduces sodium excretion, probably by a direct action on kidney filtration, and may, in fact, be a true hormone, acting to aid sodium retention. In addition, however, angiotensin contributes to sodium retention by increasing aldosterone secretion. The exact physiological significance of the renin–angiotensin system is not yet known. In one form or another the system is probably widely distributed in vertebrates.

HORMONES OF THE REPRODUCTIVE SYSTEM

The hormones of the reproductive system of vertebrates (sex hormones) are steroids that are secreted, like those of the adrenal cortex, by tissues derived from the coelomic epithelium. Both types of secretory tissues also share biosynthetic pathways (see above *Adrenocortical tissue of the cortex*).

Female hormones. The sex hormones, together with the hypothalamic region of the forebrain and the pituitary gland, form a regulatory system, which is most complex in the female mammal. It is common for sexual activity of vertebrates to be cyclical and for the cycles to be coordinated with the seasons of the year; this ensures that the young are born at the most favourable time. In mammals, however, reproduction is complicated by the need to provide for the intrauterine life of the developing fetus and to ensure that interference by another generation of embryos cannot occur.

Estrogens. Two types of gonadal hormone, estrogens and progestins (Figure 18) are secreted in the female mammal. Estrogens are substances that evoke the cyclical onset of heat, or estrus, during which the animal is sexually active and receptive to the male. Estrus in this sense is not found in human females, but estrogens contribute to the events of the menstrual cycle, bringing about cyclical changes in the reproductive system that are comparable with those accompanying estrus in other mammals.

Types of estrogens

Hormones are secreted from the mammalian ovary by the ovarian follicle, or vesicle, including the granulosa cells immediately surrounding the ovum, or egg, and the cells of the theca, which forms a supporting outer wall for the follicle. The main estrogen secreted is called β-estradiol. The close relationship between the female and the male sex hormones is revealed by the fact that testosterone (the main male hormone) is an intermediate compound in the pathway that leads to the synthesis of estradiol, although another route, which avoids the formation of testosterone, is possible. Other estrogens are also known; the most familiar ones in man and other mammals, estrone and estriol, are much less active than estradiol, estriol being the weakest. Estrone can be converted to estradiol and vice versa in the ovary and in other tissues; *e.g.,* estradiol is converted, particularly in the liver, to estriol, which is an excretory product. The metabolism of these compounds is complex; they may be combined in part with other substances, or they may pass through the bile into the intestine for reabsorption and circulation through the body before excretion in the urine occurs. Their urinary concentrations provide an important clinical index of reproductive function.

Estrogens are concerned not only with reproductive behaviour but also with the general maintenance of the sexual organization of the female. When estradiol is administered to a mammal, the hormone becomes bound to uterine tissue, where it increases the rates of protein synthesis, of uptake of water and glucose, and, eventually, of growth of the lining epithelium and underlying muscular tissue (endometrium) of the uterus. Estradiol also evokes changes in the vagina, including hardening of the epithelium, a phenomenon that, in the laboratory rat, is used to determine its sexual condition. Estradiol and other estrogens have also been found in fishes and in other lower vertebrates. As with the corticosteroids, the sex hormones evolved very early in vertebrate history. Indeed, they have even been identified in invertebrates—in the eggs of the lobster, for example, and in the ovaries of starfishes, where, however, they may be no more than by-products of the metabolism of other steroids. Estrogenic activity is not necessarily restricted to steroids; for example, the estrogen mirestrol from a Thailand plant is not a steroid, nor is the very potent synthetic estrogen, stilbestrol, which is widely used in medicine.

Progestins. Progestins, of which the most important is progesterone, are concerned with the maintenance of pregnancy. Progesterone, therefore, evolved in viviparous mammals; *i.e.,* those that produce living young. Its chemical origin is demonstrable, since it is also an important intermediate compound in the biosynthetic pathways leading to corticoid and estrogen production. Mammals thus converted to hormonal use a substance that was synthesized by vertebrates long before the evolution of terrestrial vertebrates.

Some progesterone is probably formed in the ovarian follicle, but the main site of production is the corpus luteum, which is formed by a transformation of the follicle after ovulation; the secretory cells are formed from granulosa cells. The functions of the two important follicular phases, preceding and following ovulation, therefore, are continuous. The hormone is metabolized in several ways, but one important product is pregnanediol; formed mainly in the liver, it appears in part in the urine, where it can be measured to determine the degree of ovarian function.

Role of progesterone

The transformation of the follicle into the corpus luteum is an important turning point in the diphasic menstrual cycle of women and in the ovarian cycles of other mammals, from which the human cycle evolved. Progesterone prepares the uterus for the implantation of fertilized eggs, and it is also needed for the maintenance of pregnancy once implantation has taken place. It evokes a reduction in the ability of the uterine walls to contract, a proliferation of the glands of the endometrium, and the formation of glycogen. In addition, through its feedback action upon pituitary secretion, progesterone inhibits further ovulation

(see below), thus ensuring undisturbed fetal development. Ovulation in women occurs at about the middle of the monthly cycle, and the follicular phase is succeeded by the luteal phase. The vaginal bleeding at the end of the cycle is an indication that ovulation has not been followed by implantation of a fertilized egg and is immediately followed by the inception of a new cycle. If implantation does take place, the uterus provides metabolic support for the fetus until birth (see also REPRODUCTION AND REPRODUCTIVE SYSTEMS: *Menstruation*).

The vertebrate reproductive cycle depends upon delicate interrelationships between the sex hormones and the pituitary gonadotropic hormones (FSH and LH). As mentioned, it is uncertain whether or not there are two distinct gonadotropins in lower forms, but their separate action is well defined in mammals. Broadly speaking, FSH (follicle-stimulating hormone), with some support from LH (luteinizing hormone), promotes growth and secretory activity of the follicle. The increasing output of estrogen from the ovary eventually tends, by feedback to the pituitary gland, to reduce FSH output and to stimulate the secretion of LH; it is a sudden peak release of the latter hormone that evokes ovulation in many mammals. In others, such as the cat and rabbit, however, ovulation occurs as a response to the stimulus of copulation. Although some progesterone may be secreted by the granulosa cells of the follicle, the development of the corpus luteum greatly increases its secretion. Luteotropic activity in one form or another (the action of prolactin, for example, in the rat) is important in the early stage of this phase. Progesterone, in conjunction with the estrogen that is also being secreted (in part, probably, by the corpus luteum), suppresses further ovulation. This interaction of the two hormones is the basis of the design of contraceptive pills.

The corpus luteum continues to function during pregnancy, supplemented (in eutherian, or placental, mammals but not in marsupials) by endocrine secretions of the placenta (the organ through which contact between mother and fetus is maintained). The hormonal activity of the placenta varies with the species; in man, for example, the placenta secretes two gonadotropins called human chorionic gonadotropin (HCG) and human placental lactogen (HPL). HCG, like the pituitary gonadotropins, is a glycoprotein, with a molecular weight of 25,000 to 30,000. HPL is a protein, with a molecular weight variously estimated at about 19,000 or 30,000. One or perhaps both of these hormones, which become detectable during the early weeks of human pregnancy, probably stimulate luteal secretion. After two months the human placenta begins to manufacture estrogen and progestin; as a consequence, the corpus luteum is no longer needed for the maintenance of pregnancy. Much of the estrogen, although synthesized in the placenta, is derived from a compound (dehydroepiandrosterone) formed in the adrenal glands of the fetus. The placenta and the fetus thus form an integrated endocrine complex, a striking index of the high level of specialization found in the regulation of mammalian reproduction. (See also REPRODUCTION AND REPRODUCTIVE SYSTEMS: *Pregnancy*.)

The placenta probably secretes a luteotropin in all mammalian species, thereby contributing to prolongation of the life of the corpus luteum. In the mare and the monkey the placenta also secretes estrogen and progesterone, as in man, but in the mouse and rabbit it secretes only estrogen, and in the hamster and rat it secretes neither. In these last four species and in others like them, in which the placenta cannot substitute completely for the corpus luteum, ovariectomy (removal of the ovaries) of a pregnant female leads to the termination of pregnancy unless progesterone is administered to the female.

The interrelationships between the ovarian and the pituitary hormones are based upon negative feedback involving both the cells of the pituitary and those of the hypothalamus, which contains centres that are excited or inhibited by gonadotropin released from the pituitary gland. It is the hypothalamic involvement that enables vertebrate reproductive cycles to be adjusted by the central nervous system relative to external stimuli, particularly the seasonal fluctuations of daylight and other environmental factors that determine the onset of reproduction in many vertebrate species.

Male hormones. The sex hormones of the male follow a much simpler pattern than do those of the female, although the same principle of interaction exists between the pituitary gland and the gonads. The latter organs, the testes, secrete steriods called androgens, which are responsible for the maintenance of male characteristics and behaviour. FSH (follicle-stimulating hormone) from the pituitary gland stimulates the growth of the seminiferous tubules that constitute much of the structure of the testes and promotes within them the cell divisions that result in the production of mature sperm. LH (luteinizing hormone) from the pituitary gland promotes the development within the testes of endocrine tissue, which is composed of groups of cells (interstitial tissue) between the seminiferous tubules. The interstitial tissue of certain bony fishes, however, is represented by cells, called lobule boundary cells, situated within the tubule tissue.

Androgens

Under the influence of LH (often called ICSH, or interstitial-cell-stimulating hormone, in males), the interstitial tissue secretes the steroid hormone testosterone, which is the most important vertebrate androgen. The fact that it is an intermediate compound in the metabolic pathway of estrogen synthesis accounts for the origin of some forms of abnormal sexual organization in man; for example, the testes may secrete predominantly estrogen instead of androgen, resulting in markedly female appearance and behaviour in a male. Although testosterone may be secreted by the adrenal cortex, occasionally producing sexual disturbances, the amount of secretion is not normally significant. Testosterone, which is bound to a protein as it circulates in human blood, can be converted to the compound (androstenedione) from which it is formed, especially in the liver and in muscle; both compounds are metabolized, mainly in the liver, to substances that are excreted in urine. Very small quantities of testosterone can also be excreted in urine, and the quantities of testosterone and compounds derived from it frequently are measured to provide an index of testicular condition.

In addition to promoting male characteristics, male behaviour, and the maintenance of the spermatic tubules, testosterone, in the presence of normal amounts of growth hormone, also promotes growth of the bony skeleton. The reason for rapid growth at puberty is that the secretion of androgen markedly increases. The hormone brings about the closure of the epiphyses (ends) of the long bones, which completes the process of growth (estrogens have a similar action in the female). Thus, as often occurs among animals, growth ceases before full reproductive activity is attained, and competition between two processes, both of which make heavy demands upon the resources of the body, is avoided.

HORMONES OF THE DIGESTIVE SYSTEM

In vertebrates, the muscular and secretory activities of the alimentary canal and its associated glands are regulated by nervous and hormonal mechanisms. The hormones comprise a self-contained complex that functions at a relatively primitive level of organization and is distinguished by peculiar features; for example, specialized glandular tissues that secrete the hormones cannot be identified, although certain cells that can be seen in the wall of the alimentary canal are thought to be involved in their production. In addition, the digestive hormones regulate the system that produces them and are largely independent of the rest of the endocrine system, although certain relationships may not yet have been discovered.

Functions of digestive hormones

The functions of digestive hormones are best understood in mammals, in whom at least three are well characterized; the existence of others has been postulated. The three hormones—gastrin, secretin, and cholecystokinin/pancreozymin (CCK-PZ)—are polypeptide molecules whose amino-acid sequences are known. When food enters the stomach, the wall of its pyloric end (the area at which the stomach joins the small intestine) releases a hormone called gastrin, which promotes the flow of acid from the gastric glands in the stomach. These glands also release pepsinogen, which is the inactive form of the protein-

digesting enzyme pepsin, but this process is primarily under nervous control. The entry of the acidified stomach contents into the first part of the small intestine (duodenum) releases secretin and cholecystokinin/pancreozymin. Secretin evokes the discharge of fluid and bicarbonate ions from the pancreas (hydrelatic action) and promotes the secretion of bile from the liver (chloretic action). Cholecystokinin/pancreozymin, so-called because its two main actions were formerly attributed to two separate hormones, evokes the release of enzymes from the pancreas (ecbolic action) and causes contraction of the gallbladder (cystokinetic action), thereby promoting the entry of bile into the duodenum.

Little is known regarding hormonal control of alimentary activities in lower vertebrates; however, hydrelatic, ecbolic, and cystokinetic activities are present in preparations of the alimentary tracts of both agnathans and gnathostomes, indicating that substances able to regulate digestive activity appeared very early in the evolution of the vertebrate alimentary tract. Evidence suggests that the appearance of these hormones may have resulted in molecular diversification similar to examples previously discussed. The structure of the glucagon molecule from the pancreas, for example, is similar to that of secretin in that each molecule includes the same 15 amino acids located in the same positions. It has therefore been suggested that the two hormones may have evolved from a common ancestral molecule.

ENDOCRINE-LIKE GLANDS AND SECRETIONS

In addition to the well-defined hormones, other substances, which are found in blood and in tissues and are of uncertain function, may be concerned in various ways with physiological regulation in vertebrates, although their hormonal status has not yet been established.

Blood contains kinins, which are polypeptides that originate in the blood and perhaps elsewhere; bradykinin, for example, causes contraction of most smooth muscles and has a very potent action in dilating certain blood vessels. Its function, which is not yet established, may be to regulate the rate of blood flow or to participate in the inflammatory response of an animal to injury.

Some endocrine-like glands are associated with organs. One example in mammals is the carotid bodies, which are found on the carotid arteries that supply blood to the head. The carotid glands are stimulated by a decrease in the oxygen content of the blood and are considered to be the source of a substance, the nature of which has not yet been established with certainty, that promotes the process of red-blood-cell formation (erythropoiesis).

The pineal organ is an endocrine-like body found in the brain of all vertebrates. In lower vertebrates, it contains sensory and supporting cells and functions as a light-sensitive organ; in higher vertebrates, beginning with amphibians, the pineal gland has secretory functions, and in mammals, it is exclusively a secretory organ, producing from an amino acid (tryptophan) the compound serotonin (5-hydroxytryptamine, or 5HT) and a derivative of serotonin called melatonin. Preparations of melatonin, when given to amphibians, stimulate the concentration of pigment granules in chromatophores, an effect comparable with that of intermedin (MSH) but much more powerful. The normal physiological function of melatonin in higher vertebrates has not yet been established, although involvement in the regulation of reproduction is suspected. Serotonin is widely distributed in animals, especially in the brain and alimentary tract of vertebrates; it may function as a neurohumor in the invertebrate mollusks, but its significance in other animals is not yet certain.

The thymus is essential for the normal development in mammals of the system responsible for immunological responses. Its removal in newborn mice results in a deficiency of one type of white blood cells (lymphocytes) and a consequent likelihood of early death from infection. Preparations of thymus glands from various species contain a protein component, called thymosin, that promotes the development of lymphocytes. Although thymosin is sometimes regarded as a possible thymus hormone, the evidence is not yet complete.

Urohypophysis and corpuscles of Stannius

The urohypophysis, an organ found only in elasmobranch and bony fishes, probably developed independently in each group. The neurosecretory cells comprising the urohypophysis are concentrated at the hind end of the spinal cord, where they are associated with a vascular plexus to form a neurohemal organ. The urohypophysis resembles the neurosecretory system of the hypothalamus and the neural lobe (see above *Neurohypophysis and the polypeptide hormones of the hypothalamus*), but its functions have not yet been established.

The corpuscles of Stannius, found only in bony fishes, are sac-like bodies in the kidney. Although they were once thought to be a form of adrenocortical tissue, they differ from it in embryological origin as well as in cytological characteristics; moreover, although the corpuscles of Stannius are capable of limited steroid biosynthesis, they cannot convert cholesterol into corticoids, a process that occurs in adrenocortical tissue. Evidence suggests that these corpuscles secrete some substance, as yet uncharacterized, which plays a part in maintaining ionic homeostasis, perhaps in conjunction with the corticoid hormones.

The hormones of invertebrates

Some form of endocrine regulation probably occurs in all invertebrates; in arthropods (as exemplified in insects and crustaceans) it attains a level of complexity similar to that of vertebrates.

HORMONES OF INSECTS

Insects secrete hormones from neurosecretory cells and also from endocrine glands. Important neurosecretory centres occur in the pars intercerebralis region of the brain. The several cell types found in these centres indicate that more than one hormone is produced there.

Neurohormones. One of the brain hormones is thoracotropic hormone. This is released from nerve endings located in a neurohemal organ called the corpus cardiacum; the relationship between the corpus cardiacum and the brain closely parallels that between the neural lobe of the pituitary gland and the hypothalamic region of the brain of vertebrates. The thoractropic hormone is transferred in the blood to the thoracic glands in the body region called the thorax. It stimulates the production and release from the glands of ecdysone, a hormone that initiates molting, which is the periodic shedding of the outer skeleton that typically occurs in insects and other arthropods. The thoracotropic hormone is probably a polypeptide molecule.

A neurosecretion of the insect brain distinct from the thoracotropic hormone and called bursicon acts directly on the adult cuticle (skin) of arthropods to stimulate darkening and hardening processes. Bursicon is almost certainly a polypeptide, with a molecular weight of about 40,000. The brain of insects also produces a third neurohormone, which has a hyperglycemic (increase in level of blood glucose) effect in a tissue mass called the fat body, and a fourth neurohormone, which acts on the malpighian tubules (excretory organs) and rectum to facilitate the removal of excess fluid taken in with food; these two hormones, which may actually be one hormone with two effects, are also probably polypeptides. Another more active hyperglycemic hormone is formed within the corpora cardiaca, perhaps under the control of cerebral neurosecretion.

Molting hormones. Ecdysone is a steroid compound derived from cholesterol. Two forms are found in insects—α-ecdysone and β-ecdysone; ecdysones of unknown biological significance are also present in plants. Unlike vertebrates, insects cannot synthesize cholesterol, and they thus must obtain it from their food. Evidence concerning the mode of action of ecdysone indicates that it has a direct action upon the synthesis of the ribonucleic acid (RNA) that controls protein synthesis in the cell.

Juvenile hormone

The distinction in insects between molts that occur within the larval stage of development and those that result in the transformation of larvae to other stages (pupae, adults) in the life cycle is controlled by another hormone, called juvenile hormone, which is secreted in epithelial glands, called the corpora allata, near the brain. The hormone

controls the appearance of juvenile characters in larval stages, presumably by suppressing the activity of genes concerned with the expression of adult characters; reduction in the amount of or absence of the hormone at later molts results in the appearance of mature characters. The hormone nevertheless may continue to function in adults and often is necessary for normal egg production in females. Juvenile hormone is a lipoidal (fatlike) compound of similar structure from all sources. Many synthetic compounds mimic its effect, as do certain natural products—*e.g.*, substances in the balsam fir tree (*Abies*). Substances that mimic the action of juvenile hormone sometimes are used as insecticides, for if they are present in abnormal amounts in the later stages of the life cycle, they kill the insects.

Pheromones. Pheromones are important as insect sex attractants and as regulators of the social organization of social insects; *e.g.*, bees. The sex attractant of the female silk moth (*Bombyx mori*) is called bombykol. A related compound, gyptol, is the sex attractant of the female gypsy moth (*Porthetria dispar*), and gyplure is a synthetic compound that acts as an even more powerful attractant. The use of these compounds in the chemical control of insect pests is probably more promising than is the use of juvenile hormone.

The odour of the sex attractant of the honeybee (*Apis mellifera*), 9-oxodecenoic acid, stimulates the olfactory receptors of the drones (males). Secreted by the queen bee in the hive, the pheromone inhibits the development of the ovaries of the worker bees (sterile females) but is entirely effective only when it acts in conjunction with another inhibitory pheromone, 9-hydroxydecenoic acid. Removal of the queen from the hive results in the building of new queen cells by the workers and the development of functional ovaries in the drones. The mechanism by which these inhibitory substances function is not yet understood; some effect upon the nervous system presumably is involved.

HORMONES OF CRUSTACEANS

The endocrine systems of crustaceans resemble those of insects; important differences occur, however, implying extensive independent evolution in the two groups. The main sources of neurohormones are groups of cells (the X-organs) located in the optic ganglia of the eyestalks; the most important neurohemal organ is the sinus gland beside the eyestalks. Less important neurosecretory centres and neurohemal organs also occur. The pericardial organ of decapods, for example, is a group of nerve fibres and nerve endings in the walls of the pericardium, which encloses the heart; the pericardial organ secretes a substance, perhaps a polypeptide neurohormone, that accelerates the heartbeat.

In crustaceans, neurosecretory control is exerted over many functions, including the movement of pigment in the chromatophores, which determine body colour, and in the retina of the compound eye. Neurosecretions also regulate molting and the metabolic functions associated with it by actions exerted upon the so-called Y-organ in the head; this organ so closely resembles the thoracic gland of insects that the two may share a common ancestry. In crustaceans, however, the neurosecretion inhibits secretions from the Y-organ, and the molt is initiated by the withdrawal of the inhibitory hormone (in insects, the thoracotropic hormone from the corpus cardiacum stimulates the secretion of the molting hormone, ecdysone, from the thoracic gland). Neurosecretory hormones of crustaceans have diverse chemical and biological characteristics but apparently are polypeptides, as are the neurosecretory hormones of vertebrates.

Androgenic gland

Unlike insects, crustaceans have an androgenic gland, which typically is located on the genital duct (vas deferens) of the male. The androgenic gland secretes a hormone, possibly steroid in nature, that controls both the differentiation of the gonad of the male into a testis and the male characteristics of its limbs. The absence of the androgenic gland in the female results in the formation of an ovary, which subsequently synthesizes one or more hormones that, in female amphipods, promote the development of brood chambers (in which the young are hatched) and other structures associated with reproduction.

OTHER INVERTEBRATE HORMONES

The characterization of the hormones of other invertebrates awaits further study. Evidence indicates that the brain of polychaete worms produces neurosecretions that regulate growth and reproduction; in *Nereis* and *Nephtys* the neurosecretory fibres apparently have a close and presumably functional relationship with an epithelial gland (infracerebral organ), which is formed from coelomic epithelium and is situated on the wall of the brain.

Neurosecretory cells probably are present in mollusks such as gastropods and lamellibranchs. Experimental studies indicate an endocrine relationship in gastropods between the gonad (ovotestis) and possible neurosecretory cells in the tentacles and the brain; one ganglion of the gastropod *Lymnaea* may secrete a neurohormone with a diuretic (urine producing) action. Epithelial glands in mollusks are important; in the cephalopods, which are the most advanced invertebrates in some respects, optic glands on the optic stalks (eyestalks) secrete a hormone that promotes development and maturation of the gonads. In immature cephalopods the activity of the glands is inhibited by the central nervous system, apparently by a chemical mediator that diffuses from nerve fibres.

The nerve net, which constitutes the very primitive nervous system of the coelenterates, probably the most primitive multicellular animals, apparently contains neurosecretory cells; indirect but convincing evidence suggests that the cells release a secretion that promotes growth and inhibits sexual reproduction.

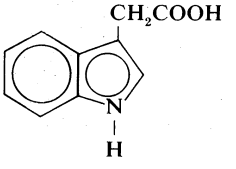
The hormones of plants

Growth in plants is regulated by four categories of phytohormones—auxins, gibberellins, cytokinins, and inhibitors.

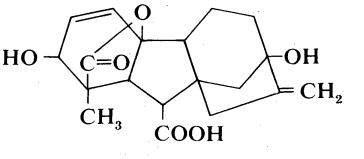
GROWTH PROMOTERS

Auxins. The distribution of auxins, which promote the lengthwise growth of plants, is correlated with the distribution of the growth regions of the plant. The most important auxin, whose structure is represented below, is β-indolylacetic acid (IAA), which is formed either from the amino acid tryptophan or from the breakdown of carbohydrates known as glycosides. The hormone affects plants by its action on chemical bonds of carbohydrates comprising plant cell walls. The process permits the cells to be irreversibly deformed and is accompanied by the entry of water and the synthesis of new cell-wall material. Many animal hormones may exert their effects by influencing protein synthesis, and evidence suggests that auxins may act in a similar way.

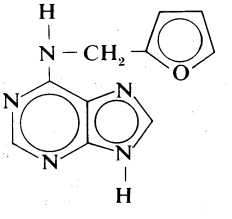
Many other naturally occurring and synthetic compounds



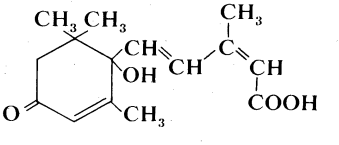
β-indolylacetic acid



gibberellic acid



kinetin (6-furfurylaminopurine)



abscisic acid

Figure 21: The structures of plant hormones.

called auxins also have growth-promoting properties, but they are not always as active as IAA. Some of these compounds, however, resist the enzymatic destruction that is the normal fate of IAA within the plant; this feature is of great value in research and in horticulture, because auxin action can be prolonged. Other auxin-like compounds are used as selective weed killers (*e.g.,* to disturb the leaf growth of dicotyledonous plants either in fields containing monocotyledonous cereal crops or on lawns) and as agents that remove leaves from dicotyledonous plants (defoliating agents).

Effects of indolylacetic acid

The hormonal characteristics of IAA are readily demonstrated in grass seedlings, in which the hormone is synthesized at the tip of the coleoptile (the protective sheath of the emerging plumule, or embryonic bud) and passes downward to its point of action in the growing region, where it evokes elongation of the coleoptile cells; growth stops if the tip is removed. The movement of the hormone downward from the tip of the coleoptile depends upon an interaction between the hormone and the cells through which the movement normally takes place.

In addition to promoting normal growth in plant length, auxins influence the growth of stems toward the light (phototropism) and against the force of gravity (geotropism). The phototropic response occurs because greater quantities of auxin are distributed to the side away from the light than to the side toward it; the geotropic response occurs because more auxin accumulates along the lower side of the coleoptile than along the upper side. The downward growth of roots is also associated with a greater quantity of auxin in their lower halves. This effect, which is the opposite to that found in coleoptiles, is attributed to an inhibitory action of auxins on root growth, but this aspect of auxin action is not yet fully understood. Auxins have actions other than those associated with promoting growth; *e.g.,* they play a role in cell division, in cell differentiation, in fruit development, in the formation of roots from cuttings, and in leaf fall (abscission). In experimental conditions, auxins tend to inhibit the progress of plant aging, perhaps because of their stimulating effect upon protein synthesis.

Gibberellins. Gibberellins are named after the fungus *Gibberella fujikuroi,* which produces excessive growth and poor yield in rice plants. One gibberellin is gibberellic acid (GA_3), which is present in higher plants as well as in fungi; many related compounds have structural variations that correlate with marked differences in effectiveness.

Gibberellins, abundant in seeds, are also formed in young leaves and in roots; they move upward from the roots in the xylem (woody tissue) and thus do not show the movement characteristic of auxins. Evidence suggests that gibberellins promote the growth of main stems, especially when applied to the whole plant. Unlike the auxins, gibberellins have little effect upon pieces of coleoptile in tissue culture. Gibberellins promote the growth of dwarf peas and are involved in the bolting (elongation) of rosette plants such as the carrot. Elongation of rosette plants occurs after exposure to certain environmental stimulation (*e.g.,* cold, or long periods of daylight), which is accompanied by an increase in the gibberellin content of the affected plant. In experimental conditions gibberellins tend, like auxins, to retard senescence.

Cytokinins. Cytokinins are compounds derived from a nitrogen-containing compound (adenine). One cytokinin is 6-furfurylaminopurine (kinetin); other compounds derived from adenine with effects similar to those of kinetin, and certain compounds derived from another nitrogen-containing compound, urea, are conveniently referred to as cytokinins, although not all are natural products. Cytokinins are synthesized in roots, from which, like the gibberellins, they move upward in the xylem and pass into the leaves and the fruit. Required for normal growth and differentiation, cytokinins act, in conjunction with auxins, to promote cell division and to retard senescence, which, at least in its early stages, is an organized phase of metabolism and not just a breakdown of tissue. An example of senescence is the yellowing of isolated leaves, which occurs as proteins are broken down and chlorophyll is destroyed. Cytokinins, which prevent yellowing by stabilizing the content of protein and chlorophyll in the leaf and the structure of chloroplasts, are used commercially in the storage of green vegetables.

Synthesis of cytokinins

GROWTH INHIBITORS

Growth inhibitors of various types have been identified in plants. The best characterized one is abscisic acid, which is chemically related to the cytokinins. It is probably universally distributed in higher plants and has a variety of actions; for example, it promotes abscission (leaf fall), the development of dormancy in buds, and the formation of potato tubers. The mode of action of abscisic acid has not yet been clarified but is thought to involve the direct inhibition of the synthesis of RNA and protein.

Another growth inhibitor is ethylene, which is a natural product of plants, formed possibly from linolenic acid (a fatty acid) or from methionine (an amino acid). Ethylene promotes abscission in senescent leaves, perhaps by facilitating the destruction of auxin. Its effects extend beyond that of inhibiting growth; in fruit, for example, ethylene is regarded as a ripening hormone. Involved in its action in fruit is another factor, perhaps auxin or another growth-regulating hormone, which influences the ethylene sensitivity of the tissues.

Hormonal interaction

The hormonal interaction conspicuous in animals is found also in plants; one example is the control of abscission, which requires the synthesis of enzymes at an abscission zone, at the base of the structure concerned, to catalyze reactions involving breakdown of cell walls. Auxin reaching the abscission zone from the tip of the structure promotes abscission; if auxin reaches the structure from the opposite direction, however, it tends to inhibit the process, probably by its influence on metabolism. Other hormones are also involved in abscission; ethylene stimulates the synthesis of the enzymes, and abscisic acid accelerates the associated senescence. Gibberellin tends to inhibit abscission by promoting growth.

Another example of hormonal interaction occurs during the germination of cereal seeds. The embryo (germ) is first activated by uptake of water, which enables it to produce gibberellin. Gibberellin acts on the living cells (aleurone layer) surrounding the food reserves (endosperm). This action induces the aleurone cells to produce enzymes that break down starch to sugars and release tryptophan from the protein of the endosperm. The tryptophan migrates to the coleoptile tip and is transformed into indolylacetic acid, which in turn moves to the growth zone and weakens the cell walls, thus permitting water uptake.

The target tissues probably play a role in such sequential actions, and it is likely that changes in their responsiveness to hormonal action, perhaps correlated with environmental stimuli, contribute to adaptive integration. The similarities in the hormonal mechanisms of plants and animals, two groups that are so profoundly different in their structure and mode of life, effectively illustrate the fundamental uniformity of biological organization.

(E.J.W.B.)

BIBLIOGRAPHY

Comprehensive works: ALBERT L. LEHNINGER, *Biochemistry* (1970); HENRY R. MAHLER and EUGENE H. CORDES, *Basic Biological Chemistry* (1968); ABRAHAM WHITE, PHILIP HANDLER, and EMIL L. SMITH, *Principles of Biochemistry,* 4th ed. (1968); and ALBERT L. LEHNINGER, *Principles of Biochemistry* (1982).

(Ed.)

Proteins: HANS NEURATH (ed.), *The Proteins,* 2nd ed., 5 vol. (1963–70), a comprehensive discussion of the structure, function, and biology of proteins; FELIX HAUROWITZ, *The Chemistry and Function of Proteins,* 2nd ed. (1963), a textbook for graduate students; M.F. PERUTZ, *Proteins and Nucleic Acids* (1963), deals with hemoglobin and the role of nucleic acids in protein biosynthesis; M.O. DAYHOFF (ed.), *Atlas of Protein Sequence and Structure* (1969, published in intervals of 1 or 2 years), describes 3-dimensional protein structure; H.E. SCHULTZE and J.F. HEREMANS, *Molecular Biology of Human Proteins,* 2 vol. (1966); L.P. CAWLEY, *Electrophoresis and Immunoelectrophoresis* (1969), a practical manual; E.M. SLAYTER, *Optical Methods in Biology* (1970); STANLEY BLACKBURN, *Amino Acid Determinations* (1968), a manual of method; and S.G. WALEY, *Mechanisms of Organic and Enzymic Reactions* (1962), a short text emphasizing enzymatic mechanisms. Supplementary material includes

Advances in Protein Chemistry (annual); *Annual Review of Biochemistry* (annual); *Amino Acids, Peptides and Proteins* (annual); A. NIEDERWIESER and G. PATAKI (eds.), *New Techniques in Amino Acid, Peptide and Protein Analysis* (1971); B. WEINSTEIN (ed.), *Chemistry and Biochemistry of Amino Acids, Peptides and Proteins* (1971); IUB, *Enzyme Nomenclature* (1965); and ALAN G. WALTON, *Polypeptides and Protein Structure* (1981).

(F.Ha./Ed.)

Carbohydrates: Advances in Carbohydrate Chemistry (published annually since 1945), multiple-authored volumes, reviewing specific areas of carbohydrate chemistry, scope now expanded to include chapters of more biological interest; F.J. BATES *et al., Polarimetry, Saccarimetry and the Sugars* (1942), practical information on sucrose and other common sugars together with methods for analysis and preparation of simple derivatives, primarily of historical interest although many of the data tables provide useful reference information; E.A. DAVIDSON, *Carbohydrate Chemistry* (1967), an intermediate college-level text emphasizing stereochemistry, conformation, and modern organic reaction mechanisms as applied to carbohydrates, and including both physical and chemical methods for structural determination; S.F. DYKE, "Chemistry of Natural Products Series," vol. 5, *The Carbohydrates* (1960), an introduction to basic reactions in the carbohydrate field, written almost in outline, schematic form; M. FLORKIN and E.H. STOTZ (eds.), *Comprehensive Biochemistry,* vol. 5, *Carbohydrates* (1963), a detailed volume covering all aspects of monosaccharide and many areas of polysaccharide structure and chemistry, although physical methods do not receive appropriate attention; R.D. GUTHRIE and J. HONEYMAN, *An Introduction to the Chemistry of Carbohydrates,* 3rd ed. (1968), a short monograph on basic carbohydrate structure and reactions; W.N. HAWORTH, *The Constitution of the Sugars* (1929), a classic description of the knowledge of sugar chemistry at that time, particularly Haworth's work in defining the ring structure of the carbohydrates—now primarily of historical importance; R.J. MCILROY, *Introduction to Carbohydrate Chemistry* (1967), a short, contemporary introduction to carbohydrate chemistry recognizing the importance of stereochemistry and conformational factors; F. MICHEEL, *Chemie der Zucker und Polysaccharide,* 2nd ed. (1956), a comprehensive work, in German, on carbohydrate chemistry; E.G.V. and E. PERCIVAL, *Structural Carbohydrate Chemistry,* 2nd ed. (1962), a standard work detailing methods of structural determination, particularly of sugar ring size and polysaccharide structure, still useful as a summary of structural methods with a heavy chemical emphasis although many modern techniques are not discussed; W. PIGMAN (ed.), *The Carbohydrates: Chemistry, Biochemistry, Physiology* (1957), a standard reference work that covers most of the classic work prior to 1957 although many of the modern developments, particularly in physical methods, stereochemistry, and reaction mechanisms are not included; M. STACEY and S.A. BARKER (eds.), *Carbohydrates of Living Tissues* (1962), a detailed account of the chemistry and biochemistry of polysaccharide substances, primarily those found in animal tissues; J. STANEK, M. CERNY, and J. PACAK (eds.), *The Monosaccharides* (1963); and *The Oligosaccharides* (1965), standard reference works translated from Czech, detailing structure, reaction, and properties of monosaccharides and oligosaccharides; R.L. WHISTLER and M.L. WOLFROM (eds.), *Methods in Carbohydrate Chemistry,* 5 vol. (1962–65), a necessary reference work for research workers in the field, detailing laboratory procedures for monosaccharide and polysaccharide preparations; and R.L. WHISTLER and C.L. SMART, *Polysaccharide Chemistry* (1953), a summary of chemical information on polysaccharides up to 1952, with emphasis on homopolymers such as starch and cellulose since few detailed structures were known at that time.

(E.A.D./Ed.)

Lipids. COMMISSION ON BIOCHEMICAL NOMENCLATURE, "The Nomenclature of Lipids," *Biochemistry,* 6: 3287–3292 (1967), nomenclature as approved by this international commission; E.J. MASORO, *Physiological Chemistry of Lipids in Mammals* (1968), a well-written introduction to lipid chemistry and metabolism in mammals; K.S. MARKLEY (ed.), *Fatty Acids,* 2nd rev. ed., pt. 1–3 (1960–64), an excellent and extensive discussion of fatty acid chemistry; L.L.M. VAN DENNEN, "Phospholipids and Biomembranes," *Progress in the Chemistry of Fats and Other Lipids,* vol. 8, pt. 1 (1965), detailed treatment of the chemistry and biochemistry of phospholipids as related to cell membranes; D. CHAPMAN, *Introduction to Lipids* (1969), a brief account of the chemistry of simple and complex lipids, with particular attention to fatty acids, glycerides, phosphoglycerides, and sphingolipids; E. TRIA and A.M. SCANU (eds.), *Structural and Functional Aspects of Lipoproteins in Living Systems* (1969), an excellent treatise describing and summarizing the status of the physical and chemical nature of lipid–protein complexes in naturally occurring systems; "Chromatography," *Progress in the Chemistry of Fats and Other Lipids,* vol. 8, pt. 3 (1969), three excellent articles on thin-layer, paper, and column chromatography; and *Advances in Lipid Research* (irreg.), collections of scholarly papers appearing approximately annually.

(D.J.H./Ed.)

Nucleic acids: E. CHARGAFF and J.N. DAVIDSON (eds.), *The Nucleic Acids: Chemistry and Biology,* 3 vol. (1955–60), a classical modern text, basic chemistry valid but biological sections out of date; E. CHARGAFF, *Essays on Nucleic Acids* (1963); D. COHEN, *The Biological Role of the Nucleic Acids* (1965), a short general monograph; J.N. DAVIDSON and W.E. COHN (eds.), *Progress in Nucleic Acid Research and Molecular Biology,* 11 vol. (1963–71), series of essays published irregularly; D.W. HUTCHINSON, *Nucleotides and Coenzymes* (1964), an elementary monograph; J.C. KENDREW, *The Thread of Life: An Introduction to Molecular Biology* (1966), a popular book based on a series of TV programs; A. KORNBERG, *Enzymatic Synthesis of DNA* (1962), three classical lectures; G. PARKER, W.A. REYNOLDS, and R. REYNOLDS, *DNA: The Key to Life* (1966), a programmed series of questions and answers; R.M.S. SMELLIE, *A Matter of Life: DNA* (1969), a popular paperback; and T.L.V. ULBRICHT, *Introduction to Nucleic Acids and Related Natural Products* (1966), an elementary monograph. College-level textbooks on this subject include: J.N. DAVIDSON, *The Biochemistry of the Nucleic Acids,* 6th ed. (1969); E. HARBERS, G.F. DOMAGK, and W. MULLER, *Introduction to Nucleic Acids: Chemistry, Biochemistry and Functions* (1968; orig. pub. in German, 1964); V.M. INGRAM, *The Biosynthesis of Macromolecules* (1965); A.M. MICHELSON, *The Chemistry of Nucleosides and Nucleotides* (1963); A.R. PEACOCKE and R.B. DRYSDALE, *The Molecular Basis of Heredity* (1965); STEPHEN NEIDLE (ed.), *Topics in Nucleic Acid Structure* (1981).

(J.N.D./R.Y.T./Ed.)

Vitamins: J.L. RODALE *et al., Complete Book of Vitamins* (1966), and J. MARKS, *Vitamins in Health and Disease* (1968), are among various popular works on vitamins. The following references are more advanced and require some background knowledge of biochemistry: G.H. BEATON and E.W. MCHENRY (ed.), *Nutrition: A Comprehensive Treatise,* vol. 2, *Vitamins, Nutrient Requirements and Food Selection* (1964), a discussion of the requirements and metabolism of vitamins and of deficiency diseases; S.A. KOSER, *Vitamin Requirements of Bacteria and Yeasts* (1968), on the requirements and metabolism of vitamins in microorganisms; F.A. ROBINSON, *The Vitamin Cofactors of Enzyme Systems* (1966), on the coenzyme function of vitamins; W.H. SEBRELL, JR. and R.S. HARRIS (eds.), vols. 1–5, P. GYORGY and W.N. PEARSON (eds.), vol. 6, 7, *The Vitamins,* 2nd ed. (1967), on the chemistry, biochemistry, and metabolism of vitamins, see esp. vol. 6 and 7 for a discussion of the methodology used in vitamin research; A.F. WAGNER and K. FOLKERS, *Vitamins and Coenzymes* (1964), a text with emphasis on molecular structure, organic synthesis and biosynthesis, chemical reactions, and metabolic roles of vitamins; ROMAN J. KUTSKY, *Handbook of Vitamins, Minerals, and Hormones,* 2nd ed. (1981).

(M.J.B.)

Hormones: E.J.W. BARRINGTON, *Hormones and Evolution* (1964), for students and general readers, concerned with the molecular structure and mode of action of hormones in relation to evolutionary theory; G.K. BENSON and J.G. PHILLIPS (eds.), *Hormones and the Environment* (1970), a wide-ranging review and research symposium, primarily for specialists but also of general interest; W.R. BUTT, *Hormone Chemistry* (1967), a correlation of research on the major mammalian hormones, requiring some knowledge of chemistry; R.E. COUPLAND, *The Natural History of the Chromaffin Cell* (1965), an integrated treatment of research on structure and function, considered at all levels of analysis, from the gross anatomical to the molecular; J. EBLING and K.C. HIGHNAM, *Chemical Communication* (1969), a concise elementary introduction for high school and first-year university students; B.E. FRYE, *Hormonal Control in Vertebrates* (1967), an elementary introduction, with emphasis on general principles and physiological adaptation; M. GABE, *Neurosecretion* (1966), an authoritative and comprehensive treatment for advanced students; G.W. HARRIS, *Neural Control of the Pituitary Gland* (1955), an authoritative monograph for students and research workers; M. PICKFORD, *The Central Role of Hormones* (1969), an elementary treatment for students and general readers with some knowledge of vertebrate biology; and C.T. SAWIN, *The Hormones: Endocrine Physiology* (1969), a clear treatment of hormone action for university students.

(E.J.W.B./Ed.)

The Biological Sciences

Biology may be defined as an area of learning that deals with all of the physicochemical aspects of life. But as a result of the modern tendency to unify scientific knowledge and investigation, there has been an overlapping of the field of biology with other scientific disciplines. Modern principles of other sciences—chemistry and physics, for example—are integrated with those of biology in such areas as biochemistry and biophysics.

Because biology is such a broad subject, it is subdivided into separate branches for convenience of study. Despite apparent differences, all the subdivisions are interrelated by basic principles. Thus, though it was once the custom to separate the study of plants (botany) from that of animals (zoology), and the study of the structure of organisms (morphology) from that of function (physiology), the current practice is to investigate those biological phenomena that all living things have in common.

Biology is often approached today on the basis of levels that deal with fundamental units of life. At the level of molecular biology, for example, life is regarded as a manifestation of chemical and energy transformations that occur among the many chemical constituents that comprise an organism. As a result of the development of more powerful and precise laboratory instruments and techniques, it is now possible to understand and define more exactly not only the invisible ultimate physiochemical organization (ultrastructure) of the molecules in living matter but also how living matter reproduces at the molecular level.

Cell biology, the study of the fundamental unit of structure and function in a living organism, may be said to have begun in the 17th century, with the invention of the compound microscope. Before that, the individual organism was studied as a whole (organismic biology), an area of research still regarded as an important level of biological organization. Population biology deals with groups or populations of organisms that inhabit a given area or region. Included at this level are studies of the roles that specific kinds of plants and animals play in the complex and self-perpetuating interrelationships that exist between the living and nonliving world, as well as studies of the built-in controls that maintain these relationships naturally.

These broadly based levels may be further subdivided into such specializations as morphology, taxonomy, biophysics, biochemistry, genetics, eugenics, and ecology.

In another way of classification, a field of biology may be especially concerned with the investigation of one kind of living thing—*e.g.*, botany, the study of plants; zoology, the study of animals; ornithology, the study of birds; ichthyology, the study of fishes; mycology, the study of fungi; microbiology, the study of microorganisms; protozoology, the study of one-celled animals; herpetology, the study of amphibians and reptiles; entomology, the study of insects; and physical anthropology, the study of man. For a list of both *Macropædia* and *Micropædia* articles on the subject areas, see *Propædia*: Part Three.

The article is divided into the following sections:

Basic concepts of biology

BIOLOGICAL PRINCIPLES

Homeostasis. The concept of homeostasis—*i.e.,* that all living things maintain a constant internal environment—was first suggested by Claude Bernard, a 19th-century French physiologist, who stated that "all the vital mechanisms, varied as they are, have only one object: that of preserving constant the conditions of life."

As originally conceived by Bernard, homeostasis applied to the struggle of a single organism to survive. The concept was later extended to include any biological system from the cell to the entire biosphere, all the areas of the Earth inhabited by living things.

Unity. All living organisms, regardless of their uniqueness, have certain biological, chemical, and physical characteristics in common. All, for example, are composed of the same basic units, or cells, and the same chemical substances, which, when analyzed, exhibit noteworthy similarities, even in such disparate organisms as bacteria and man. Furthermore, since the action of any organism is determined by the manner in which its cells interact and since all cells interact in much the same way, the basic functioning of all organisms is also similar.

There is not only unity of basic living substance and functioning but also unity of origin of all living things. According to a theory proposed in 1855 by Rudolf Virchow, a German pathologist, "all living cells arise from pre-existing living cells." This theory appears to be true for all living things at the present time under existing environmental conditions. If, however, life originated more than once in the past, the fact that all organisms have a sameness of basic structure, composition, and function would seem to indicate that only one original type succeeded.

Similarities between man and slime molds

A common origin of life would explain why in man or slime mold—and in all forms of life in between—the same chemical substance, deoxyribonucleic acid (DNA), in the form of genes accounts for the ability of all living matter to replicate itself exactly and to transmit genetic information from parent to offspring. Furthermore, the mechanisms for this transmittal follow a pattern that is the same in all organisms.

Whenever a change in a gene (a mutation) occurs, there is a change of some kind in the organism that contains the gene. It is this universal phenomenon that gives rise to the differences (variations) in populations of organisms from which nature selects for survival those that are best able to cope with changing conditions in the environment.

Evolution. In his theory of natural selection, which is discussed in greater detail later, Charles Darwin suggested that "survival of the fittest" was the basis for organic evolution (the modification of living things with time). Evolution itself is a biological phenomenon common to all living things, even though it has led to their differences. Evidence to support the theory of evolution has come primarily from the fossil record, from comparative studies of structure and function, and from studies of embryological development.

Diversity. Despite the basic biological, chemical, and physical similarities found in all living things, a diversity of life exists not only among and between species but also within every natural population. The phenomenon of diversity has had a long history of study because so many of the variations that exist in nature are visible to the eye. The fact that organisms changed during prehistoric times and that new variations are constantly evolving can be verified by paleontological records as well as by breeding experiments in the laboratory. Long after Darwin had assumed that variations existed, biologists discovered that they are caused by a change in the genetic material (DNA). This change can be a slight alteration in the sequence of the constituents of DNA (nucleotides), a larger change such as a structural alteration of a chromosome, or a complete change in the number of chromosomes. In any case, a change in the genetic material in the reproductive cells manifests itself as some kind of structural or chemical change in the offspring. The consequence of such a mutation depends upon the interaction of the mutant offspring with its environment.

Adjusting to changing conditions

It has been suggested that sexual reproduction became the dominant type of reproduction among organisms because of its inherent advantage of variability, which is the mechanism that enables a species to adjust to changing conditions. New variations are potentially present in genetic differences, but how preponderant a variation becomes in a gene pool depends upon the number of offspring the mutants or variants produce (differential reproduction). It is possible for a genetic novelty (new variation) to spread in time to all members of a population, especially if the novelty enhances the population's chances for survival in the environment in which it exists. Thus, when a species is introduced into a new habitat, it either adapts to the change by natural selection or by some other evolutionary mechanism or else it eventually dies off. Because each new habitat means new adaptations, habitat changes have been responsible for the millions of different kinds of species and for the heterogeneity within each species.

The total number of animal and plant species is estimated at between 2,000,000 and 4,500,000; authoritative estimates of the number of extinct species range from 15,000,000 up to 16,000,000,000. Although the use of classification as a means of producing some kind of order out of this staggering number of different types of organisms appears as early as the book of Genesis—with references to cattle, beasts, fowl, creeping things, trees, etc.—the first scientific attempt at classification is attributed to the Greek philosopher Aristotle, who tried to establish a system that would indicate the relationship of all things to each other. He arranged everything along a scale, or "ladder of nature," with nonliving things at the bottom; plants were placed below animals, and man was at the top. Other schemes that have been used for grouping species include large anatomical similarities, such as wings or fins, which indicate a natural relationship, and also similarities in reproductive structures.

At the present time taxonomy is based on two major assumptions: one is that similar body construction can be used as a criterion for a classification grouping; the other that, in addition to structural similarities, evolutionary and molecular relationships between organisms can be used as a means for determining classification.

Behaviour and interrelationships. As was mentioned earlier, the study of the relationships of living things to each other and to their environment is known as ecology. Because these interrelationships are so important to the welfare of Earth and because they can be seriously disrupted by man's activities, ecology is becoming one of the most important branches of biology.

Continuity. Whether an organism is man or a bacterium, its ability to reproduce is one of the most important characteristics of life. Because life comes only from preexisting life, it is only through reproduction that successive generations can carry on the properties of a species.

THE STUDY OF STRUCTURE

Living things are defined in terms of the activities or functions that are missing in nonliving things. The life processes of every organism are carried out by specific materials assembled in definite structures. Thus, a living thing can be defined as a system, or structure, that reproduces, changes with its environment over a period of time, and maintains its individuality by constant and continuous metabolism. This pattern of action or function results from and occurs in a pattern of organization.

Cells and their constituents. Knowledge of the structure and function of the cell has resulted from technological developments and methods.

Biologists once depended on the light microscope to study the morphology of cells found in higher plants and animals. The functioning of cells in unicellular and in multicellular organisms was then postulated from observation of the structure; the discovery of the chloroplastids in the cell, for example, led to the investigation of the process of photosynthesis. With the invention of the electron microscope, the fine organization of the plastids could be utilized for further quantitative studies of the different parts of this process.

Quantitative studies make use of histochemistry to identi-

fy proteins, carbohydrates, and other chemical constituents of cells. Histochemistry has also been used to identify RNA and DNA in various cell parts.

Tracing the movement of substances

A valuable method useful in tracing the movement of substances in living matter is radioautography: when radioactive nutrients, which can be incorporated into cells, are injected into animals, they give off detectable rays by which their presence and location can be determined. Thymidine, for example, can be made radioactive and, when injected, becomes part of the DNA being synthesized in the nucleus before cell division; the nuclei then can be identified by their radioactivity and the process of the origin of new DNA studied. Radioautography has been used to locate the site of protein synthesis and enzyme storage in cells.

Advanced technological developments—the microspectrophotometer, the X-ray probe, laser beam, computer, stereoscopic microscope, quartz-fibre microbalance, and television microscopy—are used to study the action of enzymes in living cells. The elucidation of such processes as lipid synthesis, active transport of large particles from the blood into cells, and continuous formation of taste cells has been dependent on similar instrumentation.

Tissues and organs. Early biologists viewed their work as a study of the organism. The organism, then considered the fundamental unit of life, is still the prime concern of some modern biologists, and the maintenance of organisms is still an important part of biological research.

In 1912 an experiment showed that cells can be kept alive indefinitely if proper conditions are maintained. Utilizing stringent laboratory techniques, workers have kept bits of chicken heart tissue alive for more than 30 years. Techniques for keeping organs alive in preparation for transplants stem from such experiments.

Modern biological research deals with the study of structure and function at all levels of biological organization from the molecule to the organism. Electronics, mathematics, and computers have become increasingly important in solving problems at all of these levels.

THE STUDY OF FUNCTION

To maintain life, an organism not only repairs or replaces (or both) its structures by a constant supply of the materials of which it is composed but also keeps its life processes in operation by a steady supply of energy. The initial source of this energy is the environment outside of the organism. The process by which the organism provides the necessary raw materials for the continuation of life is called nutrition. Plants obtain their nutrients from water, from minerals, and from the carbohydrates they manufacture. Animals, which cannot manufacture their own food, need at least the following kinds of nutrients: water, minerals, organic carbon, organic nitrogen, vitamins, certain amino acids, and fatty acids.

Many experiments have been directed toward solving the problem of biological differentiation. It has been determined that, although all genes of an organism are present in every cell, they do not all act at the same time: some genes act only at certain times during development; others never act in some cells. Whether a gene is active is sometimes the result of an interaction between cells. Cells seem to develop differently in different locations. How this is controlled is not definitely known; one possibility is the presence of an electrical communication between cells or of a substance that diffuses out of the cell. The latter idea is suggested by experiments demonstrating that the formation of the tissues of organs such as the eye, kidney, and liver are directly influenced by the tissues bordering them. Many of these experiments make use of tissue culture techniques, which permit the growth of cells outside of the body. It is possible to grow a single embryonic muscle cell into a colony of differentiated muscle. It is through such experiments that the questions about development and its implications may eventually be answered. (E.R.G.)

The history of biology

There are moments in the history of all sciences when remarkable progress is made in relatively short periods of time. Such leaps in knowledge result in great part from two factors: one is the presence of a creative mind—a mind sufficiently perceptive and original to discard hitherto accepted ideas and formulate new hypotheses; the second is the technological ability to test the hypotheses by appropriate experiments. The most original and inquiring mind is severely limited without the proper tools to conduct an investigation; conversely, the most sophisticated technological equipment cannot of itself yield insights into any scientific process.

Creative minds and technology

An example of the relationship between these two factors was the discovery of the cell. For hundreds of years there had been speculation concerning the basic structure of both plants and animals. Not until optical instruments were sufficiently developed to reveal cells, however, was it possible to formulate a general hypothesis, the cell theory, that satisfactorily explained how plants and animals are organized. Similarly, the significance of Gregor Mendel's studies on the mode of inheritance in the garden pea remained neglected for many years, until technological advances made possible the discovery of the chromosomes and the part they play in cell division and heredity. Moreover, as a result of the relatively recent development of extremely sophisticated instruments, such as the electron microscope and the ultracentrifuge, biology has moved from being a largely descriptive science—one concerned with entire cells and organisms—to a discipline that increasingly emphasizes the subcellular and molecular aspects of organisms and attempts to equate structure with function at all levels of biological organization.

THE EARLY HERITAGE

Although it is not known when the study of biology originated, early man must have had some knowledge of the animals and plants around him. His very survival depended upon the accurate recognition of nonpoisonous food plants and upon an understanding of the habits of dangerous predators. Archaeological records indicate that even before the development of civilization, man had domesticated virtually all the amenable animals available to him and had developed an agricultural system sufficiently stable and efficient to satisfy the needs of large numbers of people living together in communities. It is clear, therefore, that much of the history of biology predates the time at which man began to write and to keep records.

Earliest biological records. *Biological practices among Assyrians and Babylonians.* Much of the earliest recorded history of biology is derived from bas-reliefs the Assyrians and Babylonians made of their cultivated plants and from carvings depicting their veterinary medicine. Illustrations on certain seals reveal that the Babylonians had learned that the date palm reproduces sexually and that pollen could be taken from the male plant and used to fertilize female plants. Although a precise dating of these early records is lacking, a Babylonian business contract of the Hammurabi period (*c.* 1800 BC) mentions the male flower of the date palm as an article of commerce, and descriptions of date harvesting date back to about 3500 BC.

Another source of information concerning the extent of biological knowledge of these early peoples was the discovery of several papyri that pertain to medical subjects; one, believed to date back to 1600 BC, contains anatomical descriptions; another (*c.* 1500 BC) indicates that the importance of the heart had been recognized. Because these ancient documents, which contained mixtures of fact and superstition, probably summarized then-current knowledge, it may be assumed that some of their contents had been known by earlier generations.

Biological knowledge of Egyptians, Chinese, and Indians. Papyri and artifacts found in tombs and pyramids indicate that the Egyptians also possessed considerable medical knowledge. Their well-preserved mummies demonstrate that they had a thorough understanding of the preservative properties of herbs required for embalming; plant necklaces and bas-reliefs from various sources also reveal that the ancient Egyptians were well aware of the medicinal value of certain plants 2,000 years before Christ. Even earlier (*c.* 2800 BC), a work now ascribed to the Chinese emperor Shen Nung described the therapeutic powers of

Early use of medicinal plants

numerous medicinal plants and included descriptions of many important food plants, such as the soybean. Furthermore, the ancient Chinese not only utilized the silkworm *Bombyx mori* to produce silk for commerce but also understood the principle of biological control, employing one type of insect, an entomophagous (insect-eating) ant, to destroy insects that bored into trees.

As early as 2500 BC the people of northwestern India had a well-developed science of agriculture. The ruins at Mohenjodaro have yielded seeds of wheat and barley that were cultivated at this time. Millet, dates, melons, and other fruits and vegetables, as well as cotton, were known to this civilization. Plants were not only a source of food, however. A document, believed to date back to the 6th century BC, described the use of about 960 medicinal plants and included information on such topics as anatomy, physiology, pathology, and obstetrics.

The Greco-Roman world. Although the Babylonians, Assyrians, Egyptians, Chinese, and Indians amassed much biological information, they lived in a world believed to be dominated by unpredictable demons and spirits. Hence, learned men in these early cultures directed their studies toward an understanding of the supernatural, rather than the natural, world. Anatomists, for example, dissected animals not to gain an understanding of their structure but to study their organs in order to predict the future. With the emergence of the Greek civilization, however, these mystical attitudes began to change. Around 600 BC there arose a school of Greek philosophers who believed that every event has a cause and that a particular cause produces a particular effect. This concept, known as causality, had a profound effect on subsequent scientific investigation. Furthermore, these philosophers assumed the existence of a "natural law" that governs the universe and can be comprehended by man through the use of his powers of observation and deduction. Although they established the science of biology, the greatest contribution the Greeks made to science was the idea of rational thought.

Theories about man and the origin of life. One of the earliest Greek philosophers, Thales of Miletus (*c.* 7th century BC), maintained that the universe contained a creative force that he called physis, an early progenitor of the term physics; he also postulated that the world and all living things in it were made from water. Anaximander, a student of Thales, did not accept water as the only substance from which living things were derived; he believed that in addition to water, living things consisted of earth and a gaslike substance called *apeiron,* which could be divided into hot and cold. Various mixtures of these materials gave rise to the four elements: earth, air, fire, and water.

Early evolutionary theory

Although he was one of the first to describe the Earth as a sphere rather than as a flat plane, Anaximander proposed that life arose spontaneously in mud and that the first animals to emerge had been fishes covered with a spiny skin. The descendants of these fishes eventually left water and moved to dry land, where they gave rise to other animals by transmutation (the conversion of one form into another). Thus, an early evolutionary theory was formulated.

At Crotone in southern Italy, where an important school of natural philosophy was established by Pythagoras about 500 BC, one of his students, Alcmaeon, investigated animal structure and described the difference between arteries and veins, discovered the optic nerve, and recognized the brain as the seat of the intellect. As a result of his studies of the development of the embryo, Alcmaeon may be considered the founder of embryology.

Although the Greek physician Hippocrates, who established a school of medicine on the Aegean island of Cos around 400 BC, was not an investigator in the sense of Alcmaeon, he did recognize through observations of patients the complex interrelationships involved in the human body. He also understood how the environment can influence human nature and suggested that sharply contrasting climates tend to produce a powerful type of inhabitant, while an even, temperate climate is conducive to indolence.

Hippocrates and his predecessors were all concerned with the central philosophical question of how the cosmos and its inhabitants were created. Although they accepted the physis as the creative force, they differed with regard to the importance of the roles played by earth, air, fire, water, and other elements. Although Anaximenes, for example, who may have been a student of Anaximander, adhered to the then-popular precept that life originated in a mass of mud, he postulated that the actual creative force was to be found in the air and that it was influenced by the heat of the Sun. Members of the Hippocratic school also believed that all living bodies were made up of four humours—blood, black bile, phlegm, and yellow bile—which supposedly originated in the heart, spleen, brain, and liver, respectively. An imbalance of the humours was thought to cause an individual to be sanguine, melancholy, phlegmatic, or choleric. The persistence of these words in current vocabulary attests to the lengthy popularity of the idea of humoral influences. For centuries it was also believed that an imbalance in the humours was the cause of disease, a belief that resulted in the common practice of bloodletting to get rid of excessive humours.

Aristotelian concepts. Around the middle of the 4th century BC, ancient Greek science reached a climax with Aristotle, who was interested in all branches of knowledge, including biology. Using his own observations and theories, Aristotle was the first to attempt a system of animal classification, in which he contrasted animals containing blood with those that were bloodless. The animals with blood included those now grouped as mammals (except the whales, which he placed in a separate group), birds, amphibians, reptiles, and fishes. The bloodless animals were divided into the cephalopods, the higher crustaceans, the insects, and the testaceans, the last group being a collection of all the lower animals. His careful examination of animals led to the understanding that mammals have lungs, breathe air, are warm-blooded, and suckle their young. Aristotle was the first to show any understanding of an overall systematic taxonomy and to recognize units of different degrees within the system.

Aristotle's classification of animals

The most important part of Aristotle's work was that devoted to reproduction and the related subjects of heredity and descent. He identified four means of reproduction, including the abiogenetic origin of life from nonliving mud, a belief held by Greeks of that time. Other modes of reproduction recognized by him included budding (asexual reproduction), sexual reproduction without copulation, and sexual reproduction with copulation. Aristotle described sperm and ova and believed that the menstrual blood of viviparous organisms (those that give birth to living young) was the actual generative substance.

Although Aristotle recognized that species are not stable and unalterable and although he attempted to classify the animals he observed, he was far from developing any pre-Darwinian ideas concerning evolution. In fact, he rejected any suggestion of natural selection and sought teleological explanations (*i.e.,* all phenomena in nature are shaped by a purpose) for any given observation. Nevertheless, many important scientific principles, some of which are often thought of as 20th-century concepts, can be ascribed to Aristotle. The following are a few such: (1) Using birds as an example, he formulated the principle that all organisms are structurally and functionally adapted to their habits and habitats. (2) Nature is parsimonious; it does not expend unnecessary energy. (3) In classifying animals, Aristotle rejected the idea of dividing them solely by their external structures (*e.g.,* animals with wings and those without wings). He recognized instead a basic unity of plan among diverse organisms, a principle that is still conceptually and scientifically sound. Further, Aristotle also believed that the entire living world could be described as a unified organization rather than as a collection of diverse groups. (4) By his observations, Aristotle realized the importance of structural homology, basically similar organs in different animals, and functional analogy, different structures that serve somewhat the same function—*e.g.,* the hand, claw, and hoof are analogous structures. These principles constitute the basis for the biological field of study known as comparative anatomy. (5) Aristotle's observations also led to the formulation of the principle that general structures appear before specialized ones and that tissues differentiate before organs.

Biological principles formulated by Aristotle

Botanical investigations. Of all the works of Aristotle that have survived, none deals with what was later differentiated as botany, although it is believed that he wrote at least two treatises on plants. Fortunately, however, the work of Theophrastus, one of Aristotle's students, has been preserved to represent plant science of the Greek period. Like Aristotle, Theophrastus was a keen observer, although his works do not express the depth of original thought exemplified by his teacher. In his great work, *De historia et causis plantarum* (*The Calendar of Flora,* 1761), in which the morphology, natural history, and therapeutic use of plants are described, Theophrastus distinguished between the external parts, which he called organs, and the internal parts, which he called tissues. This was an important achievement because Greek scientists of this period had no established scientific terminology by which a specific structure could be referred to with a scientific term. For this reason, both Aristotle and Theophrastus were obliged to write very long descriptions of structures that can be described rapidly and simply today. Because of this difficulty, Theophrastus sought to develop a scientific nomenclature by giving special meaning to words that were then in more or less current use; for example, *karpos* for fruit and *perikarpion* for seed vessel.

Development of scientific terminology

Although he did not propose an overall classification system for plants, over 500 of which are mentioned in his writings, Theophrastus did unite many species into what are now considered genera. In addition to writing the earliest detailed description of how to pollinate the date palm by hand and the first unambiguous account of sexual reproduction in flowering plants, he also recorded observations on seed germination and development.

Post-Grecian biological studies. With Aristotle and Theophrastus, the great Greek period of scientific investigation came to an end. The most famous of the new centres of learning were the library and museum in Alexandria. From 300 BC until around the time of Christ all significant biological advances were made by physicians at Alexandria. One of the most outstanding of these men was Herophilus, who dissected human bodies and compared their structures to those of other large mammals. He recognized the brain, which he described in detail, as the centre of the nervous system and the seat of intelligence. Based on his knowledge, he wrote a general anatomical treatise, a special one on the eyes, and a handbook for midwives.

Erasistratus, a younger contemporary and reputed rival of Herophilus who also worked at the museum in Alexandria, studied the valves of the heart and the circulation of blood. Although he was wrong in supposing that blood flows from the veins into the arteries, he was correct in assuming that small interconnecting vessels exist. He thus suspected (but did not see) the presence of capillaries; he thought, however, that the blood changed into air, or *pneuma,* when it reached the arteries, to be pumped throughout the body.

Perhaps the last of the ancient biological scientists of note was Galen of Pergamum, a Greek physician who practiced in Rome during the middle of the 2nd century AD. His early years were spent as a surgeon at the gladiatorial arena, which gave him the opportunity to observe details of human anatomy. But this was an age when it was considered improper to dissect human bodies, and, as a result, detailed study was not possible. Thus, though Galen's research on animals was thorough, his knowledge of human anatomy was faulty. Because his work was extensive and clearly written, Galen's writings, nevertheless, dominated medicine for centuries to come.

The Arab world and the European Middle Ages. After Galen there were no further biological investigations for many centuries. It is sometimes claimed that the rise of Christianity was the cause of the decline in science; this, however, is not a tenable viewpoint, for science was already virtually dead by the end of the 2nd century AD, a time when Christianity was still an obscure sect. It is true, however, that the rise of Christianity did not favour the questioning attitude of the Greeks.

The decline of science

Arab domination of biology. During the almost 1,000 years that science was dormant in Europe, the Arabs, who by the 9th century had extended their sphere of influence as far as Spain, became the custodians of science and dominated biology, as they did other disciplines. At the same time, as the result of a revival of learning in China, new technical inventions flowed from there to the West. The Chinese had discovered how to make paper and how to print from movable type, two achievements that were to have an inestimable effect upon learning. Another important advance that also occurred during this time was the introduction into Europe from India of the so-called Arabic numerals.

From the 3rd until the 11th century biology was essentially an Arab science. Although they themselves were not great innovators, they discovered the works of such men as Aristotle and Galen, translated them into Arabic, studied them, and wrote commentaries about them. Of the Arab biologists, al-Jāḥiẓ, who died about 868, is particularly noteworthy. Among his biological writings is *Kitāb al-ḥayawān* ("Book of Animals"), which, although revealing some Greek influence, is primarily an Arabic work. In it, the author emphasized the unity of nature and recognized relationships between different groups of organisms. Because al-Jāḥiẓ believed that the Earth contained both male and female elements, he found the Greek doctrine of spontaneous generation (life emerging from mud) to be quite reasonable.

Ibn Sīnā, or Avicenna as he is better known, was an outstanding Persian scientist around the beginning of the 11th century; he was the true successor to Aristotle. His writings on medicine and drugs, which were particularly authoritative and remained so until the Renaissance, did much to bring the works of Aristotle back to Europe, where they were translated into Latin from Arabic.

Development of botany and zoology. During the 12th century the growth of biology was sporadic. Nevertheless, it was during this time that botany was developed from the study of plants with healing properties; similarly, from veterinary medicine and the pleasures of the hunt came zoology. Because of the interest in medicinal plants, herbs in general began to be described and illustrated in a realistic manner. Although Arabic science was well developed during this period and was far in advance of Latin, Byzantine, and Chinese cultures, it began to show signs of decline. Latin learning, on the other hand, rapidly increasing, was best exemplified perhaps by a mid-13th-century German scholar, Albertus Magnus (Albert the Great), who was probably the greatest naturalist of the Middle Ages. His biological writings (*De vegetabilibus,* seven books, and *De animalibus,* 26 books) were based on the classical Greek authorities, predominantly Aristotle. But in spite of this classical basis, a significant amount of his work contained new observations and facts; for example, he described with great accuracy the leaf anatomy and venation of the plants he studied.

Albert was particularly interested in plant propagation and reproduction and discussed in some detail the sexuality of plants and animals. Like his Greek predecessors, he believed in spontaneous generation; he also believed that animals were more perfect than plants because they required two individuals for the sexual act. Perhaps one of Albert's greatest contributions to medieval biology was the denial of many superstitions believed by his contemporaries, a skepticism that, together with the reintroduction of Aristotelian biology, was to have profound effects on subsequent European science.

One of Albert's pupils was Thomas Aquinas, who endeavoured to reconcile Aristotelian philosophy and the teachings of the church. Because Aquinas was a rationalist, he declared that God created the reasoning mind; hence, by true intellectual processes of reasoning, man could not arrive at a conclusion that was in opposition to Christian thought. Acceptance of this philosophy made possible a revival of rational learning that was consistant with Christian belief.

Revival of rational learning

Revitalization of anatomy. Italy, during the Middle Ages, became the most active scientific centre, although its major interests were concentrated on agriculture and medicine. A development of particular significance at this time was the introduction of dissection into medical schools, a step that revitalized the study of anatomy. Be-

cause of what it reveals about medieval anatomy in general, the work of Mondino dei Liucci, the most famous of the Italian anatomists at the beginning of the 14th century, is particularly important. First, because there was no way of preserving cadavers, organs that spoiled quickly had to be dissected rapidly. Furthermore, it was the custom for the teacher to leave the actual dissection to an underling, who, not wishing to offend the teacher, agreed with all of his statements. Thus, although Mondino performed all of his own dissections and, from his observations, could have corrected the errors of the Greeks and Arabs, he did not choose to contradict any of the authorities. Even when the authorities contradicted themselves, Mondino sought to harmonize their views. Perhaps Mondino exemplifies the difficulty that was so characteristic of the era; namely, the problem of breaking away from established authority.

The Renaissance. *Resurgence of biology.* Beginning in Italy during the 14th century there was a general ferment within the culture itself, which, together with the rebirth of learning (partly as a result of the rediscovery of Greek work), is referred to as the Renaissance. Interestingly, it was the artists, rather than the professional anatomists, who were intent upon a true rendering of the bodies of animals and men and thus were motivated to gain their knowledge firsthand by dissection. No individual better exemplifies the Renaissance than Leonardo da Vinci, whose anatomical studies of the human form during the late 1400s and early 1500s were so far in advance of the age that they included details not recognized until a century later. Furthermore, while dissecting animals and examining their structure, Leonardo compared them to the structure of man. In doing so he was the first to indicate the homology between the arrangements of bones and joints in the leg of the human and that of the horse, despite the superficial differences. Homology was to become an important concept in uniting outwardly diverse groups of animals into distinct units, a factor that is of great significance in the study of evolution.

Leonardo's discovery of homologous structures

Other factors had a profound effect upon the course of biology in the 1500s, particularly the introduction of printing around the middle of the century, the increasing availability of paper, and the perfected art of the wood engraver, all of which meant that illustrations as well as letters could be transferred to paper. In addition, after the Turks had conquered Byzantium in 1453, many Greek scholars took refuge in the West; the scholars of the West thus had direct access to the scientific works of antiquity, rather than indirect access through Arabic translations.

Advances in botany. Otto Brunfels, the German theologian and botanist, published in 1530 a book about medicinal herbs, *Herbarum vivae eicones,* which, with its fresh and vigorous illustrations, contrasted sharply with earlier texts, whose authors had been content merely to copy from old manuscripts. In addition to books on the same subject, Hieronymus Bock (Latinized to Tragus) and Leonhard Fuchs also published around the mid-1500s descriptive, well-illustrated texts about common wild flowers. The books published by the three men, who are often referred to as the German fathers of botany, may be considered the forerunners of modern botanical floras (treatises on or lists of the plants of an area or period).

Throughout the 16th century, interest in botanical study also existed in such other countries as the Netherlands, Switzerland, Italy, and France. During this time there was a great improvement in the classification of plants, which had been described in ancient herbals merely as trees, shrubs, or plants and, in later books, were either listed alphabetically or arranged in some arbitrary grouping. The necessity for a systematic method to designate the increasing number of plants being described became obvious. Accordingly, using a binomial system very similar to modern biological nomenclature, Gaspard Bauhin, a Swiss botanist of the late 16th and early 17th centuries, designated plants by a generic and a specific name. Although affinities between plants were indicated by the use of common generic names, Bauhin did not speculate on their common kinship.

Bauhin's botanical nomenclature

Pierre Belon, a French naturalist who travelled extensively in the Middle East, where he studied the flora, illustrates the wide interest of the 16th-century biologists. Although his botanical work was limited to two volumes, one on trees and one on horticulture, his books on travel included numerous biological entries, and his two books on fishes reveal much about the current state of systematics, including not only fishes but also such other aquatic creatures as mammals, crustaceans, mollusks, and worms. In his *L'Histoire de la nature des oyseaux* ("Natural History of Birds"), however, in which Belon's taxonomy was remarkably similar to that being used today, he showed a clear grasp of comparative anatomy, particularly of the skeleton, publishing the first picture of a bird skeleton beside a human skeleton to point out the homologies. Numerous other European naturalists who travelled extensively also brought back accounts of exotic animals and plants, and most of them wrote voluminous records of their excursions. Two other factors contributed significantly to the development of botany at this time: first was the establishment of botanical gardens by the universities, as distinct from the earlier gardens that had been established for medicinal plants; second was the collection of dried botanical specimens, or herbaria.

It is perhaps surprising that the great developments in botany during the 16th century had no parallel in zoology. Instead, there arose a group of biologists known as the Encyclopedists, best represented by Conrad Gesner, a 16th-century Swiss naturalist, who compiled books on animals that were illustrated by some of the finest artists of the day (Albrecht Dürer, for example). But because the descriptions of many of the animals were grossly inaccurate, in many cases continuing the legends of the Greeks, apart from their aesthetic value the books did little to advance zoological knowledge.

Advances in anatomy. Like that of botany, the beginning of the scientific study of anatomy can be traced to a combination of humanistic learning, Renaissance art, and the craft of printing. Although Leonardo da Vinci initiated anatomical studies of human cadavers, his work was not known to his contemporaries. Rather, the appellation father of anatomy must be accorded to the Belgian anatomist Andreas Vesalius, who studied at the rather conservative schools in Louvain and Paris, where he became a successful teacher very familiar with Galen's work. As a result of disagreements with his superiors, however, Vesalius moved at the end of 1537 to Padua, where he became noted for far-reaching teaching reforms. Most important, Vesalius abolished the practice of having someone else do the actual dissection; instead, he dissected his own cadavers and lectured to students from his findings. His text, *De humani corporis fabrica libri septem* (1543; "Seven Books on the Structure of the Human Body"), was the first modern book on the subject of anatomy and, as such, constituted a foundation of great importance for biology. Perhaps Vesalius' greatest contribution, however, was that he inspired a group of younger scientists to be critical and to accept a description only after they had verified it. Thus, as anatomists became more questioning and critical of the works of others, the stranglehold of Galen was finally broken. Of Vesalius' successors, Michael Servetus, a Spanish theologian and physician, discovered the pulmonary circulation of the blood from the right chamber of the heart to the lungs and stated that the blood did not pass through the central septum (wall) of the heart, as had previously been believed.

The father of anatomy

ADVANCES TO THE 20TH CENTURY

Seventeenth-century advances in biology included the establishment of scientific societies for the dissemination of ideas and progress in the development of the microscope, through which man discovered a hitherto invisible world that had far-reaching effects on biology. Systematizing and classifying, however, dominated biology throughout much of the 17th and 18th centuries, and it was during this time that the importance of the comparative study of living organisms, including man, was realized. During the 18th century the long-held idea that living organisms could originate from nonliving matter (spontaneous generation) began to crumble, but it was not until after the mid-19th century that it was finally disproved by Louis Pas-

teur. Biological expeditions added to the growing body of knowledge of plant and animal forms and led to the 19th-century development of the theory of evolution. The 19th century was one of great progress in biology: in addition to the formulation of the theory of evolution, the cell theory was established, the foundations for modern embryology were laid, and the laws of heredity were discovered.

The discovery of the circulation of blood. William Harvey, an Englishman who studied at Padua with one of Vesalius' students, is credited with the discovery of the circulation of the blood. Prior to Harvey, the Aristotelian-Galenistic theory of circulation supposed that the blood sucked up by the heart during its expansion ebbed away during contraction; further, the theory also suggested that the blood flowed through pores between the two halves of the heart and that the heart produced a vital heat, which was tempered by the air from the lungs. In his own work, however, Harvey demonstrated that the heart expands passively and contracts actively. Also, by measuring the amount of blood flowing from the heart, he concluded that the body could not continuously produce that amount. Finally, he was able to show that blood was returned to the heart through the veins, postulating a connection (the capillaries) between the arteries and veins that was not to be discovered for another century. Harvey was also interested in embryology, to which he made a significant contribution by suggesting that there is a stage (the egg) in the development of all animals during which they are undifferentiated living masses. A biological dictum, *ex ovo omnia* ("everything comes from the egg"), is a summation of this concept.

Postulation of the existence of blood capillaries

The establishment of scientific societies. A development of great importance to science was the establishment in Europe of academies or societies; they consisted of small groups of men who met to discuss subjects of mutual interest. Although some of the groups enjoyed the financial patronage of princes and other wealthy members of society, the members' interest in science was the sole sustaining force. The academies also provided freedom of expression, which, together with the stimulus of exchanging ideas, contributed greatly to the development of scientific thought. One of the earliest of these organizations was the Italian Academy of the Lynx, founded in Rome around 1603. Galileo Galilei made a microscope for the society; another of its members, Johannes Faber, an entomologist, gave the instrument its name. Other academies in Europe included the French Academy of Science (founded in 1666), a German Academy in Leipzig, and a number of small academies in England that in 1662 became incorporated under royal charter as the Royal Society of London, an organization that was to have considerable influence on scientific developments in England.

In addition to providing a forum for the discussion of scientific matters, another important aspect of these societies was their publications. Before the advent of printing there were no convenient means for the wide dissemination of scientific knowledge and ideas; hence, scientists were not well informed about the works of others. To correct this deficiency in communications, the early academies initiated several publications, the first of which, *Journal des Savants,* was published in 1665 in France. Three months later, the Royal Society of London originated its *Philosophical Transactions.* At first this publication was devoted to reviews of work completed and in progress; later, however, the emphasis gradually changed to accounts of original investigations that maintained a high level of scientific quality. Gradually, specialized journals of science made their appearance, though not until at least another century had passed.

Early publications

The development of the microscope. The magnifying power of segments of glass spheres was known to the Assyrians before the time of Christ; during the 2nd century AD, Claudius Ptolemy, an astronomer, mathematician, and geographer at Alexandria, wrote a treatise on optics in which he discussed the phenomena of magnification and refraction as related to such lenses and to glass spheres filled with water. Despite this knowledge, however, glass lenses were not used extensively until around 1300, when some anonymous person invented spectacles for the improvement of vision. This invention aroused curiosity concerning the property of lenses to magnify, and in the 16th century several papers were written about such devices. Then, near the end of the 16th century, it was discovered that if certain lenses are mounted together in a tube, they form what physicists now call a Galilean telescope when viewed through one end, and a Galilean microscope when viewed through the other. When, in the early 1600s, Galileo used this instrument to examine the stars and planets, he was able to record such new discoveries as the rings of Saturn and the four satellites of Jupiter. Although Galileo is often credited with making the first biological observations with the microscope, he did not make any further contributions to its development.

Following subsequent technological improvements in the instrument and the development of a more liberal attitude toward scientific research, five microscopists emerged who were to have a profound affect on biology: Marcello Malpighi, Antonie van Leeuwenhoek, Jan Swammerdam, Nehemiah Grew, and Robert Hooke.

The classical microscopists

Malpighi's animal and plant studies. Marcello Malpighi, an Italian biologist and physician, conducted extensive studies in animal anatomy and histology (the microscopic study of the structure, composition, and function of tissues). He was the first to describe the inner (malpighian) layer of the skin, the papillae of the tongue, the outer part (cortex) of the cerebral area of the brain, and the red blood cells. He wrote a detailed monograph on the silkworm; a further major contribution was a description of the development of the chick, beginning with the 24-hour stage. In addition to these and other animal studies, Malpighi made detailed investigations in plant anatomy. He systematically described the various parts of plants, such as bark, stem, roots, and seeds, and discussed such processes as germination and gall formation; he may even have suspected that plants were made up of cells, a concept that had not yet been introduced. Many of Malpighi's drawings of plant anatomy remained unintelligible to botanists until the structures were rediscovered in the 19th century. Although Malpighi was not a technical innovator, he does exemplify the functioning of the educated 17th-century mind, which, together with curiosity and patience, resulted in many advances in biology.

The discovery of "animalcules." Antonie van Leeuwenhoek, a Dutchman who spent most of his life in Delft, sold cloth for a living. As a young man, however, he became interested in grinding lenses, which he mounted in gold, silver, or copper plates. Indeed, he became so obsessed with the idea of making perfect lenses that he neglected his business and was ridiculed by his family and neighbours. Using single lenses rather than compound ones (a system of two or more), Leeuwenhoek achieved magnifications from 40 to 270 diameters, a remarkable feat for hand-ground lenses. Among his most conspicuous observations was the discovery in 1675 of the existence in stagnant water and prepared infusions of many protozoans, which he called animalcules. He observed the connections between the arteries and veins; gave particularly fine accounts of the microscopic structure of muscle, the lens of the eye, the teeth, and other structures; and recognized bacteria of different shapes, postulating that they must be on the order of 25 times as small as the red blood cell. Because this is the approximate size of bacteria, it indicates that his observations were correct. Leeuwenhoek's fame was consolidated when he confirmed the observations of a student that male seminal fluid contains spermatozoa. Furthermore, he discovered spermatozoa in other animals as well as in the female tract following copulation; the latter destroyed the idea held by others that the entire future development of an animal is centred in the egg, and that sperm merely induce a "vapour," which penetrates the womb and effects fertilization. Although this theory of preformation, as it is called, continued to survive for some time longer, Leeuwenhoek initiated its eventual demise.

The discovery of sperm

Leeuwenhoek's animalcules raised some disquieting thoughts in the minds of his contemporaries. The theory of spontaneous generation, held by the ancient world and passed down unquestioned, was now being criticized. Christiaan Huygens, a scientific friend of Leeuwenhoek,

hypothesized that these little animals might be small enough to float in the air and, on reaching water, reproduce themselves. At this time, however, criticism of spontaneous generation went no further.

Swammerdam's innovative techniques. In contrast to Leeuwenhoek, who was virtually unschooled, his contemporary fellow countryman Jan Swammerdam was an educated and highly systematic worker who confined his attention to studying relatively few organisms in great detail. He employed highly innovative techniques; for example, he injected wax into the circulatory system to hold the blood vessels firm, he dissected fragile structures under water to avoid destroying them, and he used micropipettes to inject and inflate organisms under the microscope. In 1669 Swammerdam published *Algemeene Verhandeling van bloedeloose diertjens* (*The Natural History of Insects,* 1792), in which he described the structure of a large number of insects as well as spiders, snails, scorpions, fishes, and worms. He regarded all of these animals as insects, distinguishing between them according to their mode of development. Although this classification was erroneous, Swammerdam did discover a great deal of information concerning insect development.

Unfortunately, Swammerdam was subject to fits of mental instability, which, combined with financial difficulties, led to periods of depression. It was while in a state of mental disturbance that he produced his classic *Ephemeri vita* ("Life of the Ephemera") in 1675, a book about the life of the mayfly noteworthy for its extremely detailed illustrations. Sometime after his death at the age of 43, Swammerdam's works were published collectively as the *Bijbel der Natuure* (1737; "Bible of Nature"), which is considered by many authorities to be the finest collection of microscopic observations ever produced by one man.

Grew's anatomical studies of plants. Nehemiah Grew was educated at Cambridge and is regarded by some as one of the founders of plant anatomy. In 1672 he published the first of his great books, *An Idea of a Philosophical History of Plants,* followed in 1682 by *The Anatomy of Plants.* Although Grew clearly recognized cells in plants, referring to them as vesicles, or bladders, their biological significance evaded him. He is best known for his recognition of flowers as the sexual organs of plants and for his description of their parts. He also described the individual pollen grains and observed that they are transported by bees, but he did not realize the significance of this observation. Twelve years after the publication of *The Anatomy of Plants,* a German physician utilized Grew's anatomical studies in experiments to verify sexual reproduction in plants.

The discovery of cells. Of the five microscopists, Robert Hooke was perhaps the most intellectually preeminent. As curator of instruments at the Royal Society of London, he was in touch with all new scientific developments and exhibited interest in such disparate subjects as flying and the construction of clocks. In 1665 Hooke published his *Micrographia,* which was primarily a review of a series of observations that he had made while following the development and improvement of the microscope. Hooke described in detail the structure of feathers, the stinger of a bee, the radula, or "tongue," of mollusks, and the foot of the fly. It is Hooke who coined the word cell; in a drawing of the microscopic structure of cork, he showed walls surrounding empty spaces and refers to these structures as cells. He described similar structures in the tissue of other trees and plants and discerned that in some tissues the cells were filled with a liquid while in others they were empty. He therefore supposed that the function of the cells was to transport substances through the plant.

Contributions of the classical microscopists

Although the work of any of the classical microscopists seems to lack a definite objective, it should be remembered that these men embodied the concept that observation and experiment were of prime importance, that mere hypothetical, philosophical speculations were not sufficient. It is remarkable that so few men, working as individuals totally isolated from each other, should have recorded so many observations of such fundamental importance. The great significance of their work was that it revealed, for the first time, a world in which living organisms display an almost incredible complexity.

Unfortunately, work with the compound microscope languished for nearly 200 years, mainly because the early lenses tended to break up white light into its constituent parts. This technical problem was not solved until the invention of achromatic lenses, which were introduced about 1830. In 1878 a modern achromatic compound microscope was produced from the design of the German physicist Ernst Abbe. Abbe subsequently designed a substage illumination system, which, together with the introduction of a new substage condenser, paved the way for the biological discoveries of that era.

The development of taxonomic principles. In 1687 in England Isaac Newton, mathematician, physicist, and astronomer, published his great work *Principia,* in which he described the universe as fixed, with the Earth and other heavenly bodies moving harmoniously in accordance with mathematical laws. This approach of systematizing and classifying was to dominate biology in the 17th and 18th centuries. One reason was that the 16th-century "fathers of botany" had been content merely to describe and draw plants, assembling an enormous and diverse number that continued to increase as explorations of foreign countries made it evident that every country had its own native plants and animals.

Aristotle began the process of classification when he used mode of reproduction and habitat to distinguish groups of animals. Indeed, the words genus and species are translations of the Greek *genos* and *eidos* used by Aristotle. As mentioned earlier, it was the Swiss botanist Bauhin who introduced a binomial system of classification, using a generic name and a specific name. Most classification schemes proposed before the 17th century were confused and unsatisfactory, however.

The use of structure for classifying organisms. Two systematists of the 17th and 18th centuries were John Ray and Carolus Linnaeus, also known as Carl von Linné. Ray, an English naturalist who studied at Cambridge, was particularly interested in the work of the ancient compilers of herbals, especially those who had attempted to formulate some means of classification. Recognizing the need for a classification system that would apply to both plants and animals, Ray employed in his classification schemes extremely precise descriptions for genera and species. By basing his system on structures, such as the arrangement of toes and teeth in animals, rather than colour or habitat, Ray introduced a new and very important concept to taxonomic biology.

The need for a common classification system

Reorganization of groups of organisms. Prior to Linnaeus, a Swedish botanist and taxonomist, most taxonomists started their classification systems by dividing all the known organisms into large groups and then subdividing these into progressively smaller groups. Unlike his predecessors, Linnaeus began with the species, organizing them into larger groups or genera, then arranging analogous genera to form families and related families to form orders and classes. Probably utilizing the earlier work of Grew and others, Linnaeus chose the structure of the reproductive organs of the flower as a basis for grouping the higher plants. Thus he distinguished between plants with real flowers and seeds (phanerogams) and those lacking real flowers and seeds (cryptogams), subdividing the former into hermaphroditic (bisexual) and unisexual forms. For animals, following Ray's work, Linnaeus relied upon teeth and toes as the basic characteristics of mammals; he used the shape of the beak as the basis for bird classification. Having demonstrated that a binomial classification system based on concise and accurate descriptions could be used for the grouping of organisms, Linnaeus established taxonomic biology as a discipline.

Later developments in classification were initiated by three French biologists, the Comte de Buffon, Jean-Baptiste Lamarck, and Georges Cuvier, all of whom made lasting contributions to biological science, particularly in comparative studies. Subsequent systematists have been chiefly interested in the relationships between animals and have endeavoured to explain not only their similarities but also their differences in broad terms that encompass, in addition to structure, composition, function, genetics, evolution, and ecology.

The development of comparative biological studies. Once the opprobrium attached to the dissection of human bodies had been dispelled in the 16th century, anatomists directed their efforts toward a better understanding of human structure. In doing so they generally ignored other animals, at least until the latter part of the 17th century, when biologists began to realize that important insights could be gained by comparative studies of all animals, including man. One of the first of such anatomists was Edward Tyson, an English physician who studied the anatomy of an immature chimpanzee in detail and compared it with that of man. In making further comparisons between the chimpanzee and other primates, Tyson clearly recognized points of similarity between these animals and man. Not only was this a major contribution to physical anthropology but also an indication—nearly two centuries before Darwin—of the existence of relationships between man and other primates.

Among those who gave comparative studies their greatest impetus was Georges Cuvier, a French naturalist who utilized large collections of biological specimens sent to him from all over the world to work out a systematic organization of the animal kingdom. In addition to establishing a connection between systematic and comparative anatomy, he believed that there was a "correlation of parts" according to which a given type of structure (*e.g.,* feathers) is related to a certain anatomical formation (*e.g.,* a wing), which in turn is related to other specific formations (*e.g.,* the collarbone), and so on. In other words, he felt that a great deal of anatomical information could be deduced about an organism even if the whole specimen were not available. This was to be of great practical importance in the study of fossils, in which Cuvier played a leading role. Indeed, the 1812 publication of Cuvier's *Recherches sur les ossemens fossiles de quadrupèdes* (translated as *Research on Fossil Bones* in 1835) laid the foundation for the science of paleontology. But in order to reconcile his scientific findings with his personal religious beliefs, Cuvier postulated a series of catastrophic events that could account for both the presence of fossils and the immutability of existing species.

The founding of paleontology

The study of the origin of life. *Spontaneous generation.* If a species can develop only from a preexisting species, then how did life originate? Among the many philosophical and religious ideas advanced to answer this question, one of the most popular was the theory of spontaneous generation, according to which, as already mentioned, living organisms could originate from nonliving matter. With the increasing tempo of discovery during the 17th and 18th centuries, however, investigators began to examine more critically the Greek belief that flies and other small animals arose from the mud at the bottom of streams and ponds by spontaneous generation. Then, when Harvey announced his biological dictum *ex ovo omnia* ("everything comes from the egg"), it appeared that he had solved the problem, at least insofar as it pertained to flowering plants and the higher animals, all of which develop from an egg. But Leeuwenhoek's subsequent disquieting discovery of animalcules demonstrated the existence of a densely populated but previously invisible world of organisms that had to be explained.

A 17th-century Italian physician and poet, Francesco Redi, was one of the first to question the spontaneous origin of living things. Having observed the development of maggots and flies on decaying meat, Redi in 1668 devised a number of experiments, all pointing to the same conclusion: if flies are excluded from rotten meat, maggots do not develop. On meat exposed to air, however, eggs laid by flies develop into maggots. But renewed support for spontaneous generation came from the publication in 1745 of a book, *An Account of Some New Microscopical Discoveries,* by John Turberville Needham, an English Catholic priest; he found that large numbers of organisms subsequently developed in prepared infusions of many different substances that had been exposed to intense heat in sealed tubes for 30 minutes. Assuming that such heat treatment must have killed any previous organisms, Needham explained the presence of the new population on the grounds of spontaneous generation. The experiments appeared irrefutable until Lazzaro Spallanzani, an Italian biologist, repeated them and obtained conflicting results. He published his findings around 1775, claiming that Needham had not heated his tubes long enough nor had he sealed them in a satisfactory manner. Although Spallanzani's results should have been convincing, Needham had the support of the influential French naturalist Buffon; hence the matter of spontaneous generation remained unresolved.

Conflicting experimental results

The death of spontaneous generation. After a number of further investigations had failed to solve the problem, the French Academy of Sciences, in January 1860, offered a prize for contributions that would "attempt, by means of well-devised experiments, to throw new light on the question of spontaneous generation." In response to this challenge, Louis Pasteur, who at that time was a chemist, subjected flasks containing a sugared yeast solution to a variety of conditions. Pasteur was able to demonstrate conclusively that any microorganisms that developed in suitable media came from microorganisms in the air, not from the air itself, as Needham had suggested. Support for Pasteur's findings came in 1876 from an English physicist, John Tyndall, who devised an apparatus to demonstrate that air had the ability to carry particulate matter. Because such matter in air reflects light when the air is illuminated under special conditions, Tyndall's apparatus could be used to indicate when air was pure. Tyndall found that no organisms were produced when pure air was introduced into media capable of supporting the growth of microorganisms. It was these results, together with Pasteur's findings, that put an end to the doctrine of spontaneous generation.

When Pasteur later showed that parent microorganisms generate only their own kind, he thereby established the study of microbiology. Moreover, he not only succeeded in convincing the scientific world that microbes are living creatures, which come from preexisting forms, but also showed them to be an immense and varied component of the organic world, a concept that was to have important implications for the science of ecology. Further, by isolating various species of bacteria and yeasts in different chemical media, Pasteur was able to demonstrate that they brought about chemical change in a characteristic and predictable way, thus making a unique contribution to the study of fermentation and to biochemistry.

The origin of primordial life. In the 1920s a Soviet biochemist, A.I. Oparin, and other scientists suggested that life may have come from nonliving matter under conditions that existed on the primitive Earth, when the atmosphere consisted of the gases methane, ammonia, water vapour, and hydrogen. According to this concept, energy supplied by electrical storms and ultraviolet light may have broken down the atmospheric gases into their constituent elements, and organic molecules may have been formed when the elements recombined.

The notion of life from atmospheric gases

Some of these ideas have been verified by advances in geochemistry and molecular genetics; experimental efforts have succeeded in producing amino acids and proteinoids (primitive protein compounds) from gases that may have been present on the Earth at its inception, and amino acids have been detected in rocks that are more than 3,000,000,000 years old. With improved techniques it may be possible to produce precursors of or actual self-replicating living matter from nonliving substances. But whether it is possible to create the actual living heterotrophic forms from which autotrophs supposedly developed remains to be seen.

Although it may never be possible to determine experimentally how life originated or whether it originated only once or more than once, it would now seem—on the basis of the ubiquitous genetic code found in all living organisms on Earth—that life appeared only once and that all the diverse forms of plants and animals evolved from this primitive creation.

Biological expeditions. Although a number of 16th- and 17th-century travellers provided much valuable information about the plants and animals in the Orient, America, and Africa, most of this information was collected by curious individuals rather than trained observers. A devel-

opment that occurred during the 18th and 19th centuries was the organization of scientific expeditions, usually under the auspices of a particular government. The most notable of these efforts were the voyages of the "Endeavour," the "Investigator," the "Beagle," and the "Challenger," all sponsored by the English government.

Captain James Cook sailed the "Endeavour" to the South Sea islands, New Zealand, New Guinea, and Australia in 1768; the voyage provided Joseph Banks, a young naturalist, with the opportunity to make a very extensive collection of plants and notes, which helped establish him as a leading biologist. Another expedition to the same area in the "Investigator" in 1801 included a botanist, Robert Brown, whose work on the plants of Australia and New Zealand became a classic; especially important were his descriptions of how certain plants adapt to different environmental conditions. Brown is also credited with discovering the cell nucleus and analyzing sexual processes in higher plants.

Darwin's voyage on the HMS "Beagle"

One of the most famous biological expeditions of all time was that of the "Beagle" in 1831, the members including Charles Darwin. Although Darwin's primary interest at the time was geology, his visit to the Galápagos Islands aroused his interest in biology and caused him to speculate about their curious insular animal life and the significance of isolation in space and time for the formation of species. During the "Beagle" voyage, Darwin collected specimens of and accumulated copious notes on the plants and animals of South America and Australia, for which he received great acclaim on his return to England.

The voyage of the "Challenger" from 1872 to 1876 was organized by the British Admiralty to study oceanography, meteorology, and natural history. Under the leadership of Charles Wyville Thomson, the chief naturalist, vast collections of plants and animals were made, the importance of plankton (minute free-floating aquatic plants and animals) as a source of food for larger marine organisms was recognized, and many new planktonic species were discovered. A particularly significant aspect of the "Challenger" voyage was the interest it stimulated in the new science of marine biology.

In spite of these expeditions, the contributions made by individuals were still very important. Such an individual was the English naturalist Alfred Russel Wallace, who undertook explorations of the Malay Peninsula from 1854 to 1862. In 1876 he published his book *The Geographical Distribution of Animals,* in which he divided the landmasses into six zoogeographical regions and described their characteristic fauna. Wallace also contributed to the theory of evolution, publishing in 1870 a book expressing his views, *Contributions to the Theory of Natural Selection.*

The development of the cell theory. Although the microscopists of the 17th century had made detailed descriptions of plant and animal structure and though Hooke had coined the term cell for the compartments he had observed in cork tissue, their observations lacked an underlying theoretical unity. It was not until 1838 that Matthias J. Schleiden, a German botanist interested in plant anatomy, stated, "the lower plants all consist of one cell, while the higher ones are composed of (many) individual cells." When Schleiden's friend, the German physiologist Theodor Schwann, extended the cellular theory to include animals, he thereby brought about a rapprochement between botany and zoology. The formation of the cell theory—all plants and animals are made up of cells—marked a great conceptual advance in biology, and it resulted in renewed attention to the living processes that go on in cells.

Uniting of botany and zoology

In 1846, after several investigators had described the streaming movement of the cytoplasm in plant cells, Hugo von Mohl, a German botanist, coined the word protoplasm to designate the living substance of the cell. The concept of protoplasm as the physical basis of life led to the development of cell physiology.

A further extension of the cell theory was the development of cellular pathology by Rudolf Virchow, who established the relationship between abnormal events in the body and unusual cellular activities. This gave a new direction to the study of pathology and resulted in advances in medicine.

The detailed description of cell division was contributed by Eduard Strasburger, a German botanist, who observed the mitotic process in plant cells and further demonstrated that nuclei arise only from preexisting nuclei. The parallel work in mammals was done by the German anatomist Walther Flemming, who published his most important findings in *Zellsubstanz, Kern und Zelltheilung* ("Cell Substance, Nucleus and Cell Division") in 1882.

The theory of evolution. As knowledge of plant and animal forms accumulated during the 16th, 17th, and 18th centuries, a few biologists began to speculate about the ancestry of these organisms, though the prevailing view was that promulgated by Linnaeus—namely, the immutability of the species. Among the early speculations voiced during the 18th century, Erasmus Darwin, an English physician and the grandfather of Charles Darwin, concluded that species descend from common ancestors and that there is a struggle for existence among animals. A French naturalist, Jean-Baptiste Lamarck, who was probably the most important of the 18th-century evolutionists, recognized the role of isolation in species formation; he also saw the unity in nature and conceived the idea of the evolutionary tree.

A complete theory of evolution was not announced, however, until the publication in 1859 of Charles Darwin's *On the Origin of Species by Means of Natural Selection or the Preservation of Favoured Races in the Struggle for Life.* In his book Darwin stated that all living creatures multiply so rapidly that, if left unchecked, they would soon overpopulate the world. According to Darwin, the checks on population size are maintained by competition for the means of life. Hence, if any member of a species differs in some way that makes it better fitted to survive, then it will have an advantage that its offspring would be likely to perpetuate. Darwin's work reflects the influence of a British economist, Thomas Robert Malthus, who in 1838 published an essay on population in which he warned that if man multiplies more rapidly than his food supply, competition for existence would result. Darwin was also influenced by a British geologist, Charles Lyell, who realized from his studies of geological formations that the relative ages of deposits could be estimated by means of the proportion of living and extinct mollusks. But it was not until after his travels in the "Beagle" in 1831, during which he observed a great richness and diversity of island fauna, that Darwin began to develop his theory of evolution. Alfred Russel Wallace had reached conclusions similar to those of Darwin following his studies of plants and animals in the Malay Peninsula. A short paper dealing with this subject sent by Wallace to Darwin finally resulted in the publication of Darwin's own theories.

Influences on Darwin's work

Conceptually, the theory was of the utmost significance, accounting as it did for the formation of new species. Following the subsequent discovery of the chromosomal basis of inheritance and the laws of heredity, it could be seen that natural selection does not involve the sharp alternatives of life or death but results from the differential survival of variants. Today, the universal principle of natural selection, which is the central concept of Darwin's theory, is firmly established.

The study of the reproduction and development of organisms. *Preformation versus epigenesis.* A question posed by Aristotle was whether the embryo is preformed and therefore only enlarges during development or whether it differentiates from an amorphous beginning. Two conflicting schools of thought had been based on this question: the preformation school maintained that the egg contains a miniature individual that develops into the adult stage in the proper environment; the epigenesis school believed that the egg is initially undifferentiated and that development occurs as a series of steps. Prominent supporters of the preformation doctrine, which was widely held until the 18th century, included Malpighi, Swammerdam, and Leeuwenhoek. In the 19th century, as criticism of preformation mounted, Karl Ernst von Baer, an Estonian embryologist, provided the final evidence against the theory. His discovery of the mammalian egg and his recognition of the formation of the germ layers out of which the embryonic organs develop laid the foundations of modern embryology.

The fertilization process. Despite the many early descriptions of spermatozoa, their essential role in fertilization was not proven until 1879, when Hermann Fol, a Swiss physician and zoologist, observed the penetration of a spermatozoon into an ovum. Prior to this discovery, during the period from 1823 to 1830, the existence of the sexual process in flowering plants had been demonstrated by Giovanni Battista Amici, an Italian astronomer and botanist, and confirmed by others. The discovery of fertilization in plants was of great importance to the development of plant hybrids, which are produced by cross-pollination between different species; it was also of great significance to the studies of genetics and evolution.

The universal occurrence and remarkable similarity of the fertilization process, regardless of the organism in which it occurs, provoked many of the leading investigators of the time to search for the underlying mechanism. It was realized that there must be some way by which the number of chromosomes is reduced before fertilization; otherwise the chromosome number would double every time a spermatozoon fused with an egg. In 1883 Edouard van Beneden, a Belgian cytologist, showed that the eggs and spermatozoa in the worm *Ascaris* contain half the number of chromosomes found in the body cells. To account for the halving of the chromosomes in the sex cells, a process that is called meiosis, in 1887 August Weismann, a German biologist, suggested that there must be two different types of cell division, and by 1900 the details of meiosis had been elucidated.

Van Beneden's work on chromosomes

The study of heredity. *Pre-Mendelian theories of heredity.* The fundamental laws of heredity were discovered in 1865 by Gregor Mendel, an Austrian monk and biologist, but his work was ignored until its rediscovery in 1900. There were, however, a number of views on the subject that had been expressed long before Mendel. The Greek philosophers, for example, believed that the traits of individuals were acquired from contact with the environment and that such acquired characteristics could be inherited by offspring. Because Lamarck was the most famous proponent of the inheritance of acquired characteristics, the theory is called Lamarckism. This concept, which emphasized the use and disuse of organs as the significant factor in determining the characteristics of an individual, postulated that any alterations in the individual could be transmitted to the offspring through the gametes. Yet the inheritance of acquired characteristics has never been experimentally verified, despite many attempts. Furthermore, many of Lamarck's examples, such as the long neck of the giraffe, can be more satisfactorily explained by means of natural selection.

In 1885 Weismann suggested that hereditary characteristics were transmitted by what he called germ plasm—as distinguished from the somatoplasm (body cells)—which linked the generations by a continuous stream of dividing germ cells. In stating definitely seven years later that the material of heredity was in the chromosomes, Weismann anticipated the chromosomal basis of inheritance.

Francis Galton, a 19th-century English anthropologist, made a number of important contributions to genetics, one of which was a study of the hereditary nature of ability, from which he developed the concept that judicious breeding could improve the human race (eugenics). Galton's most significant work was the demonstration that each generation of ancestors makes a proportionate contribution to the total makeup of the individual. Thus, he suggested that if a tall man marries a short woman, each should contribute half of the total heritage, and the resultant offspring should be intermediate between the two parents.

Mendelian laws of heredity. The fame of Gregor Mendel, the father of genetics, rests on experiments he did with garden peas, which possess sharply contrasting characteristics—*e.g.,* tall versus short; round seed versus wrinkled seed. When Mendel fertilized short plants with pollen from tall plants, he found the offspring (first filial generation) to be uniformly tall. But if he allowed the plants of this generation to self-pollinate (fertilize themselves), their offspring (the second filial generation) exhibited the characters of the grandparents in a rather consistent ratio of three tall to one short. Furthermore, if allowed to self-pollinate, the short plants always bred true—*i.e.,* never produced anything but short plants. From these results Mendel developed the concept of dominance, based on the supposition that each plant carried two trait units, one of which dominated the other. Nothing was known at that time about chromosomes or meiosis, yet Mendel deduced from his results that the trait units, later called genes, could be a kind of physical particle that was transmitted from one generation to another through the reproductive mechanism.

Discovery of dominance

Mendel's most important concept was the idea that the paired genes present in the parent separate or segregate during the formation of the gametes. Moreover, in later experiments in which he studied the inheritance of two pairs of traits, Mendel showed that one pair of genes is independent of another. Thus, the principles of segregation and of independent assortment were established.

Mendel's findings were ignored for 35 years, probably for two reasons. Because the distinguished Swiss botanist Karl Wilhelm von Nägeli failed to recognize the significance of the work after Mendel had sent him the results, he did nothing to encourage Mendel. Nägeli's great prestige and the lack of his endorsement indirectly weighed against widespread recognition of Mendel's work. Moreover, when the work was published, little was known about the cell, and the processes of mitosis and meiosis were completely unknown. Mendel's work was finally rediscovered in 1900, when three botanists independently recognized the worth of his studies from their own research and cited his publication in their work.

Elucidation of the hereditary mechanism. By 1901 it was understood how the hereditary units postulated by Mendel are distributed; it was also known that the somatic (body) cells have a double, or diploid, complement of chromosomes, while the reproductive cells have a single, or haploid, chromosome number. The experimental demonstration of the chromosomal basis for heredity had been firmly established by the German biologist Theodor Boveri soon after the turn of the century and subsequently confirmed by others. To account for the large number of observed hereditary characters, Boveri suggested that each chromosome in a pair can exchange the hereditary factors it carries with those of the other chromosome. At first the U.S. geneticist Thomas Hunt Morgan dismissed this concept, but later, when he found that it agreed with his own laboratory findings, Morgan and his collaborators assigned the hereditary units (genes) specific positions, or loci, within the chromosomes. With the genes established as the carriers of hereditary traits, William Bateson, an English biologist, coined the name genetics for the experimental study of heredity and evolution.

BIOLOGY IN THE 20TH CENTURY

Just as the 19th century can be considered the age of cellular biology, the 20th century has been characterized by developments in molecular biology.

The rise of molecular biology

Important conceptual developments. By utilizing modern methods of investigation, such as X-ray diffraction and electron microscopy, to explore levels of cellular organization beyond that visible with a light microscope—*i.e.,* the ultrastructure of the cell—new concepts of cellular function have been produced. Not only has the study of the molecular organization of the cell probably had the greatest impact upon biology during the 20th century but it also has led directly to the convergence of many different scientific disciplines in order to acquire a better understanding of life processes.

Another 20th-century development has been the realization that man is as dependent upon the Earth's natural resources as are other animals. The progressive destruction of the environment can be attributed, in part, to an increase in population pressure as well as to certain technological advances. Thus, though lifesaving advances in medicine have resulted in a dramatic drop in the death rate, they have also been a factor contributing to the explosive increase in the human population. Moreover, chemical contaminants being introduced into the environment by manufacturing processes, pesticides, automobile

emissions, and other means are seriously endangering all forms of life. It is for these reasons that biologists are beginning to pay much greater attention to the relationships of living things to each other as well as to their biotic and abiotic environments.

Intradisciplinary work. There are many important categories in the biological sciences. Botany, zoology, and microbiology deal with types of organisms and their relationships with each other. Such disciplines are subdivided into more specialized categories; for example, ichthyology is the study of fishes, algology the study of algae. All of them draw upon paleontology, taxonomy, morphology, and evolution.

Disciplines such as embryology and physiology, which deal with the development and function of an organism, may be divided further according to the kind of organism studied; for example, invertebrate embryology and mammalian physiology. In the past few decades, many developments in physiology and embryology have resulted from studies in cell biology, biophysics, and biochemistry. This has given rise to cell physiology, cytochemistry, and ultrastructural studies, which aim at correlating structure with function. Ecology, the study of the relations of a group of organisms to its environment, includes both the physical features of the environment and other organisms that may compete for food and shelter. Ecology may be subdivided according to the environment—for example, freshwater ecology and marine ecology—and draws upon animal behaviour. One aspect of cell biology, formerly called cytology, is the investigation of the structure, composition, and function of cells; biochemistry and biophysics provide important information.

Thus, biology encompasses a number of disciplines; in fact, it has become common to divide biology into its several levels of organization rather than separating the disciplines. It is useful, for example, to differentiate between organismic biology, the study of the whole organism, and cell biology. Similarly the technological advances of the 20th century have allowed increased understanding of the molecules comprising living things and their aggregation and organization into such structures as chromosomes and membranes. Knowledge of this aspect, called molecular biology, represents the molecular level of organization. The fourth level, population biology, involves the complex interaction of population of animals and plants with the environment.

Importance of technological advances

Relations with other disciplines. In the 17th century, with the invention of the microscope, which made possible study of the cellular level of organization, biology began to receive the benefits of scientific developments in physics. In the 18th century such developments in chemistry as a better understanding of the nature of oxygen, carbon dioxide, and water began to have important implications for biology. Today, through the disciplines of biochemistry and biophysics, both chemistry and physics have continued to make significant contributions to biology, particularly in the area of molecular biology.

Biology is also very closely related to the disciplines of medicine and agriculture, out of which it developed as an independent discipline. In a sense, the roles have been reversed in the 20th century, for it is basic research being conducted in biology that is contributing to major advances currently being made in medicine and agriculture. It was biological research in the structure and function of viruses, for example, that led directly to the development of a vaccine against poliomyelitis.

Another scientific discipline, that of geology, is closely related to the biological study of paleontology. The technique of radiocarbon dating, which was developed by chemists to determine the age of biological remains, has been of great use in the fields of archaeology and anthropology as well as biology. A new discipline, space biology, has arisen through the activities of the scientists and engineers concerned with the exploration of space. The conceptual framework of biology has had to be altered to accommodate newly discovered facts. In the process biology has received contributions from and made contributions to many other disciplines, in the humanities as well as in the sciences. (S.H.J./E.R.G.)

Changing social and scientific values. The biologist's role in society as well as his moral and ethical responsibility in the discovery and development of new ideas has led to a reassessment of his social and scientific value systems. A scientist can no longer ignore the consequences of his discoveries; he is as concerned with the possible misuses of his findings as he is with the basic research in which he is involved. This emerging social and political role of the biologist and all other scientists requires a weighing of values that cannot be done with the accuracy or the objectivity of a laboratory balance. As a member of society, it is necessary for a biologist now to redefine his social obligations and his functions, particularly in the realm of making judgments about such ethical problems as man's control of his environment or his manipulation of genes to direct further evolutionary development.

Coping with problems of the future. As a result of recent discoveries concerning hereditary mechanisms, genetic engineering, by which human traits are made to order, may soon be a reality. As desirable as it may seem to be, such an accomplishment would entail many value judgments. Who would decide, for example, which traits should be selected for change? In cases of genetic deficiencies and disease, the desirability of the change is obvious, but the possibilities for social misuse are so numerous that they may far outweigh the benefits.

Genetic engineering

Probably the greatest biological problem of the future, as it is of the present, will be to find ways to curb environmental pollution without interfering with man's constant effort to improve the quality of his life. Many scientists believe that underlying the spectre of pollution is the problem of surplus human population. A rise in population necessitates an increase in the operations of modern industry, the waste products of which increase the pollution of air, water, and soil. With predictions that, at the present rate of reproduction, the Earth's population will be approximately 7,000,000,000 by the year 2000, the question of how many people the resources of the Earth can support is one of critical importance.

Although the solutions to these and many other problems are yet to be found, they do indicate the need for biologists to work with social scientists and other members of society in order to determine the requirements necessary for maintaining a healthy and productive planet. For although many of man's present and future problems may seem to be essentially social, political, or economic in nature, they have biological ramifications that could affect the very existence of life itself. (E.R.G.)

The study of biological structure and function

MORPHOLOGY

Morphology is a term used in biology for the study not only of shape and structure in plants, animals, and microorganisms but also of the size, shape, structure, and relations of the parts comprising them. The term morphology connotes the general aspects of biological form and arrangement of the parts of a plant or an animal. The term anatomy also refers to the study of biological structure but usually connotes study of the details of either gross or microscopic structure. In practice, however, the two terms are used almost synonymously.

Typically, morphology is contrasted with physiology, which deals with studies of the functions of organisms and their parts; function and structure are so closely interrelated, however, that their separation is somewhat artificial. Morphologists were originally concerned with the bones, muscles, blood vessels, and nerves comprising the bodies of animals and the roots, stems, leaves, and flower parts comprising the bodies of higher plants. The development of the light microscope made possible the examination of some structural details of individual tissues and single cells; the development of the electron microscope and of methods for preparing ultrathin sections of tissues created an entirely new aspect of morphology—that involving the detailed structure of cells. Electron microscopy has gradually revealed the amazing complexity of the many structures comprising the cells of plants and animals. Other physical techniques have permitted biologists to investigate the

morphology of complex molecules such as hemoglobin, the gas-carrying protein of blood, and deoxyribonucleic acid (DNA), of which most genes are composed. Thus, morphology encompasses the study of biological structures over a tremendous range of sizes, from the macroscopic to the molecular.

A thorough knowledge of structure (morphology) is of fundamental importance to the physician, to the veterinarian, and to the plant pathologist, all of whom are concerned with the kinds and causes of the structural changes that result from specific diseases.

Historical background. Evidence that prehistoric man appreciated the form and structure of his contemporary animals has survived in the form of paintings on the walls of caves in France, Spain, and elsewhere. During the early civilizations of China, Egypt, and the Near East, as man learned to domesticate certain animals and to cultivate many fruits and grains, he also acquired knowledge about the structures of various plants and animals.

Aristotle was interested in biological form and structure, and his *Historia animalium* contains excellent descriptions, clearly recognizable in extant species, of the animals of Greece and Asia Minor. He was also interested in developmental morphology and studied the development of chicks before hatching and the breeding methods of sharks and bees. Galen was among the first to dissect animals and to make careful records of his observations of internal structures. His descriptions of the human body, though they remained the unquestioned authority for more than 1,000 years, contained some remarkable errors, for they were based on dissections of pigs and monkeys rather than of humans.

Emergence of modern morphology

Although it is difficult to pinpoint the emergence of modern morphology as a science, one of the early landmarks was the publication in 1543 of *De humani corporis fabrica* by Andreas Vesalius, whose careful dissections of human bodies and accurate drawings of his observations revealed many of the inaccuracies in Galen's earlier descriptions of the human body.

In 1661 an Italian physiologist, Marcello Malpighi, the founder of microscopic anatomy, demonstrated the presence of the small blood vessels called capillaries, which connect arteries and veins. The existence of capillaries had been postulated 30 years earlier by the English physician William Harvey, whose classic experiments on the direction of blood flow in arteries and veins indicated that minute connections must exist between them. Between 1668 and 1680, the Dutch microscopist Antonie van Leeuwenhoek used the recently invented microscope to describe red blood cells, human sperm cells, bacteria, protozoans, and various other structures.

Cellular components

Cellular components—the nucleus and nucleolus of plant cells and the chromosomes within the nucleus—and the complex sequence of nuclear events (mitosis) that occur during cell division were described by various scientists throughout the 19th century. *Organographie der Pflanzen* (1898–1901; *Organography of Plants*, 1900–05), the great work of a German botanist, Karl von Goebel, who was associated with morphology in all its aspects, remains a classic in the field. The Scot John Hunter and the Frenchman Georges Cuvier were early 19th-century pioneers in the study of similar structures in different animals—*i.e.,* comparative morphology. Cuvier in particular was among the first to study the structures of both fossils and living organisms and is credited with founding the science of paleontology. A British biologist, Sir Richard Owen, developed two concepts of basic importance in comparative morphology—homology, which refers to intrinsic structural similarity, and analogy, which refers to superficial functional similarity. Although the concepts antedate the Darwinian view of evolution, the anatomical data on which they were based became, largely as a result of the work of the German comparative anatomist Carl Gegenbaur, important evidence in favour of evolutionary change, despite Owen's steady unwillingness to accept the view of diversification of life from a common origin.

One of the major thrusts in contemporary morphology has been the elucidation of the molecular basis of cellular structure. Techniques such as electron microscopy have revealed the complex details of cell structure, provided a basis for relating structural details to the particular functions of the cell, and shown that certain cellular components occur in a variety of tissues. Studies of the smallest components of cells have clarified the structural basis not only for the contraction of muscle cells but also for the motility of the tail of the sperm cell and the hairlike projections (cilia and flagella) found on protozoans and other cells. Studies involving the structural details of plant cells, although begun somewhat later than those concerned with animal cells, have revealed fascinating facts about such important structures as the chloroplasts, which contain chlorophyll that functions in photosynthesis. Attention has also been focussed on the plant tissues composed of cells that retain their power to divide (meristems), particularly at the tips of stems, and their relationship with the new parts to which they give rise. The structural details of bacteria and blue-green algae, which are similar to each other in many respects but markedly different from both higher plants and animals, have been studied in an attempt to determine their origin.

Morphology continues to be of importance in taxonomy because morphological features characteristic of a particular species are used to identify it. As biologists have begun to devote more attention to ecology, the identification of plant and animal species present in an area and perhaps changing in numbers in response to environmental changes has become increasingly significant.

Fundamental concepts. *Homology and analogy.* Homologous structures develop from similar embryonic substances and thus have similar basic structural and developmental patterns, reflecting common genetic endowments and evolutionary relationships. In marked contrast, analogous structures are superficially similar and serve similar functions but have quite different structural and developmental patterns. The arm of a man, the wing of a bird, and the pectoral fins of a whale are homologous structures in that all have similar patterns of bones, muscles, nerves, blood vessels, and similar embryonic origins; each, however, has a different function. The wings of birds and those of butterflies, in contrast, are analogous structures—*i.e.,* both allow flight but have no developmental processes in common.

Application to molecular structure

The terms homology and analogy are also applied to the molecular structures of cellular constituents. Because the hemoglobin molecules from different vertebrate species contain remarkably similar sequences of amino acids, they may be termed homologous molecules. In contrast, hemoglobin and hemocyanin, the latter of which is present in crab blood, are termed analogous molecules because they have a similar function (oxygen transport) but differ considerably in molecular structure. Corresponding similarities occur in the structures of other proteins from different species—*e.g.,* cytochrome c and other enzymes (biological catalysts) such as the lactic dehydrogenases in birds and mammals.

Body plan and symmetry. The bodies of most animals and plants are organized according to one of three types of symmetry: spherical, radial, or bilateral. A spherically symmetrical body is similar throughout and can be cut in any plane through the centre to yield two equal halves. A few of the simplest plants and animals are spherically symmetrical—*e.g.,* protozoans such as Radiolaria and Heliozoia. Radially symmetrical bodies, such as those of starfishes and mushrooms, have a distinguishable top and bottom and usually have a cylindrical shape, with the body parts radiating from the central axis. A starfish can be cut into two equal halves by any plane that includes the line, or axis, running through its centre from top to bottom. The anterior, or oral, end usually contains the mouth; a posterior, or aboral, end may have an anus. In the bilaterally symmetrical body of higher animals including man, only a cut from head to foot exactly in the centre divides the body into equivalent halves. An anterior, or head, end and a posterior, or tail, end can be distinguished; and the dorsal, or back, side can be distinguished from the ventral, or belly, side. But because some internal organs of man are not symmetrical (*e.g.,* the heart), even the right and left halves of the human body are not exactly equivalent.

A few organisms—amoebas, slime molds, and certain sponges—with an irregular form, or one that changes as the organism moves, have no plane of symmetry.

Morphological basis of classification. The features that distinguish closely related species of plants and animals are usually superficial differences such as colour, size, and proportion. In contrast, the major divisions, or phyla, of the plant and animal kingdoms are distinguished by characteristics that, though usually not unique to a single division or phylum, occur in unique combinations in each.

One morphological feature useful in classifying animals and in determining their evolutionary relationships is the presence or absence of cellular differentiation—*i.e.,* animals may be either single celled or composed of many kinds of cells specialized to perform particular functions. Some multicellular animals have only two embryonic cell, or germ, layers: an ectoderm (outer layer) and an endoderm (inner layer), which lines the digestive tract. Other animals have these, in addition to a mesoderm, which lies between the ectoderm and endoderm. Animals may have one of two types of body cavity. The bodies of the Coelenterata (invertebrates such as the jellyfish) and other primitive many-celled animals consist of a double-walled sac surrounding a single cavity with a mouth. Higher animals have two cavities, and their bodies are constructed on a so-called tube-within-a-tube plan. An inner tube, or digestive tract, is lined with endoderm and opens at each end to form the mouth and the anus. An outer tube, or body wall, is covered with ectoderm. Between the two tubes a second cavity, or coelom, lies within the mesoderm and is lined by it. Another major distinguishing morphological feature of animal phyla is the presence or absence of segmentation. The members of several phyla have bodies characterized by the presence of a row of segments, or body units, of the same fundamental structure. Segmented animals include the vertebrates, the annelids (invertebrates such as the earthworm), and the arthropods (invertebrates such as insects); in some segmented animals such as man and most vertebrates, however, the segmental character of the body is obscured. An evolutionary tendency in many animal phyla has been the progressive differentiation of the anterior end to form a head with conspicuous sense organs and an accumulation of nervous tissues, a brain; the tendency is termed cephalization. Some morphological structures are found only in one phylum; for example, only the Coelenterata have stinging cells (nematocysts); the Echinodermata (invertebrates such as starfishes) have a peculiar water vascular system, and only the Chordates (*e.g.,* reptiles, birds) have a dorsally located, hollow nerve cord.

Segmentation and cephalization

Like animals, plants may be either single celled or composed of many kinds of specialized cells. The bodies of most of the lower plants, such as algae and fungi, are comprised of the least differentiated and least specialized type of plant cells, parenchyma cells. The embryonic tissues of higher plants, unlike those of animals, remain extremely active throughout the life of the plant. In addition, the different types of cells characteristic of the body of higher plants arise from meristems, specific regions in the plant body where cells divide and enlarge. In all but the simplest forms, the plant body is composed of various types of cells associated in more or less definite ways to form systems of units called tissue systems—*e.g.,* the vascular system consisting of conductive tissues. The arrangement of the components of the vascular system is a distinguishing morphological feature of various plant groups. The character and relative extent of the two phases in the life history of a plant—the sexual phase, or gametophyte, and the sporophyte—vary considerably among the plant groups and are useful in distinguishing them.

Areas of study. *Anatomy.* The best known aspect of morphology, usually called anatomy, is the study of gross structure, or form, of organs and organisms. It should not be inferred however, that even the human body, which has been extensively studied, has been so completely explored that nothing remains to be discovered. It was found only in 1965, for example, that the nerve to the pineal gland, which lies on the upper surface of the brain of mammals, is a branch from the sympathetic nerves; the sympathetic nerves receive nerve impulses from a small branch of the nerves that transmit impulses from the eye to the brain (optic nerves). Thus the pineal gland responds by a very indirect route to quantitative changes in the environmental lighting and secretes appropriate amounts of the substance it forms, the hormone melatonin.

Detailed comparisons of the morphological features of different animals, termed comparative anatomy, provide strong arguments for the evolutionary relationships among different species. In the course of evolution, animals and plants tend to undergo adaptive morphological changes that enable them to survive under certain environmental conditions. As a result, animals only remotely related evolutionarily may come to resemble each other superficially because of common adaptations to similar environments, a phenomenon known as convergent evolution. Structural similarities—streamlined shape, dorsal fins, tail fins, and flipper-like forelimbs and hindlimbs, for example—have evolved in such varied animal groups as the dolphins and porpoises, both of which are mammals; the extinct ichthyosaurs, which were reptiles; and both the bony and cartilaginous fishes. In a like manner, the mole, an insectivore, and the gopher, a rodent, have both evolved shovellike forelimbs, an adaptation for digging.

An opposite phenomenon, divergent evolution, occurs when animals originally closely related adapt to different environments and come to be superficially quite different. Although sea lions and seals, for example, are carnivores and thus closely related to bears, cats, and dogs, their adaptations to an aquatic existence have resulted in morphological characteristics distinct from those of the terrestrial carnivores. In the course of mammalian evolution, many features have changed to permit specific animal groups to adapt to particular environments—*e.g.,* the number and shape of the teeth, the length and number of bones in the limbs, the number and attachment sites of muscles, the thickness and colour of the hair or fur, and the length and shape of the tail.

Careful study of adaptive morphological aspects has permitted inferences about the course of the evolutionary history of various animals and of their successive adaptations to changing environments. The present-day Australian tree-climbing kangaroos, for example, are the descendents of a ground-dwelling marsupial, from whom evolved forms that began to live in trees and eventually developed limbs adapted to tree climbing. But the events may have occurred in the reverse sequence; that is, specialized limbs may have evolved before the animal adopted an arboreal mode of life. In any event, some of the tree-dwelling kangaroos subsequently left the trees, became readapted to life on the ground (*i.e.,* their hindlegs became adapted for leaping), and then went back to the trees but with legs so highly specialized for leaping as to be useless in grasping a tree trunk; consequently, present-day tree kangaroos climb by bracing their feet against a tree trunk, as do bears. Careful comparisons of the feet of the many kinds of living Australian marsupials reveal the stages in this complicated process of adaptation and re-adaptation.

Relationship of morphological phenomena to evolution

Changes in genes (mutations) constantly occur and may cause a decrease in size and function of an organ; on the other hand, a change in the environment or in the mode of life of a species may make an organ unnecessary for survival. As a result, many plants and animals contain organs or parts of organs that are useless, degenerate, undersized, or lacking some essential part when compared to homologous structures in related organisms. The human body, for instance, has more than 100 such organs—*e.g.,* the appendix, the fused tail vertebrae (coccyx), the wisdom teeth, the muscles that wiggle the ears, and the hair on the body.

Changes in genes and in environment

The parts of a seed plant include roots, stems, leaves, and reproductive organs in the flowers. The evolution of specialized conducting tissues called xylem and phloem has enabled seed plants to survive on land and to attain large sizes. Roots anchor the plant; enable it to maintain an upright position; and absorb water, minerals, and other nutrients from the soil. The roots of plants such as carrots, beets, and yams serve as sites for food storage. The stem links the roots with the leaves, where photosynthesis oc-

curs, and its xylem and phloem are continuous with those of root and leaf. The stem supports leaves, flowers, and fruits. Each year, the stems of woody plants add a layer of xylem and phloem, the annual ring, the width of which varies with climatic conditions. A leaf consists of a petiole (stalk), by which it is attached to the stem, and a blade, typically broad and flat, that contains bundles, or veins, of xylem and phloem on the undersurface. The flower contains pollen-producing anthers and egg-producing ovules. After fertilization the base of the flower, or ovary, enlarges and forms the fruit, which is a mature ovary containing seeds, or mature ovules. The bodies of ferns and mosses also are composed of roots, stems, and leaves, but those of lower plants such as mushrooms and kelps are much more simple and lack true roots, stems, and leaves.

Histology. A major trend in the evolution of both plants and animals has resulted in the specialization of cells and a division of labour among them. The cells comprising a tree or a man are quite different; each is specialized to carry out certain functions. Although specialization may permit a cell to function efficiently, it also increases the interdependence of body parts; an injury to or the destruction of one part, therefore, may result in death of the whole organism. The study of the structure and arrangement of tissues, defined as groups or layers of cells that together perform certain special functions, is termed histology. Each kind of tissue is composed of cells with characteristic features such as size, shape, and relationship to adjacent cells and may also contain noncellular material—connective tissue fibres or a bony material.

Types of tissues and their roles

Morphologists usually separate animal tissues into six groups; epithelial, connective, muscular, blood, nervous, and reproductive tissues. The cells composing epithelial tissues form a continuous layer or sheet that either covers the surface of the body or lines some cavity within the body, thus protecting the underlying cells from mechanical and chemical injury or from invasion by microorganisms. Epithelial tissues absorb nutrients and water, secrete a wide variety of substances, and may play a role in the reception of sensory stimuli. The connective tissues—bone, cartilage, ligaments, and fibrous connective tissue—support and hold together the other cells of the body. The cells of the connective tissues secrete large quantities of nonliving material (matrix), the characteristics of which largely determine the nature and the function of the specific types of connective tissue; the matrix secreted by fibrous connective tissue cells, for example, is a thick matted network of microscopic fibres surrounding the connective tissue cells. Connective tissue holds skin to muscle, keeps glands in position, makes up the tough outer walls of the blood vessels, and forms a sheath around nerve fibres and muscle cells. Tendons are flexible, cable-like cords of specialized fibrous connective tissue that join muscles to each other or muscle to bone. Ligaments are somewhat elastic cords of specialized fibrous connective tissue that join one bone to another.

Muscular tissues are composed of elongated, cylindrical, or spindle-shaped cells, each of which contains many small fibres called myofibrils. Muscle cells perform mechanical work by contracting—that is, by becoming shorter and thicker. The three types of vertebrate muscles include the cardiac muscle, which is found only in the walls of the heart; smooth muscles, which are found in the walls of the digestive tract and in other internal organs; and skeletal muscles, which make up the bulk of the muscle masses attached to the bones of the body. Skeletal and cardiac muscles have alternating light and dark stripes the relative sizes of which change during the contraction process. Evidence from electron microscopy indicates that two types of filaments occur in muscle; during contraction, one type of filament slides past the other.

Nerve tissue is made of cells, called neurons, which are specialized to conduct nerve impulses. Two or more thin hairlike fibres, called axons and dendrites, extend from the enlarged cell body containing the nucleus. The neurons extending from the spinal cord to the end of an appendage (*e.g.*, arm, leg) may extend to a metre (about three feet) or more in man and to several metres in an elephant or a whale.

Egg cells in the female and sperm cells in the male are reproductive tissues adapted for the production of offspring. The egg cell is modified by the accumulation of considerable amounts of yolk and other food reserves. The highly specialized spermatozoon contains a tail, the beating of which propels it to the egg.

Blood is composed of red cells, which are specialized for the transport of oxygen and carbon dioxide, and white cells, which engulf bacteria and produce antibodies (proteins formed in response to foreign substances called antigens). Blood also contains platelets, small fragments of cells from the bone marrow that play a key role in initiating the clotting of blood.

Types of plant tissues

The cells of higher plants may be differentiated into meristematic, protective, fundamental, and conductive tissues. Meristematic tissues, which are composed of small, thin-walled cells with few or no vacuoles (cavities), differentiate into the other types of plant tissue and are found in the rapidly growing parts of the plant—*e.g.*, at the tips of roots and stems. Protective tissues are composed of thick-walled cells that protect the underlying thin-walled cells from mechanical abrasion and dehydration; examples of protective tissues include the epidermis of leaves and the cork layers of stems and roots. The fundamental tissues comprising the body of a plant include the soft parts of the leaf, the components of the pith and the cortex of stems, the roots, and the soft parts of flowers and fruits. These tissues function in the production and storage of food. Two types of conductive tissues occur in higher plants: xylem conducts water and dissolved salts, and phloem conducts dissolved organic materials such as sugars. Both types are composed of elongated cells that fuse end to end with other cells to form the sieve tubes through which substances are transported in phloem and xylem vessels.

Cytology. The living material of most organisms is organized into discrete units termed cells; the study of their features is known as cytology. The cellular contents, when viewed through a microscope at low magnification, usually appear to consist of granules or fibrils of dense material, droplets of fatty substances, and fluid-filled vacuoles suspended in a clear, continuous, semifluid substance called cytoplasm. The remarkable structural complexity of the cell is more fully revealed at the higher magnifications attainable with the electron microscope; structural details of various cellular components, or organelles, as revealed by the technique known as X-ray diffraction analysis, have provided information concerning the relationships between the structures of the cellular components and of the molecules comprising them. Although most cells have certain features in common, the kinds and amounts of components vary considerably. Cellular components include structures such as mitochondria, chloroplasts, endoplasmic reticulum, Golgi complex, lysosomes, oil droplets, granules, and fibrils. The membrane around the cell is a three-layered structure called a unit membrane; similar membranes surround many cellular components—*e.g.*, the mitochondria.

A small spherical or oval organelle, the nucleus, is typically found near the centre of a cell. The genes within the nucleus control the development of the various traits of the cell by controlling the synthesis of specific proteins. The nuclear components are separated from those of the cytoplasm by the nuclear membrane. The structure of the nucleolus, a spherical body within the nucleus, is extremely variable in most cells. Although more than one nucleolus may occur in a nucleus, each cell of an animal or plant species has a fixed number of nucleoli. The nucleoli apparently play a role in the synthesis of the ribonucleic acid (RNA) constituent of the cellular components called ribosomes, which function in protein synthesis. Adjacent to the nucleus in the cells of animals and certain lower plants are two small, cylindrical bodies, the centrioles, which, during cell division, separate, migrate to opposite sides of the cell, and organize a structure called a spindle between them.

Within the cytoplasm of both plant and animal cells are components called mitochondria, which may be shaped like spheres, rods, or threads. Each mitochondrion is bounded by a double membrane, the outer layer of which

forms the smooth outer boundary of the mitochondrion; the inner layer, folded repeatedly into shelflike folds called cristae, contains enzymes that play an essential role in the conversion of the energy of foodstuffs into the energy used for cellular activities. The cells of most plants contain plastids, small bodies involved in the synthesis and storage of foodstuffs. The most important plastids, the chloroplasts, function in trapping the energy of sunlight during photosynthesis. They are disk-shaped structures with a platelike arrangement of tightly stacked membranes.

The cytoplasmic components important in protein synthesis, the ribosomes, are composed of nucleic acid and protein. Clusters of five or more ribosomes, termed polysomes, appear to be the functional unit in protein synthesis.

Lysosomes are membrane-bound structures containing a variety of enzymes that can break down the large molecular constituents of the cell. The membrane surrounding lysosomes presumably prevents the enzymes from digesting the cell contents before the cell dies.

Embryology. The structures and the relationships among the various parts of a mature plant or animal are usually better understood if the successive developmental stages are studied. Thus, morphologists have traditionally been interested in the study of embryos and their developmental patterns—*i.e.,* the science of embryology.

Development typically begins in animals with the cleavage, or division, of the fertilized egg (zygote) to form a hollow ball of cells called the blastula; the blastula then develops into a hollow cuplike body of two layers of cells, the gastrula, from which the embryo ultimately is formed. At one time, the techniques available to embryologists enabled them to study only whole embryos at different developmental stages. The science of experimental embryology began during the first half of the 20th century, when microsurgical techniques became available either for the removal and study of certain structures from tiny embryos or for their transplantation to other regions of the embryo. Advances in understanding the mechanism by which biological information is transferred in DNA and the means by which this information results in the production of specific proteins have led to efforts to describe development in biochemical terms. Although hypotheses regarding the reasons for the appearance of a specific enzyme or some other protein at a specific time during development have been formulated and tested, the biochemical basis of morphogenesis itself—that is, the reason for the development of particular structures—has not yet been established.

Development of microsurgical techniques

The development of the seed plant is basically different from that of an animal. The egg cell of a seed plant is retained within the enlarged lower part, or ovary, of the seed-bearing organ (pistil) of a flower; two sperm nuclei pass through a structure called a pollen tube to reach the egg. One sperm nucleus unites with the egg nucleus to form the zygote from which the new plant will develop; the second sperm nucleus unites with two nuclei, called polar nuclei, to form a body called a triploid endosperm, the cells of which divide to form a nutritive mass within the seed. The zygote undergoes several cell divisions to form the embryo, which is surrounded by the endosperm. The embryo develops one or two seed leaves, or cotyledons, which may become thick and fleshy with stored foodstuffs. The epicotyl, which consists of a growing point enclosed by a pair of folded miniature leaves, develops above the point of attachment of the seed leaves. Below the seed leaves extends the hypocotyl, the tip, or radicle, of which forms the primary root of the embryonic plant. The factors involved in initiating and controlling morphogenesis in plants have been studied by growing cells, tissues, and organs derived from plants. Indeed, an entire carrot plant has been grown from one cell of a mature carrot; this provides striking evidence that the cell from the adult plant contains all of the genetic information needed to produce an entire plant, including roots, stems, and leaves. The technique of growing plants from isolated plant parts has been useful in studies involving the characteristics of embryonic growth, the correlated growth of plant parts, and the nature of differentiation and regeneration (the replacement of lost parts).

Methods in morphology. *Chemical techniques.* The methods of investigating gross structure depend on careful dissection, or cutting apart, of an organism and on accurate descriptions of the parts. The study of the structure of tissues and cells has been extended by the techniques of autoradiography and histochemistry. In the former, a tissue is supplied with a radioactive substance and allowed to utilize it for an appropriate period of time, after which the tissue is prepared and placed in contact with a special photographic emulsion. Silver grains in the emulsion in contact with radioactive substances darken; thus, the location of the dark spots indicates the position at which the radioactive substance was concentrated in the tissue. Histochemistry involves the differential staining of cells (*i.e.,* using dyes that stain specific structural and molecular components) to reflect the chemical differences of the constituents. By choosing appropriate dyes, the histochemist is able, for example, to determine the acidity or alkalinity of the chemical compounds comprising cell components. In addition, dyes that stain specific molecular constituents such as glycogen, DNA, RNA, and protein also are used. The histochemist is able to locate a specific enzyme in a thin slice of tissue, to provide the specific substance with which the enzyme reacts to form a product, and to add a compound that reacts with the product to form an insoluble coloured compound the location of which is relatively easy to determine. In this way, information has been obtained about the specific location of enzymes within the cell.

Dissection and description

Microscopic techniques. Histologists and cytologists utilize microscopic techniques—light microscopy, phase contrast microscopy, interference microscopy, polarization microscopy, fluorescent microscopy, and electron microscopy—to investigate certain aspects of cell structure. Phase contrast microscopy is widely used to study the structure of living cells because, with such apparatus, internal structures can be observed without killing and staining the cell. In addition, motion pictures of dividing cells or moving cells can be made using phase contrast microscopy.

The interference microscope involves passing two separate beams of light through the specimen. With the appropriate instrument, the mass of material per unit area of the specimen can be determined, and contour mapping of small objects is possible.

Crystalline or fibrous elements, both of which are characterized by an orderly or layered molecular structure, are studied with a polarizing microscope; the polarizing microscope has been particularly useful in studying the detailed structure of bone.

Fluorescence microscopy

In fluorescence microscopy, the images seen are molecules of fluorescent dyes added to cells that attach to specific cellular components. Appropriate filters are required to insure that only the light of longer wavelength contributes to the image. Fluorescent antibodies have been used to locate specific kinds of proteins and other materials in certain cells of a tissue or in certain regions of a cell. The antibodies are prepared by injecting into a rabbit an antigen (*e.g.,* the protein myosin), which stimulates white blood cells called lymphocytes to synthesize antibodies that react specifically with the antigen. After the antibodies are isolated and purified, the fluorescent dye, fluorescein, becomes attached to them by a chemical reaction. If the fluorescent antibodies are spread over a tissue, they attach specifically to the molecules that stimulated their formation (myosin). The fluorescence microscope reveals the sites containing the antigen–antibody complex as bright luminescent areas in a dark background.

Electron microscopy

In the scanning electron microscope, a moving spot of electrons (negatively charged particles) is used to scan an object and to produce an image similar to that which appears on a television screen. In this manner, photographs with a three-dimensional appearance can be produced. With the transmission electron microscope, a beam of electrons passes through an object, such as a cell, and is focussed on the other side onto a fluorescent screen or a photographic plate. The beam of electrons in the scanning electron microscope is focussed and then scanned across the specimen. The electrons that leave the specimen, which are not necessarily the same electrons that strike it,

are then used to control the beam of a cathode-ray picture tube. Scanning electron microscopes allow photographs to be taken not only of large molecules such as DNA but of very small objects—individual atoms of elements such as uranium or thorium.

(C.A.V.)

PHYSIOLOGY

The word physiology was first used by the Greeks around 600 BC to describe a philosophical inquiry into the nature of things. The use of the term with specific reference to vital activities of healthy humans, which began in the 16th century, also is applicable to many current aspects of physiology. In the 19th century, curiosity, medical necessity, and economic interest stimulated research concerning the physiology of all living organisms. Discoveries of unity of structure and functions common to all living things resulted in the development of the concept of general physiology, in which general principles and concepts applicable to all living things are sought. Since the mid-19th century, therefore, the word physiology has implied the utilization of experimental methods, as well as techniques and concepts of the physical sciences, to investigate causes and mechanisms of the activities of all living things.

Historical background. The philosophical natural history that comprised the physiology of the Greeks has little in common with modern physiology. Many ideas important in the development of physiology, however, were formulated in the books of the Hippocratic school of medicine (before 350 BC), especially the humoral theory of disease in the treatise *De natura hominis* ("On the Nature of Man"). Other contributions were made by Aristotle (*Lykaion,* about 325 BC) and Galen of Pergamum (*c.* AD 130–*c.* 200). Significant in the history of physiology was the teleology of Aristotle, who assumed that every part of the body is formed for a purpose and that function, therefore, can be deduced from structure. The work of Aristotle was the basis for Galen's *De usu partium* ("On the Use of Parts") and a source for many early misconceptions in physiology. The tidal concept of blood flow, the humoral theory of disease, and Aristotle's teleology, for example, led Galen into a basic misunderstanding of the movements of blood that was not corrected until William Harvey's work on blood circulation in the 17th century.

The publication in 1628 of Harvey's *Exercitatio Anatomica de Motu Cordis et Sanguinis in Animalibus* (*An Anatomical Dissertation Upon the Movement of the Heart and Blood in Animals*) usually is identified as the beginning of modern experimental physiology. Harvey's study was based only on anatomical experiments; despite increased knowledge in physics and chemistry during the 17th century, physiology remained closely tied to anatomy and medicine. In 1747 in Berne, Switzerland, Albrecht von Haller, eminent as anatomist, physiologist, and botanist, published the first manual for physiology. Between 1757 and 1766 he published eight volumes entitled *Elementa Physiologiae Corporis Humani* (*Elements of Human Physiology*); all were in Latin and characterized his definition of physiology as anatomy in motion. At the end of the 18th century, Antoine Lavoisier wrote about the physiological problems of respiration and the production of heat by animals in a series of memoirs that still serve as a foundation for understanding these subjects.

Progress in the 19th century

Physiology as a distinct discipline utilizing chemical, physical, and anatomical methods began to develop in the 19th century. Claude Bernard in France; Johannes Müller, Justus von Liebig, and Carl Ludwig in Germany; and Sir Michael Foster in England may be numbered among the founders of physiology as it now is known. At the beginning of the 19th century, German physiology was under the influence of the romantic school of *Naturphilosophie.* In France, on the other hand, romantic elements were opposed by rational and skeptical viewpoints. Bernard's teacher, François Magendie, the pioneer of experimental physiology, was one of the first men to perform experiments on living animals. Both Müller and Bernard, however, recognized that the results of observations and experiments must be incorporated into a body of scientific knowledge, and that the theories of natural philosophers must be tested by experimentation. Many important ideas in physiology were investigated experimentally by Bernard, who also wrote books on the subject. He recognized cells as functional units of life and developed the concept of blood and body fluids as the internal environment (*milieu intérieur*) in which cells carry out their activities. This concept of physiological regulation of the internal environment occupies an important position in physiology and medicine; Bernard's work had a profound influence on succeeding generations of physiologists in France, Russia, Italy, England, and the United States.

Müller's interests were anatomical and zoological, whereas Bernard's were chemical and medical, but both men sought a broad biological viewpoint in physiology rather than one limited to human functions. Although Müller did not perform many experiments, his textbook *Handbuch der Physiologie des Menschen für Vorlesungen* and his personal influence determined the course of animal biology in Germany during the 19th century.

Methods

It has been said that, if Müller provided the enthusiasm and Bernard the ideas for modern physiology, Carl Ludwig provided the methods. During his medical studies at the University of Marburg in Germany, Ludwig applied new ideas and methods of the physical sciences to physiology. In 1847 he invented the kymograph, a cylindrical drum that still is used to record muscular motion, changes in blood pressure, and other physiological phenomena. He also made significant contributions to the physiology of circulation and urine secretion. His textbook of physiology, published in two volumes in 1852 and 1856, was the first to stress physical instead of anatomical orientation in physiology. In 1869 at Leipzig, Ludwig founded the Physiological Institute (*neue physiologische Anstalt*), which served as a model for research institutes in medical schools all over the world. The chemical approach to physiological problems, developed first in France by Lavoisier, was expanded in Germany by Justus von Liebig, whose books on *Organic Chemistry and its Applications to Agriculture and Physiology* (1840) and *Animal Chemistry* (1842) created new areas of study both in medical physiology and agriculture. German schools devoted to the study of physiological chemistry evolved from Liebig's laboratory at Giessen.

The British tradition of physiology is distinct from that of the continental schools. In 1869 Sir Michael Foster became Professor of Practical Physiology at University College in London, where he taught the first laboratory course ever offered as a regular part of instruction in medicine. The pattern Foster established still is followed in medical schools in Great Britain and the United States. In 1870 Foster transferred his activities to Trinity College at Cambridge, England, and a postgraduate medical school emerged from his physiology laboratory there. Although Foster did not distinguish himself in research, his laboratory produced many of the leading physiologists of the late 19th century in Great Britain and the United States. In 1877 Foster wrote a major book (*Textbook of Physiology*), which passed through seven editions and was translated into German, Italian, and Russian. He also published *Lectures on the History of Physiology* (1901). In 1876, partly in response to increased opposition in England to experimentation with animals, Foster was instrumental in founding the Physiological Society, the first organization of professional physiologists. In 1878, again due largely to Foster's activities, the *Journal of Physiology,* which was the first journal devoted exclusively to the publication of research results in physiology, was initiated.

Founding of the Physiological Society

Foster's teaching methods in physiology and a new evolutionary approach to zoology were transferred to the United States. in 1876 by Henry Newell Martin, a professor of biology at Johns Hopkins University in Baltimore, Md. The American tradition drew also on the continental schools. S. Weir Mitchell, who studied under Claude Bernard, and Henry P. Bowditch, who worked with Carl Ludwig, joined Martin to organize the American Physiological Society in 1887, and in 1898 the society sponsored publication of the *American Journal of Physiology.* In 1868 Eduard Pflüger, professor at the Institute of Physiology at Bonn, founded the *Archiv für die gesammte Physiologie,*

which became the most important journal of physiology in Germany.

Physiological chemistry followed a course partly independent of physiology. Müller and Liebig provided a stronger relationship between physical and chemical approaches to physiology in Germany than prevailed elsewhere. Felix Hoppe-Seyler, who founded his *Zeitschrift für physiologische Chemie* in 1877, gave identity to the chemical approach to physiology. The American tradition in physiological chemistry initially followed that in Germany; in England, however, it developed from a Cambridge laboratory founded in 1898 to complement the physical approach started earlier by Foster.

Physiology in the 20th century is a mature science; during a century of growth, physiology became the parent of a number of related disciplines, of which biochemistry, biophysics, general physiology, and molecular biology are the most vigorous examples. Physiology, however, retains an important position among the functional sciences that are closely related to the field of medicine. Although many research areas, especially in mammalian physiology, have been fully exploited from a classical-organ and organ-system point of view, comparative studies in physiology may be expected to continue. The solution of the major unsolved problems of physiology will require technical and expensive research by teams of specialized investigators. Unsolved problems include the unravelling of the ultimate bases of the phenomena of life. Research in physiology also is aimed at the integration of the varied activities of cells, tissues, and organs at the level of the intact organism. Both analytical and integrative approaches uncover new problems that also must be solved. In many instances, the solution is of practical value in medicine or helps to improve the understanding of both human beings and other animals.

Intradisciplinary work. The anatomical and medical origins of physiology still are reflected in university courses and textbooks that concentrate on functional organ systems of animals; *e.g.,* frog, dog, cat, and rat. The trend in physiology, however, is to emphasize function rather than structure; *i.e.,* comprehensive functional specializations such as nutrition, transport, metabolism, and information have replaced earlier structural studies of organ systems. This trend can be explained in part by the fact that the analysis of an organ system typically involves studies at the levels of cells and molecules, and functional emphasis accommodates such studies better than the organ-system approach.

Functional and comparative emphasis

Early in the 20th century, the emphasis on cells as units of function resulted in a view that all physiology is essentially cell physiology and that all teaching therefore should pivot around the properties of cells. In later years successful analyses of cellular mechanisms involving synthesis, control, and inheritance led to similar emphasis at a new and more fundamental level, the molecules that comprise cells. The study of physiology now encompasses molecules, cells, organs, and many types of animals, including man. The comparisons resulting from such studies not only strengthen human physiology but also generate new problems that extend into evolution and ecology. Much of the impetus for comparative physiology has resulted from the economic or medical importance to man of parasites, insects, and fishes.

Most of the physiology of microorganisms and plants developed independently of animal physiology. The concept of comparative biochemistry provided the foundations for a physiology of microorganisms that extended beyond the parasitic forms that are of medical importance and resulted in recognition of the fundamental roles of microorganisms in the biosphere. Botanists and agriculturists explore the physiology of higher plants, but fundamental differences in the modes of life of animals and plants leave little common ground above the molecular and cellular levels. In a little-known textbook, Claude Bernard stated that there is only one way to live, only one physiology of all living things. The goal of general physiology is to abstract this single physiology from the physiologies of all types of organisms. Although common or general features usually are found at the cellular and molecular levels of organization, multicellular structures also are studied. Processes that underlie cell function are emphasized in an approach based on analyses in terms of physical and chemical principles.

Areas of study. *Metabolism.* In the late 19th century the principle of conservation of energy was derived in part from observations that fermentation and muscle contraction are essentially problems in energetics. Biological energetics began with studies that established the basic equation of respiration as:

$$\text{Fuel} + \text{oxygen} \rightarrow \text{carbon dioxide} + \text{water} + \text{heat}.$$

It was realized that the heat produced in fermentation and the work performed during muscle contraction must originate in similar processes, and that fuel in the equation above is a source of potential energy. Early in the 20th century studies of animal calorimetry verified these concepts in man and other animals. Calorimetry studies showed that the energy produced by the metabolism of foodstuffs in an animal equals that produced by the combustion of these foodstuffs outside the body. After these studies, measurement of the basal metabolic rate (BMR) was used in the diagnosis of certain diseases, and data relating the composition of foodstuffs to their value as sources of metabolic energy were obtained.

Calorimetry

Early in the 20th century it was established that measurable amounts of the carbohydrate glycogen are converted to lactic acid in frog muscles contracting in the absence of oxygen. This observation and studies of alcoholic fermentation confirmed that the energy for fermentation or muscle contraction depends on a series of reactions now known as glycolysis. In order to show that the conversion of glycogen to lactic acid could provide the necessary energy for muscular contraction, extremely delicate measurements of the heat produced by contracting muscles were required. As a result of glycolysis studies, adenosine triphosphate (ATP) was recognized as an important molecule in cellular energy transfer and utilization; *e.g.,* movement, generation of electricity, transport of materials across cell membranes, and production of light by cells. Soon it was discovered that a muscle protein called myosin acts as an enzyme (organic catalyst) by liberating the energy stored in ATP and that ATP in turn can modify the physical properties of myosin molecules. It was also shown that a muscle fibre has an elaborate and ordered structure, which is based on a precise arrangement of myosin and another muscle protein called actin.

Glycolysis is an anaerobic process (*i.e.,* it does not require oxygen) and may represent one of the oldest mechanisms for cellular energy transfer, since the process could have evolved before there was free oxygen in the Earth's atmosphere. Most cells, however, derive their energy from a series of reactions involving oxygen and called the Krebs tricarboxylic acid cycle. The enzymes for the cycle are part of the structure of a mitochondrion, which is an elaborate cellular component filled with membranes and often shaped like a very small bean. In the course of the oxidation, three molecules of energy-rich ATP are generated for each oxygen atom used to form a molecule of water. The mitochondrion, therefore, is the cellular site of respiratory combustion first clearly demonstrated in whole animals by Lavoisier.

The ultimate source of foodstuffs used by animals is plants. Early 19th-century studies of photosynthesis were closely related to those of respiration and began with Joseph Priestley's demonstration that plants could restore the air used during respiration or combustion. The most important equations for living things therefore, are mutually inverse. In respiration:

$$\underset{\text{carbohydrate}}{(CH_2O)_n} + \underset{\text{oxygen}}{nO_2} \rightarrow \underset{\text{carbon dioxide}}{nCO_2} + \underset{\text{water}}{nH_2O} + \text{heat}.$$

In photosynthesis:

$$nCO_2 + nH_2O + \text{light} \rightarrow (CH_2O)_n + nO_2.$$

In the 1930s, it was shown that photosynthesis involves splitting hydrogen from water and that the oxygen liberated in photosynthesis comes from water. During the light reactions, light energy is captured by a green pigment

called chlorophyll and used to generate reactive hydrogen and ATP that are used during dark reactions in which carbohydrates and other cell constituents are synthesized.

Integration of classical, chemical, and comparative approaches

The classical fields of organ-system physiology have a role subsidiary to that of cellular metabolism. Feeding and digestion, for example, become a means for the enzyme-catalyzed breakdown of organic compounds into relatively small molecules that can be transported readily; nutrition, therefore, is a way to supply animals with sufficient sources of energy and specific substances that they cannot synthesize. Comparative animal studies, which were of practical importance in the discovery of some vitamins, led also to the general observation that the specific nutrient requirements of animals are consequences of a slow evolutionary deterioration in which synthetic abilities are lost through changes or mutations in hereditary material.

Nutrition and digestion, however, also have been important in obtaining information at the cellular and molecular levels. It was through studies of digestion, for example, that the existence and nature of enzymes were first disclosed clearly. In addition, early recognition of similarities between digestion and fermentation foreshadowed knowledge of the important role of fermentation in cellular metabolism. Finally, the study of vitamin nutrition was closely integrated with that of cellular oxidation, in which certain vitamins play an essential catalytic role.

In intact organisms, the chemical activities of individual cells do not interfere with the functions of the organism. Much of the study of physiology now is concerned with the ways by which cells obtain their nutrients and dispose of their waste products. Knowledge of the mechanism of protein synthesis and its connections with inheritance and cellular control mechanisms have initiated new inquiries into functions at all levels; *i.e.,* cells, organs, and organisms.

Circulation of blood

Transport. Many important advances in surgery and medicine have been based on the physiology of circulation, which was first studied in 1628. The measurement of blood pressure, for example, was introduced on a practicable basis late in the 19th century and has become an important part of medical diagnosis. The physiology of circulation is concerned with the origin of blood pressure in the force of the heartbeat and the regulation of heart rate, blood pressure, and the flow of blood.

Variations in heart rate that led Aristotle to consider the heart as the seat of the emotions—a myth that persists even now—were among the phenomena whose explanation revealed the existence of the autonomic nervous system. Variations in heart rate are less important to the circulatory system, however, than is the ability of the heart to adjust the strength of its beat to meet certain demands of the body.

The peripheral control of blood pressure and blood flow depends upon a maze of interacting control mechanisms, most significant of which are direct control of the diameter of small arterial branches that enlarge or dilate in response to chemical products formed during metabolism. Increased metabolic activity of tissues such as muscles or the intestine, therefore, automatically induces increased blood flow through the dilated vessels. This action, which could result in a fall in blood pressure, is offset by central-reflex controls that constrict arterial branches not dilated as a result of local chemical effects. Certain regions of the skin and the intestines serve as reservoirs for blood that may be diverted to muscles or the brain if necessary. Peripheral control may break down if excessive demands are made upon it in hot weather (heat stroke), during vigorous exercise after meals (muscle cramp), and after extensive loss of blood or tissue damage (wound shock) or extreme emotion with consequent activation of the autonomic nervous system (emotional shock). A remarkable adaptation occurs in air-breathing vertebrates—reptiles, birds, and mammals—which dive for food or protection. During a dive, the flow of blood to all parts of the body except the brain and the heart is reduced substantially. The energy for muscle contraction is provided by the anaerobic process of glycolysis because the oxygen in the blood goes to the brain and heart, which cannot function without a constant supply of oxygen.

Comparative studies have disclosed two major patterns in circulatory systems. Among vertebrates and a few invertebrates—notably annelid worms and cephalopod mollusks—the blood flows entirely in closed channels or vessels, never coming into direct contact with cells and tissues; blood pressure and the velocity of flow are high and relatively constant, and the volume of blood is small. In many invertebrates—especially arthropods and mollusks other than cephalopods—the blood flows for part of its course in large sinuses or lacunae and comes directly into contact with the tissues. Blood pressure and the velocity of flow are low and variable in these invertebrates, and the large volume of blood is comparable to the total volume of all body fluids in vertebrates.

Consideration of the blood as a transport system has centred especially on the transport of oxygen and carbon dioxide. The colour of blood changes as it passes through the lungs; venous blood is dark purple and arterial blood is bright red because of the properties of a blood pigment called hemoglobin. The complete structure of hemoglobin now has been determined, and minute variations in this structure have enabled man to study fundamental questions of heredity at the molecular level. The development of blood-banks and the techniques involved in blood transfusions depend on knowledge of the physical, chemical, and biological properties of blood. These properties include a remarkable diversity of hemoglobin, both among individuals and species and also within an individual during development. In many instances variations in protein composition better adapt a species to its circumstances.

Transport through cell membranes

Studies of membrane transport at the cellular level are an important part of general physiology. Although quantitative theories of diffusion and osmosis that developed around 1900 were applied to cell physiology, a number of phenomena (*e.g.,* movement through membranes of certain ions and other compounds of biological importance) did not behave according to established physical principles. As a result of studies of osmotic and ionic regulation in freshwater animals, the concept of active transport was formulated. Crucial to the acceptance of this concept were studies with frog skin, which can transport sodium ions against chemical and electrical forces; the transport, specific for sodium ions, is dependent on a continuing input of metabolic energy. Efforts have been directed toward establishing a molecular mechanism that may involve an enzyme found in surface membranes of cells. This enzyme breaks down ATP and releases the energy in the molecule only if sodium and potassium ions are present.

Information transfer. The physiology of animals differs from that of plants in the rapid response of animals to stimuli. René Descartes, responsible for the concept of the reflex that dominated neurophysiology for most of its history, thought a sensory impulse was "reflected" from the brain to produce a reaction in muscles. Later studies of the effects of ions on nerves suggested that a nerve must be surrounded by a membrane and that a nerve impulse results from a change in the ability of the membrane to allow passage of potassium ions. When it was shown that nerves are made up of thousands of tiny fibres, which are processes that extend from cells located in the brain or spinal cord, the nerve impulse hypothesis was applied to individual nerve fibres rather than to whole nerves. Electronic technology provided the techniques and giant nerve fibres of squids provided the experimental material that enabled two Nobel prize winners for physiology, Alan Lloyd Hodgkin and Andrew Fielding Huxley, to extend this hypothesis into a theory of the excitation of nerve cells in which sodium ions and potassium ions play principal roles.

Reflex concept

The reflex concept, however, was not dependent on understanding the molecular basis of excitation, conduction, and transmission. Early in the 20th century the role of interaction of nervous centres in controlling muscle contractions was established. The reflex now is conceived as a unit in which nerve impulses initiated in sensory neurons or nerve cells are conducted to a centre in the brain or spinal cord. In the centre, impulses initiated in motor neurons are conducted to muscles and induce a reflex response. Two processes can occur in the centre; one is

associated with central excitatory states, the other with central inhibitory states. The net effect of any stimulus or group of stimuli, therefore, can be interpreted as an interaction of these opposing states in the centre.

After the demonstration that the effects of the vagus nerve in slowing the heart are mediated by a chemical substance, subsequently identified as acetylcholine, the concept of chemical transmission of nervous impulses was extended to the central nervous system. Typically, transmission of excitation from cell to cell is accomplished by the liberation of a chemical transmitter from a nerve ending.

The reflex concept gave rise to premature attempts to develop a psychology based on reflexes. These attempts (behaviourism) were advanced by the Russian I.P. Pavlov's discovery of conditioned responses. Originally known as conditioned reflexes, these responses have been found in most animals with central nervous systems. More complex than simple reflexes, their mechanism has not yet been established with certainty.

Sensory functions at cellular level

The analysis of sensory functions also extends to the cellular level. Sense organs are diverse in structure and sensitivity to specific stimuli. It may be that the common molecular basis for the differences in sensitivity is a change in permeability of a special region of the membrane surrounding a sensory cell. This change in permeability could allow a nerve fibre to become excited and initiate a nerve impulse. Neurophysiology has borrowed from, and contributed to, the information theory used in communications engineering. The function of sense organs is to gather information both from the environment and the organism. The central nervous system integrates this information and translates it into a program of response involving the entire organism. In addition, the brain can store information previously received (memory) and has the ability to initiate actions without obvious external stimulation (spontaneity). Some aspects of memory and integrative function have been modelled in electronic computers; in fact the development of computers was closely connected with the development of ideas about the functions of the central nervous system.

The analytical interpretation of central nervous function remains, however, a complex and difficult field, even though recent progress has brought closer together the study of behaviour in terms of nerve function and behavioral models. Considerable effort now is directed to the localization of brain function. Although specific centres for reception of sensory information and integration of motor programs are known, the integrative functions that tie them together, as well as the functions of memory, are not so well established.

Regulation. The concept of internal regulations is attributed to Claude Bernard, who thought of blood as an internal environment in which cells function; according to Bernard, maintenance of the internal environment at a constant level was a major responsibility of all body functions. Bernard showed in studies of the formation and breakdown of glycogen in the liver that internal organs can secrete materials into the blood. Other investigators demonstrated such a secretion and used the word hormone to describe the substance. One classical study concerned control of the secretion of digestive fluids by the pancreas; an active substance secretin was purified, as have been a number of similar materials from the digestive tract. The field of endocrinology now is a major part of physiology.

Complementary roles of hormones and nerves

The endocrine system complements the nervous system in control and coordination. Hormones, liberated into blood and other body fluids by endocrine glands and transported throughout the body, usually act either on specific target organs or on certain activities of many organs. Nervous coordination most often is concerned with rapid responses of short duration; endocrine coordination, however, usually is involved in slower responses of longer duration. Stationary-state regulation, or homeostasis, depends on the action of hormones at many points. The hormones insulin and glucagon, both formed in specialized endocrine tissue in the pancreas, control the level of sugar in blood. Vasopressin from the pituitary gland at the base of the brain and aldosterone from the adrenal glands near the kidneys control salt and water balance of the blood. Hormonal regulation, however, is not confined to homeostasis. The cyclic events of the female reproductive cycles in mammals, for example, are determined by a complex sequence of endocrine interactions involving hormones of the pituitary gland and the ovary.

Interaction of hormones

The pervasive regulatory action of hormones is part of a large system of interactions to which the term feedback generally is applied. Hormones involved in homeostatic regulation, for example, influence their own secretion. The secretion of certain steroid hormones, which have a significant action on the conversion of amino acids to glycogen, is controlled by another hormone called the adrenocorticotropic hormone (ACTH), which is formed in the anterior pituitary gland. In turn the secretion of ACTH is controlled by a releasing factor formed in the midbrain and liberated from the stalk of the pituitary gland. ACTH liberation normally is controlled by the concentration of steroids in the blood, so that an increase in steroid concentration inhibits ACTH secretion; this negative feedback, however, may be overcome in certain conditions of intense nervous stimulation.

A similar pattern of releasing factors, by which the nervous system interacts with the endocrine system, also is known for other anterior pituitary hormones; *e.g.,* those involved in the reproductive cycle and in responses of the thyroid gland to temperature changes. In addition, neurosecretory cells—nerve cells specialized for endocrine function—liberate hormones (*e.g.,* vasopressin) that act directly on a specific target. Comparative studies show that neurosecretory cells are important in developmental and regulatory functions of most animals. Discrete endocrine glands, however, occur less frequently; in insects and crustaceans, cycles of growth, molting (shedding of the cuticle), and development are controlled by hormones. The identification of insect hormones may be useful in controlling pests through specific interference with processes of growth and development.

(B.T.S.)

TAXONOMY

Taxonomy, the science of classification in a broad sense, is usually restricted to biological classification and specifically to the classification of plants and animals. The term is derived from the Greek *taxis* ("arrangement") and *nomos* ("law"). Taxonomy is, therefore, the methodology and principles of systematic botany and zoology and sets up arrangements of the kinds of plants and animals in hierarchies of superior and subordinate groups.

Popularly, classifications of living organisms arise according to need and are often superficial. Anglo-Saxon terms such as worm and fish have been used to refer, respectively, to any creeping thing—snake, earthworm, intestinal parasite, or dragon—and to any swimming or aquatic thing. Although the term fish is common to the names shellfish, crayfish, and starfish, there are more anatomical differences between a shellfish and a starfish than there are between a bony fish and a man. Vernacular names vary widely. The American robin (*Turdus migratorius*), for example, is not the English robin (*Erithacus rubecula*), and the mountain ash (*Sorbus*) has only a superficial resemblance to a true ash.

Biologists, however, have attempted to view all living organisms with equal thoroughness and thus have devised a formal classification. A formal classification provides the basis for a relatively uniform and internationally understood nomenclature, thereby simplifying cross-referencing and retrieval of information.

The usage of the terms taxonomy and systematics with regard to biological classification varies greatly. American evolutionist Ernst Mayr has stated that "taxonomy is the theory and practice of classifying organisms" and "systematics is the science of the diversity of organisms"; the latter in such a sense, therefore, has considerable interrelations with evolution, ecology, genetics, behaviour, and comparative physiology that taxonomy need not have.

Historical background. People who live close to nature usually have an excellent working knowledge of the elements of the local fauna and flora important to them and also often recognize many of the larger groups of living

things (*e.g.*, fishes, birds, and mammals). Their knowledge, however, is according to need, and such people generalize only rarely.

From the Greeks to the Renaissance. The first great generalizer in classification was Aristotle, who virtually invented the science of logic, of which for 2,000 years classification was a part. Greeks had constant contact with the sea and marine life, and Aristotle seems to have studied it intensively during his stay on the island of Lesbos. In his writings, he described a large number of natural groups, and, although he ranked them from simple to complex, his order was not an evolutionary one. He was far ahead of his time, however, in separating invertebrate animals into different groups and was aware that whales, dolphins, and porpoises had mammalian characters and were not fish. Lacking the microscope, he could not, of course, deal with the minute forms of life.

Aristotle's classification into natural groups

The Aristotelian method dominated classification until the 19th century. His scheme was, in effect, that the classification of a living thing by its nature—*i.e.*, what it really is, as against superficial resemblances—requires the examination of many specimens, the discarding of variable characters (since they must be accidental, not essential), and the establishment of constant characters. These can then be used to develop a definition that states the essence of the living thing—what makes it what it is and thus cannot be altered; the essence is, of course, immutable. The model for this procedure is to be seen in mathematics, especially geometry, which fascinated the Greeks. Mathematics seemed to them the type and exemplar of perfect knowledge, since its deductions from axioms were certain and its definitions perfect, irrespective of whether a perfect geometrical figure could ever be drawn. But the Aristotelian procedure applied to living things is not by deduction from stated and known axioms; rather, it is by induction from observed examples and thus does not lead to the immutable essence but to a lexical definition. Although it provided for centuries a procedure for attempting to define living things by careful analysis, it neglected the variation of living things. It is of interest that the few people who understood Charles Darwin's *Origin of Species* in the mid-19th century were empiricists who did not believe in an essence of each form.

Aristotle and his pupil in botany, Theophrastus, had no notable successors for 1,400 years. In about the 12th century AD, botanical works necessary to medicine began to contain accurate illustrations of plants, and a few began to arrange similar plants together. Encyclopaedists also began to bring together classical wisdom and some contemporary observations. The first flowering of the Renaissance in biology produced, in 1543, Andreas Vesalius' treatise on human anatomy and, in 1545, the first university botanic garden, founded in Padua, Italy. After this time, work in botany and zoology flourished. John Ray summarized in the late 17th century the available systematic knowledge, with useful classifications. He distinguished the monocotyledonous plants from the dicotyledonous ones in 1703, recognized the true affinities of the whales, and gave a workable definition of the species concept, which had already become the basic unit of biological classification. He tempered the Aristotelian logic of classification with empirical observation.

Botany and zoology of the Renaissance

The Linnaean system. Carolus Linnaeus, who is usually regarded as the founder of modern taxonomy and whose books are considered the beginning of modern botanical and zoological nomenclature, drew up rules for assigning names to plants and animals and was the first to use binomial nomenclature consistently (1758). Although he introduced the standard hierarchy of class, order, genus, and species, his main success in his own day was providing workable keys, making it possible to identify plants and animals from his books. For plants he made use of the hitherto neglected smaller parts of the flower.

Linnaeus attempted a natural classification but did not get far. His concept of a natural classification was Aristotelian; *i.e.*, it was based on Aristotle's idea of the essential features of living things and on his logic. He was less accurate than Aristotle in his classification of animals, breaking them up into mammals, birds, reptiles, fishes, insects, and worms. The first four, as he defined them, are obvious groups and generally recognized; the last two incorporate about seven of Aristotle's groups.

The standard Aristotelian definition of a form was by genus and differentia. The genus defined the general kind of thing being described; the differentia gave its special character. A genus, for example, might be "Bird" and the species "Feeding in water," or the genus might be "Animal" and the species "Bird." The two together made up the definition, which could be used as a name. Unfortunately, when many species of a genus became known, the differentia became longer and longer. In some of his books Linnaeus printed in the margin a catch name, the name of the genus and one word from the differentia or from some former name; in this way he created the binomial, or binary, nomenclature. Thus, modern man is *Homo sapiens,* Neanderthal man *Homo neanderthalensis,* the gorilla *Gorilla gorilla,* and so on.

Creation of binary nomenclature

Classification since Linnaeus. Classification since Linnaeus has incorporated newly discovered information and more closely approaches a natural system. When the life history of barnacles was discovered, for example, they could no longer be associated with mollusks because it became clear that they were arthropods (jointed-legged animals such as crabs and insects). Jean-Baptiste Lamarck, an excellent taxonomist despite his misconceptions about evolution, first separated spiders and crustaceans from insects as separate classes; he also introduced the distinction, no longer accepted by all workers as wholly valid, between vertebrates—*i.e.*, those with backbones, such as fishes, amphibians, reptiles, birds, and mammals—and invertebrates, which have no backbones. The invertebrates, defined by a feature they lack rather than by those they have, constitute in fact about 90 percent of the diversity of all animals. The mixed group "Infusoria," which included all the microscopic forms that would appear when hay was let stand in water, was broken up into empirically recognized groups by the French biologist Felix Dujardin. The German biologist Ernst Haeckel proposed the term Protista in 1866 to include chiefly the unicellular plants and animals because he realized that, at the one-celled level, there could no longer be a clear distinction between plants and animals.

The process of clarifying relationships continues—only in 1898 were agents of disease discovered (viruses) that would pass through the finest filters, and it was not until 1935 that the first completely purified virus was obtained. Primitive spore-bearing land plants (Psilophyta) from the Cambrian Period, which dates from 570,000,000 years ago, were discovered in Canada in 1859. The German botanist Wilhelm Hofmeister in 1851 gave the first good account of the alterations of generations in various nonflowering (cryptogamous) plants, on which many major divisions of higher plants are based. The phylum Pogonophora (beardworms) was recognized only in the 20th century.

The immediate impact of Darwinian evolution on classification was negligible for many groups of organisms, and unfortunate for others. As taxonomists began to accept evolution, they recognized that what had been described as natural affinity—*i.e.*, the more or less close similarity of forms with many of the same characters—could be explained as relationship by evolutionary descent. In groups with little or no fossil record, a change in interpretation rather than alteration of classifications was the result. Unfortunately, some authorities, believing that they could derive the group from some evolutionary principle, would proceed to reclassify it. The classification of earthworms and their allies (Oligochaeta), for example, which had been studied by using the most complex organism easily obtainable and by then arranging progressively simple forms below it, was changed after the theory of evolution appeared. The most simple oligochaete, the tiny freshwater worm *Aeolosoma,* was considered to be most primitive, and classifiers arranged progressively complex forms above it. Later, when it was realized that *Aeolosoma* might well have been secondarily simplified (*i.e.*, evolved from a more complex form), the tendency was to start in the middle of the series, and work in both directions. Biassed names for the major subgroups (Archioligochaeta,

Neoligochaeta) were widely accepted, when in fact there was no evidence for the actual course of evolution of this and other animal groups. Groups with good fossil records suffered less from this type of reclassification because good fossil material allowed the placing of forms according to natural affinities; knowledge of the strata in which they were found allowed the formulation of a phylogenetic tree (*i.e.,* one based on evolutionary relationships), or dendrite (also called a dendrogram), irrespective of theory.

The formulation of phylogenetic trees

The long-term impact of Darwinian evolution has been different and very important. It indicates that the basic arrangement of living things, if enough information were available, would be a phylogenetic tree rather than a set of discrete classes. Many groups are so poorly known, however, that the arrangement of organisms into a dendrite is impossible. Extensive and detailed fossil sequences—the laying out of actual specimens—must be broken up arbitrarily. Many groups, especially at the species level, show great geographical variation, so that a simple definition of species is impossible. Difficulties of classification at the species level are considerable. Many plants show reticulate (chain) evolution, in which species form, then subsequently hybridize, resulting in the formation of new species. And because many plants and animals have abandoned sexual reproduction, the usual criteria for the species—interbreeding within a pool of individuals—cannot be applied. Nothing about the viruses, moreover, seems to correspond to the species of higher organisms.

The objectives of biological classification. A classification or arrangement of any sort cannot be handled without reference to the purpose or purposes for which it is being made. An arrangement based on everything known about a particular class of objects is likely to be the most useful for many particular purposes. One in which objects are grouped according to easily observed and described characteristics allows easy identification of the objects. If the purpose of a classification is to provide information unknown to or not remembered by the user but relating to something the name of which is known, an alphabetical arrangement may be best. Specialists may want a classification relating only to one aspect of a subject. A chemist analyzing the essential oils of plants, for instance, is interested only in the oil content of plants and probably requires such information in far greater detail than would anyone else.

Keys and characters in biological classification

Classification is used in biology for two totally different purposes, often in combination, namely, identifying and making natural groups. The specimen or a group of similar specimens must be compared with descriptions of what is already known. This type of classification, called a key, provides as briefly and as reliably as possible the most obvious characteristics useful in identification. Very often they are set out as a dichotomous key with opposing pairs of characters. The butterflies of a region, for example, might first be separated into those with a lot of white on the wings and those with very little; then each group could be subdivided on the basis of other characters. One disadvantage of such classifications, which are useful for well-known groups, is that a mistake may produce a ridiculous answer, since the groups under each division need have nothing in common but the chosen character (*e.g.,* white on the butterfly wings). In addition, if the group being keyed is large or given to great variation, the key may be extremely complex and may rely on characters difficult to evaluate. Moreover, if the form in question is a new one or one that is not in the key (being, perhaps, unrecorded from the region to which the key applies), it may be identified incorrectly. Many unrelated butterflies have a lot of white on the wings—a few swallowtails, the well-known cabbage whites, some of the South American dismorphiines, and a few satyrids. Should identification of an undescribed form of fritillary butterfly containing much white on the wings be desired, the use of a key could result in an incorrect identification of the butterfly. In order to avoid such mistakes, it is necessary to consider many characters of the organism: not merely one aspect of the wings but their anatomy and the features of the various stages in the life cycle.

Unfortunately, little is known about many of the vast variety of living things. In poorly known groups—and most living things are poorly known—the first objective is identification. There are, for example, about 250,000 species of beetles; many are known only from a single specimen of the adult. In such groups the tendency is to produce classifications which, though purporting to be natural ones, are actually dichotomous keys. Although most common earthworms have on each body segment four pairs of special bristles (chaetae) that are used in locomotion, some species have many chaetae arranged in a complete ring around the body on each segment (perichaetine condition). Because the chaetae are an easily observed character, the latter species were once placed together as a natural group, the family Perichaetidae; knowledge of other aspects of earthworm anatomy, however, made it obvious that several different groups had independently evolved the perichaetine condition. Many current so-called natural groups, especially those at the lower levels of classification, are probably not natural at all but are based on some easily observed characters.

A natural classification is advantageous in that it groups together forms that seem fundamentally to be related. Information utilized in the definition of a group thus need not be repeated for each constituent. This provides concision and efficient information storage. A certain amount of prediction is also possible—a new form with a few ascertained characters similar to those of a natural group probably has other similar characters. As long as no difficult intermediary forms are found, all of the different types can be classified into definite discrete categories. Biological classification has progressed from artificial or key classifications to a natural classification; it has also been realized that division into sharply separated groups often is not possible. Formal classification thus sometimes obscures actual relationships.

The taxonomic process. Basically, no special theory lies behind modern taxonomic methods. In effect, taxonomic methods depend on: (1) obtaining a suitable specimen (collecting, preserving and, when necessary, making special preparations); (2) comparing the specimen with the known range of variation of living things; (3) correctly identifying the specimen if it has been described, or preparing a description showing similarities to and differences from known forms, or, if the specimen is new, naming it according to internationally recognized codes of nomenclature; (4) determining the best position for the specimen in existing classifications and determining what revision the classification may require as a consequence of the new discovery; and (5) using available evidence to suggest the course of the specimen's evolution. Prerequisite to these activities is a recognized system of ranks in classifying; recognized rules for nomenclature; and a procedure for verification, irrespective of the group being examined. A group of related organisms to which a taxonomic name is given is called a taxon (plural taxa).

Ranks. The goal of classifying is to place an organism into an already existing group or to create a new group for it, based on its resemblances to and differences from known forms. To this end, a hierarchy of categories is recognized; for example, an ordinary flowering plant, on the basis of gross structure, is clearly one of the higher green plants, not a fungus, bacterium, or animal, it can easily be placed in the Kingdom Plantae (or Metaphyta). If the body of the plant has distinct leaves, roots, a stem, and flowers, it is placed with the other true flowering plants in the division Magnoliophyta (or Angiospermae), one subcategory of the Plantae. If it is a lily, with sword-like leaves, with the parts of the flowers in multiples of three, and with one cotyledon (the incipient leaf) in the embryo, it belongs with other lilies, tulips, palms, orchids, grasses, and sedges in a subgroup of the Magnoliophyta, which is called the class Liliatae (or Monocotyledones).

The hierarchy of categories

In this class, it is placed, rather than with orchids or grasses, in a subgroup of the Liliatae, the order Liliales; this procedure is continued to the species level. Should the plant be different from any lily yet known, a new species is named, as well as higher taxa, if necessary. If the plant is a new species within a well-known genus, a new species name is simply added to the appropriate genus. If the

plant is very different from any known monocot it might require, even if only a single new species, the naming of a new genus, family, order, or higher taxon; there is no restriction on the number of forms in any particular group. The number of ranks that is recognized in a hierarchy is a matter of widely varying opinion. Shown in Table 1 are seven ranks that are accepted as obligatory by zoologists and botanists. In botany, the term division is often used as an equivalent to the term phylum of zoology. The number of ranks is expanded as necessary by using the prefixes sub-, super-, and infra- (*e.g.,* subclass, superorder) and by adding other intermediate ranks, such as brigade, cohort, section, or tribe. Given in full, the zoological hierarchy for the timber wolf of the Canadian subarctic would be as follows:

Kingdom Animalia
Subkingdom Metazoa
Phylum Chordata
Subphylum Vertebrata
Superclass Tetrapoda
Class Mammalia
Subclass Theria
Infraclass Eutheria
Cohort Ferungulata
Superorder Ferae
Order Carnivora
Suborder Fissipeda
Superfamily Canoidea
Family Canidae
Subfamily Caninae
Tribe (none described for this group)
Genus *Canis*
Subgenus (none described for this group)
Species *Canis lupus* (wolf)
Subspecies *Canis lupus occidentalis* (northern timber wolf)

Name endings

Although the name of the species is binomial (*e.g., Canis lupus*) and that of the subspecies trinomial (*C. lupus occidentalis* for the northern timber wolf; *C. lupus lupus* for the northern European wolf), all other names are single words. In zoology, convention dictates that the names of superfamilies end in *-oidea,* and the code dictates that the names of families end in *-idae,* those of subfamilies in *-inae,* and those of tribes in *-ini.* Unfortunately, there are no widely accepted rules for other major divisions of living things because each major group of animals and plants has its own taxonomic history, and old names tend to be preserved. Apart from a few accepted endings, the names of groups of high rank are not standardized and must be memorized.

The discovery of the only living coelacanth fish of the genus *Latimeria,* in 1938, caused virtually no disturbance of the accepted classification, since the suborder Coelacanthi was already well known from fossils. When certain unusual worms were discovered in the depths of the oceans about ten years later, however, it was necessary to create a new phylum, Pogonophora, for them since they showed no close affinities to any other known animals. The phylum Pogonophora, as usually classified, has one class—the animals in the phylum are relatively similar—but there are two orders, several families and genera, and more than 100 species. Both of these examples have been widely accepted by authorities in their respective areas of taxonomy and may be considered stable taxa.

Table 1: Obligatory Hierarchy of Ranks

	animals	plants
Kingdom	Animalia	Plantae
Phylum	Chordata	Tracheophyta
Class	Mammalia	Pteropsida
Order	Primates	Coniferales
Family	Hominidae	Pinaceae
Genus	*Homo*	*Pinus*
Species	*Homo sapiens* (man)	*Pinus strobus* (white pine)

It cannot be too strongly emphasized that there are no explicit taxonomic characters that define a phylum, class, order, or other rank. A feature characteristic of one phylum may vary in another phylum among closely related members of a class, order, or some lower group. The complex carbohydrate cellulose is characteristic of two kingdoms of plants; among animals, however, cellulose occurs only in one subphylum of one phylum. It would simplify the work of the taxonomist if characters diagnostic of phylum rank in animals were always taken from one feature, the skeleton, for example; those of class rank, from the respiratory organs; and so on down the taxonomic hierarchy. Such a system, however, would produce an unnatural classification.

The taxonomist must first recognize natural groups and then decide on the rank that should be assigned them. Are seasquirts, for instance, so clearly linked by the structure of the extraordinary immature form (larva) to the phylum Chordata, which includes all the vertebrates, that they should be called a subphylum; or should their extremely modified adult organization be deemed more important, with the result that seasquirts might be recognized as a separate phylum, albeit clearly related to the Chordata? At present, this sort of question has no precise answer.

"Numerical taxonomy" as a system of quantifying characteristics

Some biologists believe that "numerical taxonomy," a system of quantifying characteristics of taxa and subjecting the results to multivariate analysis, may eventually produce quantitative measures of overall differences among groups, and that agreement can be achieved so as to establish the maximal difference allowed each taxonomic level. Although such agreement may be possible, many difficulties exist. An order in one authority's classification may be a superorder or class in another. Most of the established classifications of the better known groups result from a general consensus among practicing taxonomists. It follows that no complete definition of a group can be made until the group itself has been recognized, after which its common (or most usual) characters can be formally stated. As further information is obtained about the group, it is subject to taxonomic revision.

Nomenclature. Communication among biologists requires a recognized nomenclature, especially for the units in most common use. The internationally accepted taxonomic nomenclature is the Linnaean system, which, although founded on Linnaeus' rules and procedures, has been greatly modified through the years. There are separate international codes of nomenclature in botany (first published in 1901), in zoology (1906), and in microbiology (bacteria and viruses, 1948). The Linnaean binomial system is not employed for viruses. There is also a code, which was established in 1953, for the nomenclature of cultivated plants, many of which are artificially produced and are unknown in the wild.

The codes, the authority for each of which stems from a corresponding international congress, differ in various details, but all include the following elements: the naming of species by two words treated as Latin; a law of priority that the first validly published and validly binomial name for a given taxon is the correct one and that any others must become synonyms; recognition that a valid binomen can apply to only one taxon, so that a name may be used both in botany and in zoology but for only one plant taxon and one animal taxon; that if taxonomic opinion about the status of a taxon is changed, the valid name can change also; and, lastly, that the exact sense in which a name is used be determined by reference to a type. Rules are also given for the obligate categories of the hierarchy and for what constitutes valid publication of a name. Finally, recommendations are given on the process of deriving names.

Linnaeus believed that there were not more than a few thousand genera of living things, each with some clearly marked character, and that the good taxonomist could memorize them all, especially if their names were well chosen. Thus, although the naming of the species might often involve much research, the genus at least could be easily found.

At the present time, in many taxa, the species has a definite biological meaning—it is defined as a group of

Ability to interbreed as a criterion of species

individuals that can breed among themselves but do not normally breed with other forms. Among microorganisms and other groups in which sexual reproduction need not occur, this criterion fails.

In botanical practice, matters more usually resemble the Linnean situation. Many sorts of chromosomal variants (individuals with different arrangements of chromosomes, or hereditary material, which prevent interbreeding) and marked ecotypes (individuals whose external form is affected by the conditions of soil, moisture, and other environmental factors), as well as other forms, that would clearly be classified as separate species by the zoologist may be lumped together unrecognized or considered subspecies by the botanist. Botanists commonly use the terms variety and form to designate genetically controlled variants within plant populations below the subspecies level.

The use of a strictly biological species definition would enormously increase rather than reduce the number of names in use in botany. A recognized species of flowering plant may consist of several "chromosomal races"—*i.e.,* identical in external appearance but genetically incompatible and, thus, effectively separate species. Such various forms are often identifiable only by cytological examination, which requires fresh material and extensive laboratory work. Many botanists have said that there has been so little stability in the accepted nomenclature that further upheavals would be intolerable and render identification impossible for many applied botanists who may not require such refinements. To postpone recognition of such forms, however, will probably cause upheaval in the future.

Some species of birds are widespread over the archipelagos of the Southwest Pacific, where nearly every island may have a form sufficiently distinct to be given some kind of taxonomic recognition. For example, 73 races are currently recognized for the golden whistler (*Pachycephala pectoralis*). Before the realization that species could vary geographically, each island form was named as a separate species (as many of the races of *P. pectoralis* actually were). It is often believed—and often it is only belief rather than fact—however, that all of these now genetically isolated populations arose as local differentiations of a single stock; thus, they are now usually classed in zoological usage as subspecies of one polytypic species. The term

Polytypic species

polytypic indicates that a separate description (and type specimen) is needed for each of the distinct populations, instead of one for the entire species. The use of a trinomial designation for each subspecies (*e.g., Pachycephala pectoralis bougainvillei*) indicates that it is regarded as simply a local representative (in this case, on Bougainville Island in the Solomons) of a more widely distributed species. The decision on whether to consider such island forms as representatives of one species depends partly on whether, in the judgment of the taxonomist, populations from adjacent islands are sufficiently similar to allow free interbreeding.

Verification and validation by type specimens. The determination of the exact organism designated by a particular name usually requires more than the mere reading of the description or the definition of the taxon to which the name applies. New forms, which may have become known since the description was written, may differ in characteristics not originally considered; or later workers may discover, by inspection of the original material, that the original author inadvertently confused two or more forms. No description can be guaranteed to be exhaustive for all time. Validation of the use of a name requires examination of the original specimen. It must, therefore, be unambiguously designated.

At one time an author might have taken his description from a series of specimens, or partly (or even wholly) from other authors' descriptions or figures, as Linnaeus often did. Much of the controversy over the validity of certain names in current use, especially those dating from the late 18th century, stems from the difficulty in determining the identity of the material used by the original authors. In modern practice, a single type specimen must be designated for a new species or subspecies name. The type should always be placed in a reliable public institution, where it can be properly cared for and made available to taxonomists. For many microorganisms, type cultures are maintained in qualified institutions. Because of the short generation time of microorganisms, however, they may actually evolve during storage.

Nomenclature associated with type specimens

A complex nomenclature is applied to the different sorts of type specimens. The holotype is a single specimen designated by the original describer of the form (a species or subspecies only) and available to those who want to verify the status of other specimens. When no holotype exists, as is frequently the case, a neotype is selected and so designated by someone who subsequently revises the taxon; the neotype occupies a position equivalent to that of the holotype. The first type validly designated has priority over all other type specimens. Paratypes are specimens used, along with the holotype, in the original designation of a new form; they must be part of the same series (*i.e.,* collected at the same immediate locality and at the same time) as the holotype. For a taxon above the species level, the type is a taxon of the next lower rank (for a genus, for instance, it is a species); from the level of the genus to that of the superfamily there are rules regarding the formation of a group name from the name of the type group. The genus *Homo* (man) is the type genus of the family Hominidae, for example, and the code forbids its removal from the family Hominidae as long as the Hominidae is treated as a valid family and the name *Homo* is taxonomically valid. Whatever the remainder of its contents, the family that contains the genus *Homo* must be the Hominidae.

Indiscriminate collecting is of little use today, but huge areas of the Earth still are poorly known biologically, at least as far as many groups are concerned, and there remain many groups for which the small number of properly collected and prepared specimens precludes any thorough taxonomic analysis. Even in well-studied groups, such as the higher vertebrates, new methods of analyzing material often necessitate special collecting. The determination of variation within species or populations may necessitate the study of more specimens than are available, even when (as is usual) the specialist can utilize material from many institutions. Usually, collecting is done to fill gaps (in geographical range, geological formations, or taxonomic categories) already brought to light by specialists reviewing the available material. The well-informed collector of living things knows where to go, what to look for, and how to spot anything especially valuable or extraordinary.

The actual techniques of collecting and preserving vary greatly from one group of organisms to another—soil protozoa, fungi, or pines are neither collected nor preserved in the same manner as birds. Some animals can be preserved only in weak alcohol; others macerate (decompose) in it. Certain earthworms "preserved" in weak alcohol simply flow out of their own skins when lifted out. Special methods are used after long experience to preserve characters of special value in taxonomy. Some methods make specimens difficult to observe; this is especially true of material that has to be sectioned or otherwise made into preparations suitable for microscopic observation.

After taxonomic material has been collected and preserved, its value can be lost unless it is accurately and completely labelled. Only rarely is unlabelled or insufficiently labelled material of any use. The taxonomist normally

Labelling the specimens

must know the locality of collection of each specimen (or lot of specimens), often the habitat (*e.g.,* type of forest, marsh, type of seawater), the date, the name of the collector, and the original field number given to the specimen or lot. To this information is added the catalog number of the collection and the sex (if not already determined in the field and if relevant). The scientific identity of the specimen, as determined by an acknowledged specialist, is usually added to the label at the museum. Also included is the name of the specialist who identified the specimen. Later revisions of the classification and additional knowledge of the organism may result in later alterations of the scientific name, but the original labels must still be kept unaltered.

Other information may also be required; for example, the males and females of some insect groups are extremely different in appearance, and males and females of the

same species may have to be identified. The capture of a pair in the wild actually in copulation gives a strong (but, surprisingly, not absolute) indication that the male and female belong to the same species; the labels of each specimen (if they are separated) indicate the specimen with which it was mating.

Evaluating taxonomic characters. Comparison of material depends to some extent on the purposes of the comparison. For mere identification, a suitable key, with attention given only to the characters in it, may be enough in well-known groups. If the form is likely to be a new one, its general position is determined by observing as many characters as possible and by comparing them with the definitions and descriptions in a natural classification. The new specimen is compared with its nearest known relatives, usually with reference to type material. Any character may be of taxonomic use. In general, taxonomists tend to work from preserved material, so that their findings can be checked. For extinct forms, of course, only preserved material (fossils) is available.

Many biochemical, physiological, or behavioral characters may be at least as good as anatomical characters for discriminating between closely related species or for suggesting relationships. There has been a tendency in recent years rather to discount anatomical characters, but when they are obtainable in quantity (as for most plants and animals), they probably represent as large a sample of the effects of the organism's heredity as can be got, short of complete genetic analysis. Enthusiasts in genetics often stress that the only real basis for classification is the actual genotype of each organism—*i.e.,* the hereditary information by which the organism is formed. It is impossible to obtain such information for extinct forms, and the time required to obtain it for most existing ones would be enormous, even if the techniques were available. An important development, however, has been the hybridization technique employing deoxyribonucleic acid (DNA), the substance by which hereditary information is coded. With this technique, it has been possible to determine similarities in parts of DNA molecules from different organisms but not the nature of their differences.

The role of convergent resemblances

In making comparisons, resemblances resulting from convergence must be considered. Whales and bony fishes, for example, have similar body shapes for the same function—progression through water; their internal features, however, are widely different. In this case, the convergence is evident because of the large number of characters that link whales to other mammals and not to the fishes and because the fossil record for the vertebrates provides a fair indication of the actual evolutionary sequence from primitive fishes through primitive amphibians to primitive reptiles, mammal-like reptiles, and mammals. In the absence of a good fossil record it may be difficult or impossible to identify positively a case of convergence; yet it has been asserted that the occurrence of convergence must not be stated unless it has been "proved." To obey this assertion would be to make the method of analysis dictate in part the results achieved.

In some forms, especially internal parasites, great modification has occurred in adapting to a parasitic way of life. The "root system" of the "tumour" (in reality a parasite) found under the abdomen ("tail") of some crabs, for example, penetrates through the crab's body; the parasite is unrecognizable as a close relative of the barnacles (crustaceans not far removed from the crabs themselves) without the free-swimming larval stage, which shows its affinities. Transient or inconspicuous characters may be of great importance in indicating affinities; the complete life cycle of a specimen may have to be observed before its affinities can be determined. Although such characters may be useless for identification and for definition of a natural group if only a few forms in a group show them, they may be of the utmost importance in understanding relationships. Characters are therefore weighted to some extent by the taxonomist according to their utility for different purposes. Any characters intrinsic to the organism can be used in classification. Extrinsic characters, including the position of fossils in a geological sequence and geographical distribution of fossil and recent forms, may force the taxonomist to look more closely at the intrinsic characters.

Weighting of characters

Weighting or nonweighting (*i.e.,* by the degree of importance) of characters has been a subject of great dispute. On the one hand it has been pointed out that weighting is often demonstrably arbitrary and always imprecise. On the other, it has been said that if characters were actually examined without weighting, some obvious cases of extreme convergence would have to be classed with each other instead of in their proper place. A classification based on unweighted characters is called a phenetic one (based on appearances) as opposed to a phyletic one, in which characters are weighted by their presumed importance in indicating lines of descent. The quarrel results in part from a misunderstanding of aims.

At present, the classification of living things is a rough, non-quantitative sketch of their diversity. A properly surveyed map of this diversity would advance classification enormously. If, on such a map, the diving petrels (*Pelecanoides*) of the Southern Hemisphere and the little auk (*Plautus*) of the Northern Hemisphere were closer to each other than to their own phylogenetic relatives (the other petrels, fulmars, and albatrosses; and the guillemots, terns, gulls, and shorebirds, respectively), this would show the extent of their convergence, which is certainly great, but it would not be a reason for combining them in a separate group. In recent years numerical techniques have been developed for estimating overall resemblance or phenetic distance. For these methods, it is necessary to use large numbers of characters taken from each form and, as far as possible, at random; this involves enormous labour. The mathematical techniques are not as yet wholly satisfactory, some having been borrowed from statistical analysis and applied to taxonomic problems without any consideration of whether they were designed to answer the questions asked by the taxonomist.

It is worth noting that if there were a complete fossil record for any group, then simply placing any form nearest to those most like it (which must be its immediate ancestors or descendants) would produce an arrangement in which all cases of parallelism and convergence would be revealed. Since evolution occurs by descent with modification, this arrangement would presumably reflect the greatest use of the information available about the group and thus would also be the most useful general arrangement. For such groups, the phenetic arrangement is the phyletic one also.

Making a classification. When some idea has been obtained of the constituent forms in a group and of the similarity and dissimilarity that they bear to each other, it is necessary to fit a hierarchical system to them. As already indicated, for groups with good fossil records, a dendritic, or branching, arrangement is desired, and classification must be partly arbitrary because of lack of knowledge.

Compact and scattered groups of species

If the taxonomist has two compact groups of species, those within each group agreeing closely with each other in many characters and differing sharply from members of the other group in others, there is no difficulty in classification except in ranking. If each group contains a scattering of forms, any one close to another but the most divergent members in each group less like each other than they are like certain of the other group, breaking up the groups into definite classes becomes arbitrary.

A particularly difficult case arises when these forms also occur in time series: the present-day dogs, cats, hyenas, and other carnivores differ greatly from each other, but at one time their ancestors were much alike; presumably, therefore, they came from one ancestral stock. Paleontologists trace back each taxonomic line and are inclined to carry their separations of taxonomic groups as far backward in time as possible, until the earliest members of related groups are far more like each other than each is to the rest of the later members of the group to which it is assigned. This separation of groups is extreme phyletic splitting, but cutting off a large basal group containing all the primitive members may require arbitrary breaks in the many lines of descent and will obscure the evolutionary relationships. There is no answer to this dilemma except to avoid extremes.

A similar difficulty arises when the same character complex has arisen independently in related lines. The American paleontologist George Gaylord Simpson, for example, has pointed out that mammalian characters such as the single jawbone (dentary) have arisen several times in groups of the extinct mammal-like reptiles. To use Sir Julian Huxley's useful terminology, the definition of the Mammalia expresses a grade of organization (the attainment of a particular level of advancement), not a clade (a single phyletic group or line). Some taxonomists insist that in an evolutionary classification every group must be truly monophyletic—that is, spring from a single ancestral stock. Usually, this cannot be ascertained; the fossil material is insufficient or, as with many soft-bodied forms, nonexistent. Definite convergence must not be overlooked if it can be detected.

How far groups should be split to show phyletic lines and what rank should be given each group and subgroup thus are matters for reasonable compromise. The resulting classification, if fossils are unknown, may be frankly "natural" or phenetic, as is often explicitly the case with the flowering plants and is actually the case with many animal groups. If sufficient fossils are available, the resulting classification may be consonant with what is known about the evolution of the group or with what is merely conjectured. In reality, many classifications are conjectural or tendentious, and simpler and more natural ones might be closer to the available facts.

Form group classification of fossil remains

Even when only mere fragments are dealt with, a classification of some sort may still be necessary. Large numbers of leaves, some stems, trunks or roots, many seeds, and few flowers are known as fossils and may be of interest to the evolutionist. It may be many decades before a particular sort of fossilized leaf can be associated with a particular sort of branch, let alone trunk, flower, or seed. It is customary to construct form groups (*i.e.,* a genus or species name is assigned to the fossilized material based on its structure) in order to classify fossilized remains and to give them valid binomial names. When (if ever) two or more bits of fossil material are identified as belonging to one organism, one name only is retained. This procedure is best known for plants, but one phylum of animals (the Conodonta) is made up of enigmatic structures that are obviously some part of something animal. (A.J.Ca.)

Current systems of classification. *Division of organisms into kingdoms.* As long as the only known plants were those that grew fixed in one place and all known animals moved about and took in food, the greater groups of organisms were obvious. Even in the time of Linnaeus, however, many biologists wondered about such animal groups as corals and sponges, which were fixed in position and in some ways even flowerlike. Were they zoophytes—animal-plants—intermediate between the two kingdoms?

A more serious problem of classification arose with the invention of the microscope and the discovery of microscopic forms of life. It became apparent that many of these microorganisms held both animal and plant characteristics and could not simply be classified in either kingdom. For example, *Euglena* is a unicellular organism with chlorophyll characteristic of a plant, yet with such animal features as an eyespot and locomotion by means of a flagellum.

Some microorganisms are parasitic inside animals and ingest complicated materials as food, while related microorganisms obtain their nutrients through photosynthesis. It has been proposed that the unicellular forms of microorganisms be placed in a separate kingdom, the Protista. Some biologists do not find this to be a happy solution, however, as some of the "unicellular" plants occur in "colonies" of various numbers of cells and may even have specialized reproductive cells.

In the mid-20th century, biologists recognized two vastly different cell types, procaryote (prokaryote) and eucaryote (eukaryote), and based a division of the living and extinct world on these two broad categorizations. The divisions were based primarily on the absence or presence, respectively, of a membrane-bound nucleus containing the genetic material of the cell, as well as on other organizational and structural features. Many classifications of living organisms adopted such a division and further created two superkingdoms, Prokaryota and Eukaryota. Within the Prokaryota was placed the kingdom Monera (the bacteria, blue-green algae, and a recently described bacterial group called the Archaebacteria [also called Archaeobacteria]). The Eukaryota comprised all other living organisms.

Viruses are far more difficult to classify. They are known only as parasites; no free-living forms have been found. They have a far simpler structure than bacteria and reproduce by injecting their hereditary material, which is either deoxyribonucleic acid (DNA) or ribonucleic acid (RNA) but not both (as in all other living things), into cells of other organisms. In effect, viruses utilize the host's protein-synthesizing mechanism to reproduce. The individual virus particle (virion), therefore, does not grow and divide by fission as do bacteria. Some biologists have speculated that viruses are genes that have gotten out of control and become parasitic; others have denied that viruses can be considered living at all. Many are highly important disease producers in plants, animals, and bacteria.

The principal characteristic shared by bacteria and viruses is that the hereditary material is not contained within a special nuclear membrane. Such a procaryotic condition might be postulated by evolutionists as primitive when compared with forms with a complex nucleus, as in eucaryotic organisms. Viruses, as they now exist, may be the simplest of living things, but it is not known how much they are modified from ancestral forms that are assumed not to have been parasitic and that were evidently on the main line of evolution; nor is their relation to bacteria known.

The blue-green algae

Another procaryotic group, the blue-green algae, is traditionally placed with the other algae (*e.g.,* seaweeds) and studied more by botanists than by microbiologists. Blue-green algae may be either unicellular or filamentous, and they behave like true plants, photosynthesizing in a way that resembles green plants rather than bacteria. Many move by gliding, as do some bacteria and some true unicellular algae. They are often extremely abundant around hot springs or at the edges of muddy ponds, and, though they are resistant to harsh environments, blue-green algae are killed by many drugs (*e.g.,* antibiotics) used against bacteria. Perhaps they are best regarded as representing a group close to the main evolutionary line that gave rise to the eucaryotic plants.

Another problem relates to the position of the fungi, a large group including such familiar forms as mushrooms, toadstools, molds, and yeasts. (Although some authorities place the true slime molds [Myxomycetes] with the fungi, others point to the many characteristics they share with the protists.) The fungi are eucaryotic, lack chlorophyll (and therefore cannot photosynthesize as do green plants), and have rigid walls to the "cells," or filaments (hyphae) that sometimes contain cellulose, as do green plants. Some fungi walls or filaments are made of chitin, the major constituent of the external skeleton of insects and other arthropods, or even of other structural compounds. A fungal "cell" usually contains many nuclei. Asexual and sexual spores are usually produced; some produce motile spores with flagella, like the spores of some algae. The sexual cycle is often very complex. Because fungi in general grow and produce "fruit" like ordinary plants, they have traditionally been included with them; but the differences between the fungi and the plants seem considerable.

The preceding considerations exemplify the difficulties inherent in producing a generally accepted classification, even at the highest levels. Since the earliest attempts at classifying the living world into two kingdoms, Plantae and Animalia, biologists have debated the relationships among all organisms. Most biologists, however, accept the fundamental differences in cell structure that separates the superkingdoms Eukaryota and Prokaryota.

The two-kingdom classification of organisms has not been a suitable alternative since the discovery of a microscopic group of organisms. One four-kingdom classification (Table 2) recognizes the kingdoms Virus, Monera, Plantae, and Animalia within the superkingdoms Prokaryota and Eukaryota. Separate kingdoms are not recognized for the microorganisms (Protista) or for the fungi, which

Table 2: The Four-Kingdom Scheme of Classification

kingdom	members
Virus	
Monera	bacteria, blue-green algae, archaebacteria, and prochlorophytes
Plantae	algae, slime molds, true fungi, bryophytes (mosses, liverworts, and hornworts), ferns, psilophytes, lycopodiophytes, conifers, gnetophytes, ginkgophytes, cycads, and flowering plants
Animalia	protozoans, sponges, corals, flatworms, tapeworms, arthropods, mollusks, lamp shells, annelids, bryozoans, echinoderms, hemichordates, and chordates, including the vertebrates

are placed in the plant kingdom. Another classification recognizes Protista (including the fungi and protozoans) rather than viruses.

A classification of living organisms. Recent advances in biochemical and electron microscopic techniques, as well as in testing that investigates the genetic relatedness among species, have redefined previously established taxonomic relationships and have fortified support for a five-kingdom classification of living organisms. This alternative scheme is presented below and is used in the major biological articles in the *Macropædia.* In it, the procaryotic Monera continue to comprise the bacteria, although techniques in genetic homology have defined a new group of bacteria, the Archaebacteria, that some biologists believe may be as different from bacteria as bacteria are from other eucaryotic organisms. The eucaryotic kingdoms now include the Plantae, Animalia, Protista, and Fungi, or Mycota.

The protists are predominantly unicellular, microscopic, nonvascular organisms that do not generally form tissues. Exhibiting all modes of nutrition, protists are frequently motile organisms, primarily using flagella, cilia, or pseudopodia. The fungi, also nonvascular organisms, exhibit an osmotrophic type of heterotrophic nutrition. Although the mycelium may be complex, they also exhibit only simple tissue differentiation, if any at all. Their cell walls usually contain chitin, and they commonly release spores during reproduction. The plants are multicellular, multitissued, autotrophic organisms with cellulose-containing cell walls. The vascular plants possess roots, stems, leaves, and complex reproductive organs. Their life cycle shows an alternation of haploid (gametophyte) and diploid (sporophyte) generations. The animals are multicellular, multitissued, heterotrophic organisms whose cells are not surrounded by cell walls. Animals generally are independently motile, which has led to the development of organ and tissue systems. The monerans, the only procaryotic kingdom in this classification scheme, is principally made up of the bacteria. They are generally free-living, unicellular organisms that reproduce by fission. Their genetic material is concentrated in a non-membrane-bound nuclear area. Motility in bacteria is by a flagellar structure that is different from the eucaryotic flagellum. Most bacteria have an envelope that contains a unique cell wall, peptidoglycan, the chemical nature of which imparts a special staining property that is taxonomically significant (*i.e.,* gram-positive, gram-negative, acid-fast).

The use of "division" by botanists and "phylum" by zoologists for equivalent categories leads to a rather awkward situation in the Protista, a group of interest to both botanists and zoologists. As used below, the terms follow prevailing usage: phylum for the primarily animal-like protozoa and division for other protistan groups that are more plantlike and of interest primarily to botanists.

The discussion above shows the difficulty involved in classification; for example, one traditional classification of the Aschelminthes, presented below and in the *Macropædia* article ASCHELMINTHS, divides the phylum Aschelminthes into five classes: Rotifera, Gastrotricha, Kinorhyncha, Nematoda, and Nematomorpha. An alternative classification elevates these classes to phyla, and still another classification establishes different relationships between the groups: phylum Gastrotricha, phylum Rotifera, phylum Nematoda (containing classes Adenophorea, Secernentea, and Nematomorpha), and phylum Introverta (containing classes Kinorhyncha, Loricifera, Priapulida, and Acanthocephala). The true relationships between these pseudocoelomates remain to be established.

KINGDOM MONERA (bacteria, archaebacteria, and blue-green algae)

KINGDOM PROTISTA (protists)
Algae other than the blue-green algae, protozoa, and slime molds.

KINGDOM FUNGI (fungi)

KINGDOM PLANTAE (METAPHYTA or EMBRYOPHYTA; nonvascular and vascular plants)
Includes mosses, liverworts, hornworts, whisk ferns, club mosses, horsetails, ferns, cycads, conifers, gnetophytes, ginkgophytes, and flowering plants.

Division Bryophyta (mosses, liverworts, and hornworts)

Division Psilotophyta (whisk ferns or psilopsids)

Division Lycophyta (club mosses and quillworts)

Division Sphenophyta (horsetails)

Division Polypodiophyta (ferns)

Division Coniferophyta
Includes pines, yews, spruces, firs, junipers, redwoods, and others.

Division Ginkgophyta (ginkgoes)

Division Cycadophyta (cycads)

Division Gnetophyta (gnetophytes)

Division Magnoliophyta (flowering plants)
Includes monocots (grasses, rushes, sedges, cattails and pondweeds, palms, pineapple and other bromeliads, lilies, bananas, ginger, orchids, and others) and dicots or broad-leaved plants (most trees, buttercups, poppies, roses, violets, cacti, mints, squashes, sunflowers, and many others).

KINGDOM ANIMALIA (or METAZOA)

Subkingdom Parazoa (sponges)

Phylum Porifera (sponges)

Subkingdom Eumetazoa

Phylum Mesozoa (mesozoans)

Phylum Cnidaria (or *Coelenterata;* cnidarians)

Phylum Ctenophora (ctenophores)

Phylum Platyhelminthes (flatworms)

Phylum Nemertea (or *Rhynchocoela;* ribbonworms)

Phylum (or class) Acanthocephala (spiny-headed worms)

Phylum Aschelminthes

Phylum Priapulida (priapulids)

Phylum Annelida (annelid worms)

Phylum Tardigrada

Phylum Onychophora

Phylum Arthropoda (arthropods)

Phylum Mollusca (mollusks)

Phylum Bryozoa (or *Ectoprocta;* bryozoans)

Phylum Phoronida (phoronid worms)

Phylum Brachiopoda (brachiopods)

Phylum Sipuncula (sipunculid worms)

Phylum Chaetognatha (arrowworms)

Phylum Echiurida (spoonworms)

Phylum Echinodermata (echinoderms)

Phylum Hemichordata (hemichordates)

Phylum Pogonophora (beardworms)

Phylum Chordata (chordates)

(Ed.)

BIOPHYSICS

The birth of biophysics

In its broadest sense, biophysics is concerned with the solution of biological problems in terms of the concepts of physics and other physical sciences. The relatively recent emergence of biophysics as a scientific discipline may be attributed, in particular, to the spectacular success of biophysical tools in unravelling the molecular structure of deoxyribonucleic acid (DNA), the fundamental hereditary material, and in establishing the precisely detailed

structure of proteins such as hemoglobin in order that the position of each atom may be known. Biophysics and the intimately related subject molecular biology now are firmly established as cornerstones of modern biology.

Historical background. The origin of biophysics antedates the division of natural sciences into separate disciplines. Bioluminescence must be considered among the most ancient objects of biophysical exploration, because the emission of light by living organisms has long stimulated the curiosity of natural philosophers. Perhaps the first scientific investigation of animal luminescence was that of Athanasius Kircher, a 17th-century German Jesuit priest, who devoted two chapters of his book *Ars Magna Lucis et Umbrae* to bioluminescence. In the midst of his more scientific observations, Kircher found time to expose as a fallacy the notion that an extract made from fireflies could be used to light houses.

The relation between electricity and biology became a subject of speculation in the 17th century and one of intense exploration in the 18th and 19th. Sir Isaac Newton in the *Principia* (1687) wrote of "a certain most subtle spirit which pervades and lies hid in all gross bodies," and that "all sensation is excited, and the members of animal bodies move at the command of the will, namely, by the vibrations of this spirit, mutually propagated along the solid filaments of the nerves, from the outward organs of sense to the brain, and from the brain into the muscles." Man's fascination with animal electricity is illustrated in a letter written by John Walsh in 1773 to the American inventor and statesman Benjamin Franklin; Walsh wrote the details of his discovery of the electrical nature of the discharge from the torpedo or electric ray:

> I am concerned that other engagements have prevented me from giving to the Royal Society, before their recess, a complete account of my experiments on the electricity of the torpedo; a subject not only serious in itself, but opening a large field of interesting inquiry, both to the electrician in his walk of physics, and to all who consider, particularly or generally, the animal oeconomy.

Typical of the unity of science that then prevailed were the advances sometimes made either by professors of physics who were interested in biological phenomena or professors of anatomy, a subject that at that time included physiology. Thus Abbé Giovanni Beccaria, professor of physics in Turin and Italy's leading student of electricity in the mid-18th century, carried out experiments on the electrical stimulation of muscles. Albrecht von Haller, professor of anatomy and surgery at Göttingen, discussed "the nervous fluid" and conjectured as to whether "electrical matter" and "animal spirits" were the same. In 1786 Luigi Galvani, a physician in Bologna, made the crucial experiment that helped end this controversy. Galvani supposedly was performing experiments with a machine in the company of friends, when, by chance, one member of the party idly probed with a knife the nerves of the thigh of a skinned frog to be used for soup. As the muscles of the frog leg suddenly and unexpectedly contracted, Galvani's wife noted that a spark had been produced by the electrical machine and "fancied that there was an agreement in point of time." Although Galvani's own account of the occurrence differed somewhat in detail from the preceding, it is certain that the experiment was repeated and verified, setting the stage for a long controversy between the advocates of Galvani's view that current generated by an animal can cause contraction and those of Alessandro Volta, who claimed that the frog leg served only as a detector of minute differences in electrical potential external to it. The Galvani partisans performed an experiment in which no external sources of electricity were present, thus proving that current generated by an animal could cause the muscle contraction. But it was also possible to cause contraction by contact with metals; Volta performed such investigations, and they culminated in his invention of the electrical battery, which was so important that it overshadowed Galvani's research. As a result, the study of electrical potential in animals disappeared from scientific consideration until 1827.

Electrical stimulation of muscles

Because for many years the frog leg was the most sensitive detector of differences in electrical potential, final acceptance of the view that currents can be generated by living tissues had to await the construction of galvanometers sensitive enough to measure the minute currents generated in muscles and the small potential differences across nerve membranes. Galvanometers were built by the great German 19th-century electrophysiologist Du Bois-Reymond, professor of physiology in Berlin. His investigations of muscular current and electrical potential of nerves depended upon a galvanometer of his own devising that required 3.17 miles (5.10 kilometres) of wire wound in 24,000 turns. Research in this subject, called neurophysiology, grew in stature with increased understanding of both electrical phenomena and cellular physiology; it served as one point of origin for biophysics.

Biophysics also grew out of investigations on diffusion gradients and osmotic pressure—two forces responsible for the passive flow of matter in living organisms. Osmotic pressure, the pressure that develops in a solution separated from a solvent by a membrane permeable only to solvent, was first described by Abbé J.A. Nollet, who became professor of experimental physics at the College of Navarre. The semipermeable membranes required to produce the fluid flow that characterizes osmotic phenomena initially came from biological sources; French scientist René Dutrochet wrote in 1828, "it appears from these new studies that the endosmotic and exosmotic phenomena, which I discovered, belong to a new class of physical phenomena, whose powerful intervention in the vital phenomenon is no longer doubtful." Following the first quantitative measurements by the botanist W.F.P. Pfeffer, the fundamental laws governing diffusion were enunciated by Adolf Fick, who in 1856 published what is probably the first biophysics text, *Die medizinische Physik* ("Medical Physics"). Fick developed the laws of diffusion not from experiment but by analogy with the laws governing the flow of heat; subsequent laboratory experiments proved the analogy to be quantitatively exact.

The development of physical chemistry

Physical and chemical investigation coalesced in physical chemistry, a subject that began to develop with the emergence of the *Zeitschrift für Physikalische Chemie* in 1887, a journal founded by Dutch chemist Jacobus van't Hoff and German chemist Wilhelm Ostwald. The first volume contains contributions from the most noted physical chemists of the time, including van't Hoff, Ostwald, François Raoult, and Svante Arrhenius. They were concerned with reactions in solution, a central topic in biology because the interior milieu of all living cells is aqueous, and the chemical reactions that sustain life take place in water. The scientific interests of van't Hoff in particular transcended the boundaries between disciplines. He stressed the importance of the laws of osmosis, which he had clearly delineated, to the economy of all living processes.

Biophysics matured in the 20th century. British biophysicist A.V. Hill described the modern biophysicist in these terms:

> Biological phenomena, like many others, show aspects and relations susceptible of physical analysis and interpretation. It is by the choice of problems and by the intellectual processes with which they are formulated and attacked, more than by the particular techniques employed, that a subject can be most clearly defined. There are people to whom physical intuitions come naturally, who can state a problem in physical terms, who can recognize physical relations when they turn up, who can express results in physical terms. These intellectual qualities, more than any special facility with physical instruments and methods, are essential to the make-up of a biophysicist. Equally essential, however, are the corresponding qualities, intuitions, and experience of the biologist. A physicist who cannot develop the biological approach, who has no curiosity about vital processes and functions, who is not willing to spend time in learning the habits of living things, who regards biology simply as a branch of physics has no important future in biophysics. (From *Science,* Dec. 21, 1956.)

Most biophysical research has been carried out by physicists with an interest in biology; therefore, there must be a way by which scientists educated in physics and physical chemistry can find their way into biology and become familiar with problems that may be open to a physical interpretation. Although classically oriented biology departments often offer positions to biophysicists, they are

not substitutes for centres in which biophysical research is of central importance.

Tools of the biophysicist

The biophysicist possesses the ability to separate biological problems into segments that are amenable to exact physical interpretation and to formulate hypotheses that can be tested by experiment. The primary tool of the biophysicist is an attitude of mind. To this might be added the ability to use complex physical theory to study natural objects—for example, that involved in the X-ray diffraction techniques used to determine the structure of large molecules such as proteins. The biophysicist usually recognizes the utility of new physical tools—*e.g.*, nuclear magnetic resonance and electron spin resonance—in the study of specific problems in biology. But he may also, through previous experience in building specialized equipment to solve physical problems, not have to rely on commercially built instruments.

The development of instruments for biological purposes is an important aspect of a new area—applied biophysics. Biomedical instrumentation is probably most widely used in hospitals. Applied biophysics is important in the field of therapeutic radiology, in which the measurement of dose is critical to treatment, and in diagnostic radiology, particularly with techniques involving isotope localization and whole body scanning to aid in tumour diagnosis. As aids in diagnosis and patient care, computers are of increasing importance. Automation of the chemical analyses routinely carried out in hospitals will soon be a reality. The opportunities for the applications of biophysics seem limitless because the lengthy delay between the development of a research instrument and its application means that many scientific instruments based on physical principles already known will be shown to have important potential for medicine.

Interdisciplinary work. The biophysical approach is unified by a consideration of biological problems in the light of physical concepts, so that biophysics is, perforce, interdisciplinary. Biophysics may be thought of as the central circle in a two-dimensional array of overlapping circles, which include physics, chemistry, physiology, and general biology. Relations with chemistry are mediated through biochemistry and chemistry; those with physiology, through neurophysiology and sensory physiology. Biology, which may be viewed as a general subject pervading biophysical study, is evolving from a purely descriptive science into a discipline increasingly devoted to understanding the nature of the prime movers of biological events. The evolution of biology in these directions has received great impetus from the biophysical and biochemical discoveries of the 20th century. An understanding of the physical principles governing biological effects is the proper end of biophysics.

Areas of study. The content and methods of biophysics are illustrated by examining several notable contributions to science.

Protein structure. Within two days after the initial publication of Wilhelm Röntgen's discovery of X-rays in 1895, a surgeon in Scotland used X-rays to observe a needle as he extracted it from the palm of an unfortunate seamstress. Although this medical application resulted in the development of radiological diagnosis and treatment of disease by radiation, physical aspects of Röntgen's discovery also provided the means for elucidating the structure

The significance of X-rays

of proteins and other large molecules. The laws governing the diffraction of X-rays were discovered by the two Braggs, Sir William and Sir Lawrence, who were father and son. At the Cavendish Laboratory at the University of Cambridge, where Sir Lawrence was professor, J.D. Bernal was studying the use of X-ray diffraction for the determination of the structure of large biological molecules. He had already used X-rays to define the size and shape of the tobacco mosaic virus and showed it to have a regular internal structure. At the Cavendish Laboratory the group that formed around Bernal, a man of wide public and scientific interests, included the Nobel Prize winners Max Perutz and John Kendrew, who in 1937 began to use X-rays to analyze two proteins fundamental to life, myoglobin and hemoglobin, both of which function in the transport of gases in the blood. Twenty-two years passed before the structures of these proteins were established; the significance of the work is that it provided the basis for an understanding of the mechanism of the action of enzymes and other proteins, an active and fruitful subject of modern investigation.

Deoxyribonucleic acid. Interest in biophysics at the Cavendish Laboratory resulted in another important discovery, the structure of deoxyribonucleic acid (DNA), the genetic material. This achievement by a British biophysicist, Francis H.C. Crick, and by a U.S. biochemist, James Watson, was based on X-ray data obtained by Maurice Wilkins at King's College, London. When Crick first went to the Cavendish Laboratory for education in biophysics, he worked under Perutz's direction; when Watson went to the Cavendish, he and Crick began the collaboration that led to the establishment of the structure of DNA, for which Watson, Crick, and Wilkins later were awarded a Nobel Prize.

Much impetus for biophysical investigation following World War II came from the desire of physicists to move away from physics and into biology; this drive was strengthened by the publication in 1944 of Erwin Schrödinger's book *What Is Life?* Schrödinger, the Austrian physicist who contributed substantially to the development of wave mechanics, was anxious to determine whether biological events could be accounted for in terms of known laws of physics and chemistry, or whether a full explanation would require the formulation of physical laws not yet known to exist. Because biological reproduction seemed to pose intractable problems, he devoted a chapter of his book to a consideration of the gene. The discussion was based on the model put forward by Max Delbrück, a physicist who had for some years been studying the genetics of viruses that infect bacteria (bacteriophages). Delbrück's summer course on bacteriophages in 1945 at Cold Spring Harbour in New York set in motion the chain of events that led to understanding the genetic code by which the sequence of the nucleotides in DNA is translated into the sequence of amino acids in a protein. The use of bacteriophage also provided an opportunity for experiments with a primitive living organism that could be studied without anatomic complexities. This aspect of biophysics has become more biochemically oriented as it has developed and is now known as molecular biology; sometimes it is considered a distinct discipline, and other times it is subsumed under the biophysical sciences.

The nerve impulse. Important aspects of biophysics have been derived from physiology, especially in studies involving the conduction of nerve impulses. One important scientific product of World War II—the development of vastly improved electronics—largely resulted from radar devices that had been used primarily for locating aircraft. Another product, the atomic bomb, was constructed by way of nuclear reactors that could, in peace time, provide an abundant supply of radioactive isotopes, which are now of great value not only in biophysical research but also in biochemistry and medicine. These two disparate advances were important to the work of two Nobel Prize winners, Alan Hodgkin and Andrew Huxley, who showed how the flow of sodium and potassium across the membranes of nerves can be coupled to produce the action potential, a brief electrical event that initiates the action potential, which propagates the nervous signal.

A model of the nerve axon proposed by Hodgkin and Huxley grew from a 19th-century confluence of ideas. Julius Bernstein, an experimental neurophysiologist, used physical chemical theories to develop a membrane theory of nervous conduction; Hodgkin's initial experiments were designed to test specific predictions of the Bernstein hypothesis. Early in 1938 Hodgkin learned of the important results of a newly developed technique that allowed examination of the time course of nervous conduction. After World War II, Hodgkin, joined by Huxley, again took up the research. They presented their explanation of the mechanism of nervous conduction in five scientific papers between October 1951 and March 1952.

Biological membranes. The availability of radioactive isotopes provided the technology necessary for understanding how molecules are transported across biological

membranes, which are the very thin boundaries of living cells; the environment maintained by membranes in cells differs from the external environment and permits cellular function. The Danish physiologist August Krogh laid the groundwork in this subject; his pupil, Hans Ussing, developed the conceptual means by which the transport of ions (charged atoms) across membranes can be identified. Ussing's definition of active transport made possible an understanding, at the cellular level, of the way in which ions and water are pumped into and out of living cells in order to regulate the ionic composition and water balance in cells, organs, and organisms. The molecular mechanism by which these processes occur, however, remains to be discovered.

Conceptual explanation of active transport

In addition to the function of transport, membranes also are utilized as templates on which such molecules as enzymes, which must function in a sequential fashion, can be kept in the requisite order. Although great progress has been made in understanding the mechanisms by which specific atoms are assembled into large biological molecules, the principles involved in the assembly of molecules into membranes, which are organized structures of a higher degree of complexity than large molecules, are not yet very well understood. There is reason to believe that the incorporation of a molecule into a membrane endows it with properties that differ from those of a molecule in solution. A primary task of biophysics is to understand the physical character of these cooperative interactions that are essential to life.

Muscle contraction. A.V. Hill developed exquisitely sensitive temperature sensors for measuring heat generated during muscular contraction; he initiated studies relating this heat to the thermodynamic parameters responsible for it. The electron microscope in the years following World War II made possible the description of muscular contraction at a structural level, though the mechanisms involved in the flow of heat during the process are not yet known. Simultaneously, in the 1960s, but independently, various physicists postulated the sliding-filament theory of muscular contraction, according to which muscles contract by the sliding of one filament along another and not by a springlike coiling. Remarkable advances, based on the use of techniques such as X-ray diffraction and electron microscopy, have made it possible to visualize many of the molecules involved in the process. The entire process of muscular contraction, in terms of an identification of the molecules and a description of the chemical reactions in the muscle fibre, has been almost completely explained.

Sensory communication. The above comprise a few specific examples of the scope of biophysics. One area, difficult to discuss in specific terms, is that of sensory communication. Because stimuli, particularly those of a visible or auditory nature, can easily be specified in exact physical terms, they have excited the interest of physical scientists since before 1850. Modern electronic techniques make it relatively easy to distinguish true signals from noise; in addition, computers make possible the performance of significant experiments concerning the complex relationship between stimulus and action. Quantitative analysis of sensory response is very difficult, however, because it involves a synthesis of the action of many cells. It has been pointed out that

> An adequate theory of sensory function implies an adequate theory of brain function. And an adequate theory of brain function in its turn requires that the nervous system's behavioral repertory be predictably related to the behaviour of the elements that compose it.

(A.K.S.)

BIOCHEMISTRY

Biochemistry is the study of substances found in living organisms and of the changes they undergo during development and life of the organism. It deals with the chemistry of life, and as such it draws on the techniques of analytical, organic, and physical chemistry, as well as those of physiologists concerned with the molecular basis of vital processes. All chemical changes within the organism—either the degradation of substances, generally to gain necessary energy, or the buildup of complex molecules necessary for life processes—are collectively termed metabolism. These chemical changes depend on the action of organic catalysts known as enzymes, and enzymes, in turn, depend for their existence on the genetic apparatus of the cell. It is not surprising, therefore, that biochemistry enters into the investigation of chemical changes in disease, drug action, and other aspects of medicine, as well as in nutrition, genetics, and agriculture.

The term biochemistry is synonymous with two somewhat older terms: physiological chemistry and biological chemistry. Those aspects of biochemistry that deal with the chemistry and function of very large molecules (*e.g.*, proteins and nucleic acids) are often grouped under the term molecular biology. Biochemistry is a young science, having been known under that term only since about 1900. Its origins, however, can be traced much further back; its early history is part of the early history of both physiology and chemistry.

Historical background. The particularly significant past events in biochemistry have been concerned with placing biological phenomena on firm chemical foundations.

Prelude to biochemistry

Before chemistry could contribute adequately to medicine and agriculture, however, it had to free itself from immediate practical demands in order to become a pure science. This happened in the period from about 1650 to 1780, starting with the work of Robert Boyle and culminating in that of Antoine-Laurent Lavoisier, the father of modern chemistry. Boyle questioned the basis of the chemical theory of his day and taught that the proper object of chemistry was to determine the composition of substances. His contemporary John Mayow observed the fundamental analogy between the respiration of an animal and the burning, or oxidation, of organic matter in air. Then, when Lavoisier carried out his fundamental studies on chemical oxidation, grasping the true nature of the process, he also showed, quantitatively, the similarity between chemical oxidation and the respiratory process. Photosynthesis was another biological phenomenon that occupied the attention of the chemists of the late 18th century. The demonstration, through the combined work of Joseph Priestley, Jan Ingenhousz, and Jean Senebier, that photosynthesis is essentially the reverse of respiration was a milestone in the development of biochemical thought.

In spite of these early fundamental discoveries, rapid progress in biochemistry had to wait upon the development of structural organic chemistry, one of the great achievements of 19th-century science. A living organism contains many thousands of different chemical compounds. The elucidation of the chemical transformations undergone by these compounds within the living cell is a central problem of biochemistry. Clearly, the determination of the molecular structure of the organic substances present in living cells had to precede the study of the cellular mechanisms, whereby these substances are synthesized and degraded.

There are few sharp boundaries in science, and the boundaries between organic and physical chemistry, on the one hand, and biochemistry, on the other, have always shown much overlap. Biochemistry has borrowed the methods and theories of organic and physical chemistry and applied them to physiological problems. Progress in this path was at first impeded by a stubborn misconception in scientific thinking—the error of supposing that the transformations undergone by matter in the living organism were not subject to the chemical and physical laws that applied to inanimate substances and that consequently these "vital" phenomena could not be described in ordinary chemical or physical terms. Such an attitude was taken by the vitalists, who maintained that natural products formed by living organisms could never be synthesized by ordinary chemical means. The first laboratory synthesis of an organic compound, urea, by Friedrich Wöhler in 1828, was a blow to the vitalists but not a decisive one. They retreated to new lines of defense, arguing that urea was only an excretory substance—a product of breakdown and not of synthesis. The success of the organic chemists in synthesizing many natural products forced further retreats of the vitalists. It is axiomatic in modern biochemistry that the chemical laws that apply to inanimate materials are equally valid within the living cell.

At the same time that progress was being impeded by a misplaced kind of reverence for living phenomena, the practical needs of man operated to spur the progress of the new science. As organic and physical chemistry erected an imposing body of theory in the 19th century, the needs of the physician, the pharmacist, and the agriculturalist provided an ever-present stimulus for the application of the new discoveries of chemistry to various urgent practical problems.

The contributions of Liebig and Pasteur

Two outstanding figures of the 19th century, Justus von Liebig and Louis Pasteur, were particularly responsible for dramatizing the successful application of chemistry to the study of biology. Liebig studied chemistry in Paris and carried back to Germany the inspiration gained by contact with the former students and colleagues of Lavoisier. He established at Giessen a great teaching and research laboratory, one of the first of its kind, which drew students from all over Europe.

Besides putting the study of organic chemistry on a firm basis, Liebig engaged in extensive literary activity, attracting the attention of all scientists to organic chemistry and popularizing it for the layman as well. His classic works, published in the 1840s, had a profound influence on contemporary thought. Liebig described the great chemical cycles in nature. He pointed out that animals would disappear from the face of the Earth if it were not for the photosynthesizing plants, since animals require for their nutrition the complex organic compounds that can be synthesized only by plants. The animal excretions and the animal body after death are also converted by a process of decay to simple products that can be re-utilized only by plants.

In contrast with animals, green plants require for their growth only carbon dioxide, water, mineral salts, and sunlight. The minerals must be obtained from the soil, and the fertility of the soil depends on its ability to furnish the plants with these essential nutrients. But the soil is depleted of these materials by the removal of successive crops; hence the need for fertilizers. Liebig pointed out that chemical analysis of plants could serve as a guide to the substances that should be present in fertilizers. Agricultural chemistry as an applied science was thus born.

In his analysis of fermentation, putrefaction, and infectious disease, Liebig was less fortunate. He admitted the similarity of these phenomena but refused to admit that living organisms might function as the causative agents. It remained for Pasteur to clarify that matter. In the 1860s Pasteur proved that various yeasts and bacteria were responsible for "ferments," substances that caused fermentation and, in some cases, disease. He also demonstrated the usefulness of chemical methods in studying these tiny organisms and was the founder of what came to be called bacteriology.

Later, in 1877, Pasteur's ferments were designated as enzymes, and, in 1897, the German chemist E. Buchner clearly showed that fermentation could occur in a press juice of yeast, devoid of living cells. Thus a life process of cells was reduced by analysis to a nonliving system of enzymes. The chemical nature of enzymes remained obscure until 1926, when the first pure crystalline enzyme (urease) was isolated. This enzyme and many others subsequently isolated proved to be proteins, which had already been recognized as high-molecular-weight chains of subunits called amino acids.

The mystery of how minute amounts of dietary substances known as the vitamins prevent diseases such as beriberi, scurvy, and pellagra became clear in 1935, when riboflavin (‡vitamin B_2) was found to be an integral part of an enzyme. Subsequent work has substantiated the concept that many vitamins are essential in the chemical reactions of the cell by virtue of their role in enzymes.

In 1929 the substance adenosine triphosphate (ATP) was isolated from muscle. Subsequent work demonstrated that the production of ATP was associated with respiratory (oxidative) processes in the cell. In 1940 F.A. Lipmann proposed that ATP is the common form of energy exchange in many cells, a concept now thoroughly documented. ATP has been shown also to be a primary energy source for muscular contraction.

Isotopes as tools

The use of radioactive isotopes of chemical elements to trace the pathway of substances in the animal body was initiated in 1935 by two U.S. chemists, R. Schoenheimer and D. Rittenberg. That technique provided one of the single most important tools for investigating the complex chemical changes that occur in life processes. At about the same time, other workers localized the sites of metabolic reactions by ingenious technical advances in the studies of organs, tissue slices, cell mixtures, individual cells, and, finally, individual cell constituents, such as nuclei, mitochondria, ribosomes, lysosomes, and membranes.

In 1869 a substance was isolated from the nuclei of pus cells and was called nucleic acid, which later proved to be deoxyribonucleic acid (DNA), but it was not until 1944 that the significance of DNA as genetic material was revealed, when bacterial DNA was shown to change the genetic matter of other bacterial cells. Within a decade of that discovery, the double helix structure of DNA was proposed by Watson and Crick, providing a firm basis for understanding how DNA is involved in cell division and in maintaining genetic characteristics.

Advances have continued since that time, with such landmark events as the first chemical synthesis of a protein, the detailed mapping of the arrangement of atoms in some enzymes, and the elucidation of intricate mechanisms of metabolic regulation, including the molecular action of hormones.

Areas of study. A description of life at the molecular level includes a description of all the complexly interrelated chemical changes that occur within the cell—*i.e.,* the processes known as intermediary metabolism. The processes of growth, reproduction, and heredity, also subjects of the biochemist's curiosity, are intimately related to intermediary metabolism and cannot be understood independently of it. The properties and capacities exhibited by a complex multicellular organism can be reduced to the properties of the individual cells of that organism, and the behaviour of each individual cell can be understood in terms of its chemical structure and the chemical changes occurring within that cell. When all the chemical changes within a cell are completely described and understood, man will have achieved as complete an understanding of life as can be achieved by the intellect alone. Living processes are sufficiently complex, however, to guarantee the biochemist enough unsolved problems to last into the unforeseeable future.

Chemical composition of living matter. Every living cell contains, in addition to water and salts or minerals, a large number of organic compounds, substances composed of carbon combined with varying amounts of hydrogen and usually also of oxygen. Nitrogen, phosphorus, and sulfur are likewise common constituents. In general, the bulk of the organic matter of a cell may be classified as (1) protein, (2) carbohydrate, and (3) fat, or lipid. Nucleic acids and various other organic derivatives are also important constituents. Each class contains a great diversity of individual compounds. Many substances that cannot be classified in any of the above categories also occur, though usually not in large amounts.

Proteins are fundamental to life, not only as structural elements (*e.g.,* collagen) and to provide defense (as antibodies) against invading destructive forces but also because the essential biocatalysts are proteins. The chemistry of proteins is based on the researches of the German chemist Emil Fischer, whose work from 1882 demonstrated that proteins are very large molecules, or polymers, built up of about 24 amino acids. Proteins may vary in size from small—insulin with a molecular weight of 5,700 (based on the weight of a hydrogen atom as 1)—to very large—molecules with molecular weights of more than 1,000,000. The first complete amino acid sequence was determined for the insulin molecule in the 1950s. By 1963 the chain of amino acids in the protein enzyme ribonuclease (molecular weight 12,700) had also been determined, aided by the powerful physical techniques of X-ray-diffraction analysis. In the 1960s, Nobel Prize winners J.C. Kendrew and M.F. Perutz, utilizing X-ray studies, constructed detailed atomic models of the proteins hemoglobin and myoglobin (the respiratory pigment in muscle), which were later

Protein size

confirmed by sophisticated chemical studies. The abiding interest of biochemists in the structure of proteins rests on the fact that the arrangement of chemical groups in space yields important clues regarding the biological activity of molecules.

Carbohydrates include such substances as sugars, starch, and cellulose. The second quarter of the 20th century witnessed a striking advance in the knowledge of how living cells handle small molecules, including carbohydrates. The metabolism of carbohydrates became clarified during this period, and elaborate pathways of carbohydrate breakdown and subsequent storage and utilization were gradually outlined in terms of cycles (*e.g.,* the Embden–Meyerhof glycolytic cycle and the Krebs cycle). The involvement of carbohydrates in respiration and muscle contraction was well worked out by the 1950s. Refinements of the schemes continue.

Fats, or lipids, constitute a heterogeneous group of organic chemicals that can be extracted from biological material by nonpolar solvents such as ethanol, ether, and benzene. The classic work concerning the formation of body fat from carbohydrates was accomplished during the early 1850s. Those studies, and later confirmatory evidence, have shown that the conversion of carbohydrate to fat occurs continuously in the body. The liver is the main site of fat metabolism. Fat absorption in the intestine, studied as early as the 1930s, still is under investigation by biochemists. The control of fat absorption is known to depend upon a combination action of secretions of the pancreas and bile salts. Abnormalities of fat metabolism, which result in disorders such as obesity and rare clinical conditions, are the subject of much biochemical research. Equally interesting to biochemists is the association between high levels of fat in the blood and the occurrence of arteriosclerosis ("hardening" of the arteries).

Significance of nucleic acids

Nucleic acids are large, complex compounds of very high molecular weight present in the cells of all organisms and in viruses. They are of great importance in the synthesis of proteins and in the transmission of hereditary information from one generation to the next. Originally discovered as constituents of cell nuclei (hence their name), it was assumed for many years after their isolation in 1869 that they were found nowhere else. This assumption was not challenged seriously until the 1940s, when it was determined that two kinds of nucleic acid exist: deoxyribonucleic acid (DNA), in the nuclei of all cells and in some viruses; and ribonucleic acid (RNA), in the cytoplasm of all cells and in most viruses.

The profound biological significance of nucleic acids came gradually to light during the 1940s and 1950s. Attention turned to the mechanism by which protein synthesis and genetic transmission was controlled by nucleic acids (see below *Genes*). During the 1960s, experiments were aimed at refinements of the genetic code. Promising attempts were made during the late 1960s and early 1970s to accomplish duplication of the molecules of nucleic acids outside the cell—*i.e.,* in the laboratory. By the mid-1980s genetic engineering techniques had accomplished, among other things, in vitro fertilization and the recombination of DNA (so-called gene splicing).

Nutrition. Biochemists have long been interested in the chemical composition of the food of animals. All animals require organic material in their diet, in addition to water and minerals. This organic matter must be sufficient in quantity to satisfy the caloric, or energy, requirements of the animals. Within certain limits, carbohydrate, fat, and protein may be used interchangeably for this purpose. In addition, however, animals have nutritional requirements for specific organic compounds. Certain essential fatty acids, about ten different amino acids (the so-called essential amino acids), and vitamins are required by many higher animals. The nutritional requirements of various species are similar but not necessarily identical; thus man and the guinea pig require vitamin C, or ascorbic acid, whereas the rat does not.

That plants differ from animals in requiring no preformed organic material was appreciated soon after the plant studies of the late 1700s. The ability of green plants to make all their cellular material from simple substances—carbon dioxide, water, salts, and a source of nitrogen such as ammonia or nitrate—was termed photosynthesis. As the name implies, light is required as an energy source, and it is generally furnished by sunlight. The process itself is primarily concerned with the manufacture of carbohydrate, from which fat can be made by animals that eat plant carbohydrates. Protein can also be formed from carbohydrate, provided ammonia is furnished.

Similarity of chemical changes in living things

In spite of the large apparent differences in nutritional requirements of plants and animals, the patterns of chemical change within the cell are the same. The plant manufactures all the materials it needs, but these materials are essentially similar to those that the animal cell uses and are often handled in the same way once they are formed. Plants could not furnish animals with their nutritional requirements if the cellular constituents in the two forms were not basically similar.

Digestion. The organic food of animals, including man, consists in part of large molecules. In the digestive tracts of higher animals, these molecules are hydrolyzed, or broken down, to their component building blocks. Proteins are converted to mixtures of amino acids, and polysaccharides are converted to monosaccharides. In general, all living forms use the same small molecules, but many of the large complex molecules are different in each species. An animal, therefore, cannot use the protein of a plant or of another animal directly but must first break it down to amino acids and then recombine the amino acids into its own characteristic proteins. The hydrolysis of food material is necessary also to convert solid material into soluble substances suitable for absorption. The liquefaction of stomach contents aroused the early interest of observers, long before the birth of modern chemistry, and the hydrolytic enzymes secreted into the digestive tract were among the first enzymes to be studied in detail. Pepsin and trypsin, the proteolytic enzymes of gastric and pancreatic juice, respectively, continue to be intensively investigated.

The products of enzymatic action on the food of an animal are absorbed through the walls of the intestines and distributed to the body by blood and lymph. In organisms without digestive tracts, substances must also be absorbed in some way from the environment. In some instances simple diffusion appears to be sufficient to explain the transfer of a substance across a cell membrane. In other cases, however (*e.g.,* in the case of the transfer of glucose from the lumen of the intestine to the blood), transfer occurs against a concentration gradient. That is, the glucose may move from a place of lower concentration to a place of higher concentration.

In the case of the secretion of hydrochloric acid into gastric juice, it has been shown that active secretion is dependent on an adequate oxygen supply (*i.e.,* on the respiratory metabolism of the tissue), and the same holds for absorption of salts by plant roots. The energy released during the tissue oxidation must be harnessed in some way to provide the energy necessary for the absorption or secretion. This harnessing is achieved by a special chemical coupling system. The elucidation of the nature of such coupling systems has been an objective of the biochemist.

Extent of knowledge about blood

Blood. One of the animal tissues that has always excited special curiosity is blood. Blood has been investigated intensively from the early days of biochemistry, and its chemical composition is known with greater accuracy and in more detail than that of any other tissue in the body. The physician takes blood samples to determine such things as the sugar content, the urea content, or the inorganic-ion composition of the blood, since these show characteristic changes in disease.

The blood pigment hemoglobin has been intensively studied. Hemoglobin is confined within the blood corpuscles and carries oxygen from the lungs to the tissues. It combines with oxygen in the lungs, where the oxygen concentration is high, and releases the oxygen in the tissues, where the oxygen concentration is low. The hemoglobins of higher animals are related but not identical. In invertebrates, other pigments may take the place and function of hemoglobin. The comparative study of these compounds constitutes a fascinating chapter in biochemical investigation.

The proteins of blood plasma also have been extensively investigated. The gamma-globulin fraction of the plasma proteins contains the antibodies of the blood and is of practical value as an immunizing agent. An animal develops resistance to disease largely by antibody production. Antibodies are proteins with the ability to combine with an antigen (*i.e.,* an agent that induces their formation). When this agent is a component of a disease-causing bacterium, the antibody can protect an organism from infection by that bacterium. The chemical study of antigens and antibodies and their interrelationship is known as immunochemistry.

Metabolism and hormones. The cell is the site of a constant, complex, and orderly set of chemical changes collectively called metabolism. Metabolism is associated with a release of heat. The heat released is the same as that obtained if the same chemical change is brought about outside the living organism. This confirms the fact that the laws of thermodynamics apply to living systems just as they apply to the inanimate world. The pattern of chemical change in a living cell, however, is distinctive and different from anything encountered in nonliving systems. This difference does not mean that any chemical laws are invalidated. It instead reflects the extraordinary complexity of the interrelations of cellular reactions.

Hormones, which may be regarded as regulators of metabolism, are investigated at three levels, to determine (1) their physiological effects, (2) their chemical structure, and (3) the chemical mechanisms whereby they operate. The study of the physiological effects of hormones is properly regarded as the province of the physiologist. Such investigations obviously had to precede the more analytical chemical studies. The chemical structures of thyroxine and adrenaline are known. The chemistry of the sex and adrenal hormones, which are steroids, has also been thoroughly investigated. The hormones of the pancreas—insulin and glucagon—and the hormones of the hypophysis (pituitary gland) are peptides (*i.e.,* compounds composed of chains of amino acids). The structures of most of these hormones has been determined. The chemical structures of the plant hormones, auxin and gibberellic acid, which act as growth-controlling agents in plants, are also known.

tatus of
ormonal
tudy

The first and second phases of the hormone problem thus have been well, though not completely, explored, but the third phase is still in its infancy. It seems likely that different hormones exert their effects in different ways. Some may act by affecting the permeability of membranes; others appear to control the synthesis of certain enzymes. Evidently some hormones also control the activity of certain genes.

Genes. Genetic studies have shown that the hereditary characteristics of a species are maintained and transmitted by the self-duplicating units known as genes, which are composed of nucleic acids and located in the chromosomes of the nucleus. One of the most fascinating chapters in the history of the biological sciences contains the story of the elucidation, in the mid-20th century, of the chemical structure of the genes, their mode of self-duplication, and the manner in which the deoxyribonucleic acid (DNA) of the nucleus causes the synthesis of ribonucleic acid (RNA), which, among its other activites, causes the synthesis of protein. Thus, the capacity of a protein to behave as an enzyme is determined by the chemical constitution of the gene (DNA) that directs the synthesis of the protein. The relationship of genes to enzymes has been demonstrated in several ways. The first successful experiments, devised by the Nobel Prize winners George W. Beadle and Edward L. Tatum, involved the bread mold *Neurospora crassa;* the two men were able to collect a variety of strains that differed from the parent strain in nutritional requirements. Such strains had undergone a mutation (change) in the genetic makeup of the parent strain. The mutant strains required a particular amino acid not required for growth by the parent strain. It was then shown that such a mutant had lost an enzyme essential for the synthesis of the amino acid in question. The subsequent development of techniques for the isolation of mutants with specific nutritional requirements led to a special procedure for studying intermediary metabolism.

Evolution and origin of life. The exploration of space beginning in the mid-20th century intensified speculation about the possibility of life on other planets. At the same time, man was beginning to understand some of the intimate chemical mechanisms used for the transmission of hereditary characteristics. It was possible, by studying protein structure in different species, to see how the amino acid sequences of functional proteins (*e.g.,* hemoglobin and cytochrome) have been altered during phylogeny (the development of species). It was natural, therefore, that biochemists should look upon the problem of the origin of life as a practical one. The synthesis of a living cell from inanimate material was not regarded as an impossible task for the future.

The coming of space biochemistry

Applied biochemistry. An early objective in biochemistry was to provide analytical methods for the determination of various blood constituents because it was felt that abnormal levels might indicate the presence of metabolic diseases. The clinical chemistry laboratory now has become a major investigative arm of the physician in the diagnosis and treatment of disease and is an indispensable unit of every hospital. Some of the older analytical methods directed toward diagnosis of common diseases are still the most commonly used—for example, tests for determining the levels of blood glucose, in diabetes; urea, in kidney disease; uric acid, in gout; and bilirubin, in liver and gallbladder disease. With development of the knowledge of enzymes, determination of certain enzymes in blood plasma has assumed diagnostic value, such as alkaline phosphatase, in bone and liver disease; acid phosphatase, in prostatic cancer; amylase, in pancreatitis; and lactate dehydrogenase and transaminase, in cardiac infarct. Electrophoresis of plasma proteins is commonly employed to aid in the diagnosis of various liver diseases and forms of cancer. Both electrophoresis and ultracentrifugation of serum constituents (lipoproteins) are used increasingly in the diagnosis and examination of therapy of atherosclerosis and heart disease. Many specialized and sophisticated methods have been introduced, and machines have been developed for the simultaneous automated analysis of many different blood constituents in order to cope with increasing medical needs.

Diagnostic tests

Analytical biochemical methods have also been applied in the food industry to develop crops superior in nutritive value and capable of retaining nutrients during the processing and preservation of food. Research in this area is directed particularly to preserving vitamins as well as colour and taste, all of which may suffer loss if oxidative enzymes remain in the preserved food. Tests for enzymes are used for monitoring various stages in food processing.

Biochemical techniques have been fundamental in the development of new drugs. The testing of potentially useful drugs includes studies on experimental animals and man to observe the desired effects and also to detect possible toxic manifestations; such studies depend heavily on many of the clinical biochemistry techniques already described. Although many of the commonly used drugs have been developed on a rather empirical (trial-and-error) basis, an increasing number of therapeutic agents have been designed specifically as enzyme inhibitors to interfere with the metabolism of a host or invasive agent. Biochemical advances in the knowledge of the action of natural hormones and antibiotics promise to aid further in the development of specific pharmaceuticals.

Methods in biochemistry. Like other sciences, biochemistry aims at quantifying, or measuring, results, sometimes with sophisticated instrumentation. The earliest approach to a study of the events in a living organism was an analysis of the materials entering an organism (foods, oxygen) and those leaving (excretion products, carbon dioxide). This is still the basis of so-called balance experiments conducted on animals, in which, for example, both foods and excreta are thoroughly analyzed. For this purpose many chemical methods involving specific colour reactions have been developed, requiring spectrum-analyzing instruments (spectrophotometers) for quantitative measurement. Gasometric techniques are those commonly used for measurements of oxygen and carbon dioxide, yielding respiratory quotients (the ratio of carbon dioxide to oxygen). Some-

what more detail has been gained by determining the quantities of substances entering and leaving a given organ and also by incubating slices of a tissue in a physiological medium outside the body and analyzing the changes that occur in the medium. Because these techniques yield an overall picture of metabolic capacities, it became necessary to disrupt cellular structure (homogenization) and to isolate the individual parts of the cell—nuclei, mitochondria, lysosomes, ribosomes, membranes—and finally the various enzymes and discrete chemical substances of the cell in an attempt to understand the chemistry of life more fully.

Centrifugation and electrophoresis. An important tool in biochemical research is the centrifuge, which through rapid spinning imposes high centrifugal forces on suspended particles, or even molecules in solution, and causes separations of such matter on the basis of differences in weight. Thus, red cells may be separated from plasma of blood, nuclei from mitochondria in cell homogenates, and one protein from another in complex mixtures. Proteins are separated by ultracentrifugation—very high speed spinning; with appropriate photography of the protein layers as they form in the centrifugal field, it is possible to determine the molecular weights of proteins.

Exploiting electrical charges of molecules

Another property of biological molecules that has been exploited for separation and analysis is their electrical charge. Amino acids and proteins possess net positive or negative charges according to the acidity of the solution in which they are dissolved. In an electric field, such molecules adopt different rates of migration toward positively (anode) or negatively (cathode) charged poles and permit separation. Such separations can be effected in solutions or when the proteins saturate a stationary medium such as cellulose (filter paper), starch, or acrylamide gels. By appropriate colour reactions of the proteins and scanning of colour intensities, a number of proteins in a mixture may be measured. Separate proteins may be isolated and identified by electrophoresis, and the purity of a given protein may be determined. (Electrophoresis of human hemoglobin revealed the abnormal hemoglobin in sickle-cell anemia, the first definitive example of a "molecular disease.")

Chromatography and isotopes. The different solubilities of substances in aqueous and organic solvents provide another basis for analysis. In its earlier form, a separation was conducted in complex apparatus by partition of substances in various solvents. A simplified form of the same principle evolved as "paper chromatography," in which small amounts of substances could be separated on filter paper and identified by appropriate colour reactions. In contrast to electrophoresis, this method has been applied to a wide variety of biological compounds and has contributed enormously to research in biochemistry.

The general principle has been extended from filter paper strips to columns of other relatively inert media, permitting larger scale separation and identification of closely related biological substances. Particularly noteworthy has been the separation of amino acids by chromatography in columns of ion-exchange resins, permitting the determination of exact amino acid composition of proteins. Following such determination, other techniques of organic chemistry have been used to elucidate the actual sequence of amino acids in complex proteins. Another technique of column chromatography is based on the relative rates of penetration of molecules into beads of a complex carbohydrate according to size of the molecules. Larger molecules are excluded relative to smaller molecules and emerge first from a column of such beads. This technique not only permits separation of biological substances but also provides estimates of molecular weights.

Perhaps the single most important technique in unravelling the complexities of metabolism has been the use of isotopes (heavy or radioactive elements) in labelling biological compounds and "tracing" their fate in metabolism. Measurement of the isotope-labelled compounds has required considerable technology in mass spectroscopy and radioactive detection devices.

A variety of other physical techniques, such as nuclear magnetic resonance, electron spin spectroscopy, circular dichroism, and X-ray crystallography, have become prominent tools in revealing the relation of chemical structure to biological function. (E.H.St./B.V.)

GENETICS

Genetics is the branch of biology concerned with heredity. Since prehistoric times, man has recognized the influence of heredity and has applied its principles to the improvement of cultivated crops and domestic animals. A Babylonian tablet more than 6,000 years old, for example, shows pedigrees of horses and indicates possible inherited characteristics; other old carvings show cross-pollination of date palm trees. Most of the mechanisms of heredity, however, remained a mystery until the 20th century, when scientifically supported information became available.

Genetics may be defined as the study of the way in which genes operate and the way in which they are transmitted from parents to offspring. Modern genetics involves study of the mechanism of gene action—the way in which the genetic material (deoxyribonucleic acid, or DNA) affects physiological reactions within the cell. Although genes determine the features an individual may develop, the features that actually develop depend upon the complex interaction between genes and their environment. Normal green plants, for example, have genes containing the information necessary to synthesize the chlorophyll that gives them their green colour, and chlorophyll is synthesized in an environment containing light; *i.e.,* the gene for chlorophyll is expressed. If the plant is placed in a dark environment, chlorophyll synthesis stops; *i.e.,* the gene is no longer expressed.

The study of gene action

Genetics overlaps many different branches of biology and many other sciences; *e.g.,* chemistry, physics, mathematics, sociology, psychology, and medicine. Microbiologists who study inheritance in microorganisms are called microbial geneticists; cytologists who study the genetics of cells are called cytogeneticists. Biochemical, or molecular, geneticists investigate the chemical nature of the gene and its methods of action. Some physicists have applied their techniques to molecular genetics, and mathematicians may specialize in population genetics. Behavioral scientists also look to genetics to solve certain problems of human and animal behaviour. Specialists in medical genetics or genetic counselling act on the knowledge that many of man's afflictions are hereditary.

Historical background. The Greek philosopher Pythagoras speculated around 500 BC that human life begins with a blend of male and female fluids, or semens, originating in body parts. Aristotle later postulated that the semens are purified blood and that blood, therefore, is the element of heredity. That this later concept persisted in the Western world is indicated by such common phrases as blue blood, blood-will-tell, blood relative, bad blood, and royal blood.

About 1651, William Harvey disproved the Greek concept; his discovery that deer embryos have the appearance of a tiny ball during early developmental stages and resemble a deer only later in development led him to conclude that the origin of the tiny ball was a small egg. Before the end of the 17th century, it had been suggested that the female structures called ovaries are the source of eggs and that sperm might carry the hereditary material of the male.

Early in the 19th century, Jean-Baptiste Lamarck suggested that acquired characteristics are inherited. Around 1865 Gregor Mendel reported his discoveries on inheritance in garden peas. A few years later, the DNA component of genes was isolated from pus cells, and it was discovered that salmon sperm also contain considerable amounts of DNA. Late in the 19th century, a German physician, August Weismann, showed that reproductive cells (germ plasm) are independent of other body cells (somatoplasm), thus refuting earlier hypotheses of inheritance of acquired characteristics.

20th-century discoveries

The concept of sudden changes in heredity (mutations) was introduced in the beginning of the 20th century. Discoveries concerning sex determination in insect chromosomes and gene linkage on a chromosome of sweet peas were made soon afterward in the United States and England. In 1908 an English mathematician and a German

physician formulated the so-called Hardy–Weinberg principle, which provided the foundation for population genetics. The study of biochemical genetics was begun in 1909 in England with an effort to discover the way by which gene-induced enzyme deficiencies cause abnormalities.

Hermann J. Muller, a U.S. geneticist, induced mutations in the fruit fly with X-rays in 1927. Experiments on the mold *Neurospora* by George W. Beadle and Edward L. Tatum proved that the function of most genes is to direct the synthesis of enzymes, which thus are the expression of many hereditary traits. By 1944 DNA had been proved to be the substance of heredity, and in 1953 James D. Watson and F.H.C. Crick reported a structure of DNA compatible with the capability for self-duplication. Two French Nobel Prize winners, François Jacob and Jacques Monod, discovered the mechanism by which hereditary information is transferred from genes to the site of protein (enzyme) synthesis. Their work resulted in the discovery of the genetic code, by which DNA is translated into protein. Barbara McClintock, who received a Nobel Prize in 1983, was cited for her discovery of mobile genetic elements—some of the mechanisms that account for mutation.

Areas of study. *Classical genetics.* Classical genetics, which remains a basis for all other topics in genetics, is concerned primarily with the method by which genetic traits classified as dominant (always expressed), recessive (subordinate to a dominant trait), intermediate (partially expressed), or polygenic (due to multiple genes) are transmitted in plants and animals. These traits may be sex-linked (result from the action of a gene on the sex, or X, chromosome) or autosomal (result from the action of a gene on a chromosome other than a sex chromosome). Classical genetics began with Mendel's study of inheritance in garden peas and continues with studies of inheritance in many different plants and animals.

Cytogenetics. Cytogenetics blends the skills of cytologists, who study the structure and activities of cells, with those of geneticists, who study the relationship between the mechanism of heredity and cellular activities. Cytologists discovered chromosomes and the way in which they duplicate and separate during cell division at about the same time that geneticists began to understand the behaviour of genes at the cellular level. The close correlation between the two disciplines led to their combination.

Plant and animal cytogenetics

Plant cytogenetics early became an important subdivision of cytogenetics because, as a general rule, plant chromosomes are larger than those of animals. Animal cytogenetics became important after the development of the so-called squash technique in which entire cells are pressed flat on a piece of glass and observed through a microscope; the human chromosomes were numbered using this technique.

Microbial genetics. Microorganisms were generally ignored by the early geneticists because they are small in size and were thought to lack variable traits and the sexual reproduction necessary for a mixing of genes from different organisms. After it was discovered that microorganisms can have different physiological characteristics and also are able to reproduce sexually, they became objects of great interest to geneticists because they reproduce more rapidly than larger organisms; *i.e.,* a mutation, or change, occurs in a gene about one time in 10,000,000 gene duplications, and one bacterium may produce 10,000,000,000 offspring, among which are numerous mutants, in 48 hours.

Many discoveries in microbial genetics have been applied to other areas of genetics; for example, the way in which genes produce enzymes that function in turn to produce genetic traits has important applications to human genetics. Much of microbial genetics also applies to the study of the genetics of viruses.

Molecular genetics. Molecular genetics includes the study of the molecular nature of the gene and the method by which genes control the activities of the cell. Molecular geneticists have studied the molecular structure of a gene (*e.g.,* that involved in the synthesis of the human blood pigment, hemoglobin) and determined the exact sequence of its components; in addition, they have created a synthetic gene by joining the components comprising a known gene in the correct sequence. Genetic engineering had become a commercial enterprise by the early 1980s.

Synthetic genes

Population genetics. A study of genes in populations of animals provides information on past migrations, evolutionary relationships and extents of mixing among different varieties and species, and methods of adaptation to the environment. Statistical methods are used to analyze gene distributions and chromosomal variations in populations.

Human population geneticists have traced migration and invasion routes of man; genetic studies of present-day Europeans, for example, reveal routes of human migrations that occurred hundreds or thousands of years ago. The origin of the people inhabiting South Pacific islands and the degree of intermingling among mixed races also are studied by human population geneticists.

Behavioral genetics. Another aspect of genetics is the study of the influence of heredity on behaviour. Many characteristics once considered to be acquired behavioral patterns actually are of a hereditary nature. The role of heredity in instinctive patterns of behaviour among animals has long been recognized, but many of man's actions also have a hereditary explanation. The effect of various drugs (*e.g.,* the hallucinogenic drug lysergic acid diethylamide, or LSD) on behavioral patterns in animals, including man, is of particular interest.

Human genetics. Some geneticists specialize in human genetics. When classical geneticists first determined the principles of heredity in plants, fruit flies, mice, and other forms of life, they tried to interpret man's heredity in a similar way but found many traits that did not fit the patterns. As techniques improved, it was found that the method of inheritance of human characteristics is the same as that for other living things.

Some human geneticists, called genetic counsellors, advise individuals concerning the probabilities for the appearance of serious hereditary defects in their children. The counsellors usually have medical training because many traits are recognizable only after special diagnostic procedures. Medical genetics is another important application of human genetics. Many medical schools devote entire departments to medical genetics, which is the study of the treatment and prevention of inherited afflictions in man.

It is possible that man may someday control his heredity; even now functional genes can be transferred from one organism to another, and certain treatments are able to cause specific kinds of mutations. Such manipulation of genes eventually may be useful in solving many human hereditary diseases; *e.g.,* stopping the function of genes that are out of control, starting the function of nonfunctioning ones. Activation of nonfunctioning genes in some types of tissue may enable them to replace body parts that have been injured or destroyed. It is conceivable that man may someday learn how to change harmful genes into normal ones.

Methods in genetics. *Experimental breeding.* When animals that differ with respect to *one* primary trait are bred, and their offspring then are bred among themselves to give a second generation, the method of inheritance of the trait can be determined; the process is known as a monohybrid cross. A dihybrid cross involves breeding individuals that differ with respect to *two* traits; the results of such crosses show whether the genes are linked on the same chromosome or are on different chromosomes. If the genes are linked, the distance between them can be determined by the number of recombinations of traits obtained, an indication of the amount of crossing over between genes. By such crosses, geneticists have established elaborate chromosome maps of many organisms showing the location of many genes on the chromosomes.

Hybrids

A test cross may be used to determine if animals carry recessive genes; *e.g.,* cocker spaniel dogs may be of solid colours or parti-coloured (spotted). Since the gene for parti-coloured is recessive, it may be carried (but not expressed) by some solid-coloured dogs. If a solid-coloured dog is suspected of carrying a recessive gene for parti-coloured, it is bred to a parti-coloured dog. Parti-coloured offspring indicate that the solid-coloured dog carries the recessive gene. This technique is used by animal breeders to eliminate undesirable recessive genes.

Experimental breeding is most successful in organisms

with large numbers of offspring, a relatively short life cycle, and a number of variable characteristics. The fruit fly, *Drosophila,* meets these requirements and has been used extensively in breeding experiments; mice also have been used extensively.

Cytogenetic techniques. Cytogenetic techniques are closely associated with experimental breeding. Older cytogenetic techniques involve placing cells in paraffin wax, slicing thin sections, and preparing them for microscopic study. The newer and faster squash technique involves squashing entire cells and studying their chromosomes. Dyes that selectively stain various parts of the cell are used; the genes, for example, may be located by selectively staining the DNA of which they are composed. Radioactive compounds also are valuable in determining the location of various components of the cell. Tissue-culture techniques may be used to grow cells before squashing; white blood cells can be grown from samples of human blood and studied with the squash technique.

Biochemical techniques. Biochemical techniques are used to determine the activities of genes within cells. Radioactive compounds are valuable in studies involving gene duplication and cell metabolism. Thymine is a compound found only in genes; if radioactive thymine is placed in a tissue-culture medium in which cells are growing, genes use it to duplicate themselves. When cells containing radioactive thymine are analyzed, the results show that, during duplication, genes split in half, and each half synthesizes its missing components. When radioactive uracil, a compound found only in the RNA component of cells, is incorporated into the RNA messengers of genes, their pathway from the chromosomes to the site of protein synthesis in the cytoplasm (ribosomes) is revealed.

Radioactive thymine

Chemical tests are used to distinguish certain inherited characteristics of man; *e.g.,* urinalysis and blood analysis reveal the presence of certain inherited abnormalities—phenylketonuria (PKU), cystinuria, alkaptonuria, gout, and galactosemia. Special techniques (*e.g.,* chromatography, electrophoresis) are used to separate the components of proteins, so that inherited differences in their structures can be revealed; for example, more than 100 different kinds of human hemoglobin molecules have been identified.

Physiological techniques. Physiological techniques also are used in genetic investigations. In microorganisms, most genetic variations involve some important cell function. Some strains of one bacterium (*Escherichia coli*), for example, are able to synthesize the vitamin thiamine from simple compounds; others, which lack an enzyme necessary for this synthesis, cannot survive unless thiamine is present. The two strains can be distinguished by placing them on a thiamine-free mixture; those that grow have the gene for the enzyme, those that fail to grow do not. The technique also is applied to human cells since many inherited human abnormalities are caused by a faulty gene that fails to produce a vital enzyme; albinism, which results from an inability to produce the pigment melanin in the skin, hair, or iris of the eyes, is an example of an enzyme deficiency in man.

Immunological techniques. Many substances (*e.g.,* proteins) are antigenic; *i.e.,* when introduced into a vertebrate body, they stimulate the production of specific proteins called antibodies. Various antigens exist in red blood cells, including those that comprise the major blood groups of man (A, B, AB, O). Blood antigens of man include inherited variations, and the particular combination of antigens in an individual is almost as unique as fingerprints. Immunological techniques are used in the blood-group determinations that precede blood transfusions and in determining Rh incompatibility in childbirth.

Evolutionary relationships can be determined by immunological techniques. If protein from a fruit fly is injected into a guinea pig, and the guinea pig produces antibodies and its blood serum is then mixed with proteins from the fly, antigens and antibodies react to produce a cloudy mixture. Mixtures of guinea pig blood serum and proteins from other fruit-fly species cause various degrees of cloudiness, depending on their evolutionary relationship to the original species; *e.g.,* the closer the relationship, the greater degree of cloudiness.

Mathematical techniques. Mathematical techniques are used extensively in genetics. The laws of probability are applicable to crossbreeding and are used to predict ratios concerning the appearance of specific traits in offspring. Geneticists also use statistical methods to determine the significance of deviations from expected results. In investigations involving possible mutagenic effects of factors such as high-energy radiation and drugs, statistical tests are used to establish the validity of conclusions; statistics are used in studies of the possible effects of LSD in producing chromosome aberrations in man, for example, to show whether differences found in cells of users and nonusers of the drug are significant.

Statistical methods

Mathematics is used by population geneticists to evaluate the distribution of genes in populations. The Hardy–Weinberg principle, for example, is important in studying animals that carry a recessive gene; when the actual number of carriers is much greater than that calculated, it is concluded that some environmental factor favours the carriers. The gene for sickle-cell anemia in black Africans, for instance, is found in more people than the frequency of those who have the anemia would indicate because people who carry the gene are more resistant to malaria than noncarriers and, therefore, have a better chance of survival.

Applied genetics. *Medicine.* Genetic techniques are used in medicine to diagnose and treat inherited human disorders. Knowledge of a family history of cancer or tuberculosis may indicate a hereditary tendency to develop these afflictions. Cells from embryonic membranes reveal certain genetic abnormalities, including enzyme deficiencies, that may be present in newborn babies, and thus permit early treatment. Many countries require a blood test of newborn babies to determine the presence of an enzyme necessary to convert an amino acid, phenylalanine, into simpler products. Phenylketonuria, which results from lack of the enzyme, causes permanent brain damage if not treated soon after birth. The presence of approximately 100 different types of human genetic diseases can be detected in embryos as young as 12 weeks; the procedure, called amniocentesis, involves removal and testing of a small amount of fluid from around the embryo.

Agriculture and animal husbandry. Agriculture and animal husbandry apply genetic techniques to improve plants and animals. Plant geneticists produce new species by special treatment; *e.g.,* a hybrid grain has been produced from wheat and rye, and plants resistant to destruction by insect pests have been developed.

Plant breeders use the techniques of budding and grafting to maintain desirable gene combinations originally obtained from crossbreeding. The use of the chemical compound colchicine, which causes chromosomes to double in number, has resulted in many new varieties of fruits, vegetables, and flowers.

Budding and grafting

Animal breeders use artificial insemination to propagate the genes of prize bulls. Prize cows can transmit their genes to hundreds of offspring by hormone treatment, which stimulates the release of many eggs that are collected, fertilized, and transplanted to foster mothers.

Industry. Various industries employ geneticists; the brewing industry, for example, may use geneticists to obtain strains of yeast that produce large quantities of alcohol. The pharmaceutical industry has developed strains of molds, bacteria, and other microorganisms high in antibiotic yield. (A.M.W.)

EUGENICS

Eugenics is the study of human improvement in all aspects by genetic means. Its goal is to increase the proportion of persons with better than average genetic endowment. It draws from the science of genetics for an understanding of the processes of heredity; from psychology for analysis of the part played by differences in heredity and environment in the development of personality, intelligence, and character; from medicine and medical genetics for information on hereditary defects and susceptibility to disease; and from sociology, and more particularly demography, for the rates of births and deaths, the mating habits, and the social and physical factors that determine the relative

increase or decline of persons with different characteristics. In its scientific aspect it is essentially the study of trends and causal factors in human evolution.

A companion science, euthenics, is also concerned with the improvement of mankind, but by adjustment of the environment. Obviously both concerns are essential to the well-being of man, and since neither science can stand alone, the term "humanics" has been suggested to embrace both aspects of human improvement.

Historical background. The idea of applying knowledge of heredity to the improvement of the human race goes back to earliest times. References to eugenic ideals appear in the Old Testament, and Plato's *Republic* idealizes a society in which there is constant selection for the improvement of the human stock. Thomas Robert Malthus noted the struggle for existence; Charles Darwin saw in it the means of evolution. It remained for Francis Galton, a scientist of many talents and a cousin of Darwin, to see that the theory of evolution implied that man might in part direct his own evolutionary future. Galton's first important work, *Hereditary Genius* (1869), contained the results of his studies of the families of eminent men as evidence for his belief that "it would be quite practical to produce a highly gifted race of men by judicious marriages during several consecutive generations." In 1883 he published *Inquiries into Human Faculty,* in which he coined the term eugenics. In *Natural Inheritance* (1889), Galton pioneered the development and application of advanced statistical methods to the study of man.

While Galton's thinking was much ahead of his time, he retained some of the prejudices of an English gentleman in regard to social class and race. In his studies of the families of eminent men Galton underestimated the effects of a favourable environment. But his faults in this respect were far less than those of many of his followers. He developed and used many statistical methods for the study of populations and was the first to recognize the value of the study of twins for research in heredity. He repudiated the Lamarckian view, then widely held, that acquired characteristics were inherited. Although Galton attached religious significance to the eugenic movement, he did not think of it in revolutionary terms, but rather as ". . . the new duty which is supposed to be exercised concurrently with and not in opposition to, the old ones upon which the social fabric depends."

he slow
ceptance
f eugeni-
al ideas

For all his wisdom and high ability, Galton was not able to gain any wide acceptance for eugenics, largely because much of the scientific and technical foundation was lacking. Psychology was not sufficiently advanced to help him in his studies of individual differences or in assessing the importance of the environment; there was no cultural anthropology to provide an outline of the growth of civilizations; and birth control was so little recognized or accepted that he did not take it into account as a possible factor in the control and distribution of births.

From Galton to World War II. Galton had endowed a research fellowship in eugenics in 1904 and, in his will, provided funds for a chair of eugenics at University College, London University. The fellowship and later the chair at University College were both occupied by Karl Pearson, a brilliant mathematician who helped to create the science of biometry, the statistical aspects of biology. Pearson was a controversial figure who believed that environment had little to do with the development of mental or emotional qualities. He felt that the high birth rate of the poor was a threat to civilization and that the "higher" races must supplant the "lower." His views gave countenance to those who believed in racial and class superiorities. Thus, Pearson shares the blame for the discredit later brought on eugenics in the United States and for making possible the dreadful misuse of the word eugenics in Hitler's propaganda. The English Eugenics Society, founded by Galton in 1907 as the Eugenics Education Society, opposed Pearson's views, but was itself slow in throwing off the prevalent biases of that time.

ugenics
liscredited
y early
acist views

In the United States the American Eugenics Society was founded in 1926 by men whose views were coloured by the times in which they lived. They believed that the white race was superior to other races, and, further, that the "Nordic" white was superior to other whites. They thought of races as "pure" groups, quite separate from each other. They did not know that all races are mixtures of many types, the distribution of genes among the races varying in proportions rather than in kind, as evidenced by the distribution of the various blood groups among all of the races. They did not realize that environments are so uncontrollable that scientists cannot reasonably put forward views on the genetic differences in the performance of different races. Furthermore, they believed that the upper classes had superior hereditary qualities that justified their being the ruling class.

The science of that day supported extreme views on feeblemindedness and "criminal types." Intelligence tests, introduced in the early 1900s by a French psychologist, Alfred Binet, were thought, contrary to Binet's views, to be measures of innate, genetic intelligence. People whose test performance (*e.g.,* intelligence quotient, or IQ) gave them a mental age of 12 or less could be classified as feebleminded or morons, even though the term would then include most of those brought up in deprived environments, such as the southern Appalachian highlands or the blighted rural and urban slums occupied by the descendants of slaves. Criminality was generally considered a concomitant of feeblemindedness. Studies on degenerate family stocks were taken to prove that hundreds of persons in each of these families were feebleminded or criminal types because of the inheritance they received from a single ancestor five or six generations back. Claims were made that immigrants from southern and eastern Europe, besides being socially inferior, included a remarkable number of criminal and defective stocks. There was much in the intellectual atmosphere of the United States. that fostered an extreme hereditarian racist view.

Even before 1905 eugenists had been active in the effort to restrict immigration. Their arguments, along with those of others who advocated restrictions for different reasons, culminated with the passage of the Immigration Act of 1924, which limited quota immigrants to about 150,000 annually, with no more than 2 percent of each nationality according to the number of persons of their national origin in the United States as of 1890, and providing that beginning July 1, 1929, the quota of any country shall have the same ratio to 150,000 as the number of persons of that national origin living in the United States had to the total population of the United States, as determined from the 1920 census. In later years it became clear that the material the eugenists had presented to congressional hearings had little scientific foundation.

Movement to sterilize defectives

Eugenists at this time laid great stress on the importance of sterilizing defective persons. By 1931 sterilization laws had been enacted by 27 states in the United States, and by 1935 sterilization laws had been passed in Denmark, Switzerland, Germany, Norway, and Sweden. Most of these laws provided for the voluntary or compulsory sterilization of certain classes of people thought to be insane, idiotic, imbecilic, feebleminded, or epileptic; some applied equally to habitual criminals, moral perverts, or the feebleminded. In most cases the purpose was clearly eugenic, though some laws tacitly permitted sterilization for social rather than genetic reasons. In the United States most states did not generally enforce these rather extreme measures, and the number of sterilizations was seldom over 100 per year. Exceptions were California, where sterilizations averaged over 350 cases per year, with a total of 9,931 by 1935, and some of the southern states, with fairly high sterilization rates relative to their populations.

With the advance of science the background for these laws began to be questioned. In 1935 a committee of the American Neurological Association reported on an extensive investigation and evaluation of the "facts and theories which constitute the subject matter of the inheritance of the mental diseases, feeblemindedness, epilepsy and crime." The committee concluded that there was a serious question whether "many of the eugenic proposals now current take into account the newer genetic data" and that "though the avowed purpose of such sterilization invariably starts out by being a eugenic one, it often ends by becoming a social one." They exposed many fallacies

in the reasoning behind the sterilization laws and concluded that there was no sound basis for sterilization on the basis of immorality or character defect. They recommended sterilization only with the consent of the patient, or those responsible for him, and only in the case of certain specified diseases or defects, namely: (1) Huntington's chorea, hereditary optic atrophy, familial cases of Friedreich's ataxia, and a few other rare hereditary defects; (2) feeblemindedness common in a family; (3) schizophrenia; (4) extreme cases of manic depressive psychosis; and (5) epilepsy, mainly on the ground of its social aspects. Their recommendation for concerted, coordinated, and planned long-term research was not followed seriously until after 1945, when medical schools began active research into medical genetics.

In Europe the studies of criminal types were discredited, and serious work in eugenics continued in the form of family studies by government institutes. Only in Nazi Germany did eugenics serve racial and political purposes. There, Hitler, exploiting the concepts of the master race and the superman as expounded by some German scholars, developed racism into a powerful political weapon.

Eugenics after 1945. With the development of atomic weapons, the general public and physicians alike became educated to the dangers of genetic mutations produced by radioactive fallout from nuclear explosions. Medical X-rays came under new scrutiny, and the findings of studies on hereditary factors in diabetes and other diseases received new attention. Medicine and public health, professions previously little interested in heredity, recognized a new responsibility. By the 1960s the major medical schools in the United States had included geneticists on their research staffs, and a number had set up departments of medical genetics. Couples planning marriage became more aware of their responsibility to potential offspring, and heredity clinics associated with medical schools began to provide genetic counselling.

Still other factors arose to increase public interest in heredity. The great increase in births that took place in the United States and most European countries after World War II aroused interest in the type of children being born and the kind of homes they were being brought up in. Birth control, given worldwide publicity as a solution to the problem of too rapid growth of populations, raised questions as to the type of people who would use it, and the kind of children they would have. Medical solutions to the problems of male and female infertility—such as artificial insemination, in vitro fertilization, and surrogate mothering—have raised questions about such issues as the heredity as well as the legal status of children so conceived, and about the effect of the psychological environment on the marriage relationship and on the child. New scientific findings were widely disseminated, and the public has begun to take a more balanced view of the contributions made by both heredity and environment to good health, intelligence, and character.

Scientific basis of eugenics broadened

Genetics made unparalleled advances after 1950, opening up a whole new field of the biochemistry of gene structure and gene behaviour. At the same time, the scientific basis for the new eugenics was being further broadened by advances in psychology and in the field of demography. The Institut National d'Études Démographiques in France sponsored important investigations of human populations. In England the Royal Commission on Population made its report in 1949, and the Population Investigating Committee at the London School of Economics gave attention to quantitative and qualitative aspects of population trends in its publication *Population Studies.* In the United States a number of the larger universities established facilities for training and research in demography, and several nationwide surveys on factors affecting fertility were taken. The atmosphere was appropriate to a balanced reconsideration of eugenic problems.

Modern trends. Recent trends in the distribution of births suggest the possibility of the development in modern industrial societies of eugenic selection of a wholly voluntary and unselfconscious sort, in which the only controls might be certain moderate changes in social institutions, most of which changes are already desired for other than genetic reasons. Eugenic thinking seems to be developing along these lines, but the public and many scientists are still disturbed by the older conception of a controlled or authoritarian eugenics that would threaten other social and ethical values.

The most immediate and most publicized procedure that can be eugenically applied is the use of insemination techniques. The insemination of women whose husbands are infertile is quite frequently practiced in the United States and Europe. A proposal was made by Herman Muller that the sperm for insemination be collected from men of high attainment, of family stocks as free as possible from defective strains, and that it be kept, frozen and stored, for as long as possible, perhaps beyond the lifetime of the donor. It would then be available for women who desired to assure themselves of superior children. It was Muller's belief that the advantages of this procedure would result in its rapid and widespread adoption, and an organization was formed to carry on his work.

The "sperm bank"

Other scientific possibilities, not yet in the stage of application, are already being discussed by scientists for their possible effect on the genetic qualities of human beings. Changing hereditary factors in the human cell by chemical means is not an impossibility, but it lies in the future. Another possibility is the implanting of a specified segment of genes into a chromosome. Preliminary experiments have been performed on mice, but it will be a long time indeed before such methods can be used for the improvement of human germ cells. A nearer possibility is vegetative, or clonal, reproduction, similar to the reproduction of cuttings from plants. If this not improbable technique becomes practicable with humans, the way would be open for the reproduction, in any number desired, of individuals as similar genetically as are identical twins. Human beings could then be propagated for diverse valuable characteristics of every sort. Few persons would regard this proposal as desirable or even thinkable. The technical difficulties and social and moral objections to the "mass production" of such twin types make its application unrealistic even in the remote future.

For all practical purposes the measures discussed above are premature. They are just beyond the realm of theory, but far short of general acceptance even as theories. They would violate many established social attitudes. They raise many moral and legal questions. The population at large generally regards such proposals with suspicion. Their fears cannot be dismissed lightly. Who can assure them that eugenical procedures will not be used ulteriorly? Who can promise that there is no danger?

The public's fear of eugenics

Authoritarian controls would have to be employed and extremely difficult value judgments would have to be made. Who, for example, would determine a superior quality in man? Who would make the decisions whether certain qualities should be preserved or extinguished? Modern eugenics is little interested in authoritarian controls. Rather, it believes that the social and economic environment can be shaped so as to influence a eugenic distribution of births throughout entire populations in a voluntary and largely unconscious process of selection.

Areas of study. *Genetic defects.* It is generally accepted that genes causing defects originate in mutations that occasionally take place in one or another of the many thousands of genes present in pairs in every human cell. Usually the mutated gene is recessive; that is, it does not have a harmful effect unless the other gene in the pair has the same characteristics. The chances of two such genes meeting as a pair depend on the number of such genes distributed throughout the intermarrying population. Thus under random mating, if a particular deleterious gene is carried by one person in a hundred, the chances of the mating of persons with such genes are one in 10,000. While the chances of such an unfortunate mating seem slim, in a large population such a ratio results in a considerable number of persons with defects.

With an increase in the proportion of carriers of a particular defective gene, there is an even greater increase in the likelihood of a mating in which two such genes will be paired in the same fertilized cell. When this happens, there will be a defect. If the defect is lethal (*i.e.*, results in

stillbirth or infant death), the two deleterious genes will be removed from the population. If the defect is minor, but of a sort to make marriage or reproduction less likely, the genes are to that extent less likely to survive. Thus at some point nature establishes a balance in which, for every new deleterious gene brought into circulation, a similar gene is lost from circulation.

Geneticists believe that most people carry at least a few deleterious genes; some estimate the average to be as high as 8 percent, and at least 2 percent of the babies born carry some genetic defect, major or minor.

Most scientists believe that the proportion of carriers of deleterious genes may now be increasing, though the results in the form of an increase in defect would not be apparent for a number of generations. Greater exposure to X-rays and to fallout from atomic explosions is currently accelerating the rate of mutations, while medicine is making it possible for an increasing number of persons with some kind of genetic defect to live long enough to rear children unfortunate in having that parental defect. Diabetics, for instance, formerly died at an early age, and the genes for diabetes were lost with them. After an insulin treatment was found, these people led longer useful lives, but were not generally able to have children. Now safe delivery of the children of diabetic mothers is commonplace. The deleterious genes remain in circulation, while new ones are constantly added by mutations. Thus, with good medical care, diabetics live normally and often bear diabetic children of their own. The accumulation of certain severe types of deleterious genes could conceivably reach a point that might seriously threaten man's future.

he genetic urden

To support and maintain defective members in the population is not at odds with the goal of genetic improvement of mankind so long as such persons do not reproduce defective offspring.

While present social and medical conditions favour the increase of many genes that cause physical defects and susceptibility to certain diseases, other genes that affect mental disease and mental defect are probably diminishing as a result of the low reproductive rate and the institutionalization of the carriers. It is probable that efforts at control will be strengthened by increased use of sterilization in some areas, by increased use of more effective means of birth control, and by a more eugenic use of heredity counselling clinics. Each year increasing numbers of defects are being defined, the type of transmission understood, and, in many cases, means found for spotting carriers. The immediate solution to the dilemma seems to lie in educating those persons known to be carrying a seriously defective gene to their societal responsibility, and thereby, hopefully, to convince them to surrender their right to bear children who might be defective and burden them as well as society. The possibility of "correcting" nature's mistakes at their source, the deoxyribonucleic acid molecule (DNA) of the gene, is a hope for the future. Such genetic engineering has been proposed by several prominent scientists and termed "algeny."

Psychological traits. With responsibility for work on mental and physical defects transferred to the professions of medicine and public health, eugenists are now giving major attention to the genetic factors that provide the base for the development of such complexly inherited traits as intelligence, character, and personality. The genetic base for these psychological traits is provided by multiple genes in various combinations. They may be broken up and recombined with other genes from one generation to another, making them difficult to trace in family lines. It takes time to locate these genes and trace the processes involved in their transmission. Fortunately, the efficacy of selection does not depend on a full knowledge of the genetic mechanism. For the present, at least, eugenists will be guided by familial incidence of intelligence and personality and by the survival of family stocks of different types.

Massive evidence from studies on identical and fraternal twins, nontwin siblings, and adopted children reared away from their natural parents indicates that differences in genetic factors play an important part in psychological differences between individuals. These were Galton's findings, and they have been confirmed by larger and more carefully controlled studies. Identical twins are extremely similar; even when reared apart from infancy, they resemble each other in physical as well as mental characteristics to a degree greater than that shown by fraternal twins reared together.

Children tend to be like their parents in intelligence not only because of their common heredity, but also because they usually share a common environment. The evidence on differences in personality is not so clear, largely because these traits are more difficult to measure. All studies indicate, however, that personality, like intelligence, is the product of interaction between environment and the genetic base. This is shown in the extremes of personality, like schizophrenia, which can often be measured.

Environmental and socioeconomic factors. As work in the field of individual differences has progressed, scientists have come to feel that the old distinction between heredity and environment is no longer useful; both are a part of the processes that begin when the cell is fertilized and starts its separate life; both operate throughout the individual's life; and both contribute to making each individual different from every other individual. Genetic factors can be given value only in relation to specific environments; some qualities that are socially valuable in one environment may be harmful in another.

Apparent racial differences in intelligence

Racial groups differ in their average response to intelligence tests, but these tests invariably involve environmental factors that favour those culturally more fortunate. Even when differences in such factors as place of residence, income, or length of schooling can be controlled, the results may still be influenced by factors affecting motivation and behaviour.

In studies of socioeconomic, or other, groups, there is no evidence of genetic differences, except perhaps in the case of comparisons between certain of the professional classes and the rest of the population, in which there is some indication that the genetic factors for response to intelligence tests may be slightly superior. Differences between socioeconomic classes in intelligence-test response are obviously strongly influenced by differences in environment in much the same way as occurs among racial groups. It is as yet impossible to measure the relative genetic potentialities of any large population groups. But differences between individuals within each group can be measured, and it is these differences that are the proper concern of eugenists today.

Until very recently, from 30 to 50 percent of all children died before maturity, and death was an important factor in the survival of particular stocks. The evidence from some areas where people are still living under primitive conditions seems to favour the belief that under conditions of life as it was everywhere before the mid-18th century, the general tendency was toward the survival of more competent stocks. If so, this situation changed in the industrialized countries, when the death rate began an accelerated decline, dropping to 12 or less per 1,000 by 1940. In the mid-20th century, about 95 percent of all children lived to age 30, and differential deaths had little effect on differential survival.

Social-class differences in family size

While this change was taking place, social-class differences in births were widening with the introduction of birth-control measures, which reached first into the more educated and wealthy classes. With differential births greatly favouring the less educated, who made little effective use of birth control, the inverse relation between births and education or socioeconomic status widened rapidly. In the United States this inverse relationship reached its peak with the postponement of births and marriages in the last years of the Great Depression. For the period 1935–40 women with one to three years of college education had a net reproduction rate 33 percent below that required for replacement of their numbers, and women with four years of high school had a rate 26 percent below replacement, while women at the elementary school level had a rate twice that of the college women and well above that required for replacement.

The upsurge of births after World War II and the spread of birth control among the less educated changed this trend. By 1960, among women 35 to 40 years old who had prac-

tically completed their childbearing, high-school women and women with one to three years of college had borne more children than the number needed for replacement. Women with less than eight years of elementary school education had had about the same number of children as those who had completed their childbearing during the Depression. This trend seems likely to continue. Extensive studies made about 1960 on national samples of the U.S. population indicate that the number of children actually desired by couples in the "blue collar" (factory or industrial) group is less than the number desired by couples in the "white collar" (office or business) group. When both the husband and the wife were college graduates, only 7 percent of the last pregnancies were unwanted, while among the couples at grade school level 33 percent had not wanted their last pregnancy. Of the college couples, 92 percent effectively used contraception, while only 69 percent of the grade school couples used contraception, and they used it less effectively.

It is expected that birth control may soon be used as widely and as effectively at the lowest educational levels as at the highest. Should that occur, present trends suggest that educational and social class differentials in births may soon bear a direct relation to size of family. The public generally, and many scientists, have assumed that an inverse relation between size of family and education or socioeconomic class must represent an unfavourable, or dysgenic, trend, even though there is little evidence for genetic differences between classes. If the relation between size of family and educational level becomes direct, with the more educated groups having the greater proportion of children, the trend would be generally taken as favourable, or eugenic. But its greatest significance would be an increase in the proportion of children brought up in above-average home surroundings. The major opportunity for genetic improvement will be found in selection of individuals within each group.

There is some limited evidence that at the present time in the United States the abler and the more intelligent members of any group tend to have the most children. In studies made as early as 1927 on graduates of Yale and Harvard universities, and again on Princeton University graduates in 1939, the men were rated by a committee of classmates for degree of success some years after graduation. In each group the highest-rated married men averaged over two children apiece, while the lowest-rated married men averaged less than one and one-half children. Many studies have indicated that the probability of being childless increases as IQ decreases. There is thus some evidence that in attempting to increase the comparative fertility of the more intelligent and successful individuals, eugenists are working with and not against present trends, as has been claimed by some.

Individual voluntary selection on the increase

A number of changes taking place in industrial societies today are favourable to individual selection. There are more opportunities for finding work appropriate to one's genetic capacities. There is a wider choice of mates in a highly mobile society, which favours the production of novel types and offers greater possibilities if there is selection. The trend toward more education for all classes would help make birth control selective within each class. Also, the continuing improvement of birth control methods should improve voluntary forms of selection at all levels of education. To further these trends, eugenists urge better training for choice of mate and for marriage, a more general individual and community sense of responsibility for the parents who have more than one or two children, and continuing study of all the psychological, social, and economic pressures and rewards that would tend toward a voluntary and favourable selection of births within every occupational group.

(F.H.O.)

ECOLOGY

Long unfamiliar to the public, and relegated to a second-class status by many in the world of science, ecology emerged in the late 20th century as one of the most popular and most important aspects of biology. It has become painfully evident that the most pressing problems in the affairs of men—expanding populations, food scarcities, environmental pollution, and all the attendant sociological and political problems—are to a great degree ecological.

The word ecology was coined by a German zoologist, Ernst Haeckel, who applied the term *oekologie* to the "relation of the animal both to its organic as well as its inorganic environment." The word comes from the Greek *oikos,* meaning "household, home, or place to live." Thus ecology deals with the organism and its environment. The word environment includes both other organisms and physical surroundings. It involves relationships between individuals within a population and between individuals of different populations. These interactions between individuals, between populations, and between organisms and their environment form ecological systems, or ecosystems. Ecology has been defined variously as "the study of the interrelationships of organisms with their environment and each other," as "the economy of nature," and as "the biology of ecosystems."

Historical background. Ecology had no firm beginnings. It evolved from the natural history of the Greeks, particularly Theophrastus, a friend and associate of Aristotle. He first described the interrelationships between organisms and between organisms and their nonliving environment. Later foundations for modern ecology were laid in the early work of plant and animal physiologists.

In the early and mid-1900s two groups of botanists, one in Europe and the other in America, studied plant communities from two different points of view. The European botanists concerned themselves with the study of the composition, structure, and distribution of plant communities. The American botanists studied the development of plant communities, or succession. Both plant and animal ecology developed separately until American biologists emphasized the interrelation of both plant and animal communities as a biotic whole.

During the same period interest in population dynamics developed. The study of population dynamics received special impetus in the early 19th century, after Thomas Malthus called attention to the conflict between expanding populations and the capability of the earth to supply food. R. Pearl (1920), A.J. Lotka (1925), and V. Volterra (1926) developed mathematical foundations for the study of populations, and these studies led to experiments on the interaction of predators and prey, competitive relationships between species, and the regulation of populations. Investigations of the influence of behaviour on populations was stimulated by the recognition in 1920 of territoriality in nesting birds. Concepts of instinctive and aggressive behaviour were developed by K. Lorenz and N. Tinbergen, and the role of social behaviour in the regulation of populations was explored by V.C. Wynne-Edwards.

While some ecologists were studying the dynamics of communities and populations, others were concerned with energy-budgets. In 1920, August Thienemann, a German freshwater biologist, introduced the concept of trophic, or feeding, levels, by which the energy of food is transferred through a series of organisms, from green plants (the producers) up to several levels of animals (the consumers). An English animal ecologist, C.E. Elton (1927), further developed this approach with the concept of ecological niches and pyramids of numbers. Two American freshwater biologists, E. Birge and C. Juday, in the 1930s, in measuring the energy budgets of lakes, developed the idea of primary production, *i.e.,* the rate at which food energy is generated, or fixed, by photosynthesis. Modern ecology came of age in 1942 with the development, by R.L. Lindeman of the United States, of the trophic-dynamic concept of ecology, which details the flow of energy through the ecosystem. Quantified field studies of energy flow through ecosystems were further developed by Eugene and Howard Odum of the United States; similar early work on the cycling of nutrients was done by J.D. Ovington of England and Australia.

The study of both energy flow and nutrient cycling was stimulated by the development of new techniques—radioisotopes, microcalorimetry, computer science, and applied mathematics—that enabled ecologists to label, trace, and measure the movement of particular nutrients and

energy through the ecosystems. These modern methods encouraged a new stage in the development of ecology—systems ecology, which is concerned with the structure and function of ecosystems.

he unifying concept f ecology

Until the late 20th century ecology lacked a strong conceptual base. Modern ecology, however, is now focussed on the concept of the ecosystem, a functional unit consisting of interacting organisms and all aspects of the environment in any specific area. It contains both the nonliving (abiotic) and living (biotic) components through which nutrients are cycled and energy flows. To accomplish this cycling and flow, ecosystems must possess a number of structured interrelationships between soil, water, and nutrients, on the one hand, and producers, consumers, and decomposers on the other. Ecosystems function by maintaining a flow of energy and a cycling of materials through a series of steps of eating and being eaten, of utilization and conversion, called the food chain. Ecosystems tend toward maturity, or stability, and in doing so they pass from a less complex to a more complex state. This directional change is called succession. Whenever an ecosystem is used, and that exploitation is maintained—as when a pond is kept clear of encroaching plants or a woodland is grazed by domestic cattle—the maturity of the ecosystem is effectively postponed. The major functional unit of the ecosystem is the population. It occupies a certain functional niche, related to its role in energy flow and nutrient cycling. Both the environment and the amount of energy fixation in any given ecosystem are limited. When a population reaches the limits imposed by the ecosystem, its numbers must stabilize or, failing this, decline from disease, starvation, strife, low reproduction, or other behavioral and physiological reactions. Changes and fluctuations in the environment represent selective pressure upon the population to which it must adjust. The ecosystem has historical aspects: the present is related to the past and the future to the present. Thus the ecosystem is the one concept that unifies plant and animal ecology, population dynamics, behaviour, and evolution.

Areas of study. Of necessity, ecology is a multidisciplinary science. It involves plant and animal biology, taxonomy, physiology, genetics, behaviour, meteorology, pedology, geology, sociology, anthropology, physics, chemistry, mathematics, and electronics. Often it is difficult to draw a sharp line between ecology and any of these, for all impinge on it. The same situation exists also within ecology. In understanding the interactions between the organism and the environment or between organisms, it is often difficult to separate behaviour from population dynamics, behaviour from physiology, adaptation from evolution and genetics, animal ecology from plant ecology.

Ecology developed along two lines: the study of plants and the study of animals. Plant ecology concerns the relationships of plants to other plants and their environment. The approach is largely descriptive of the vegetational and floristic composition of an area and usually ignores the influence of animals on the plants. Animal ecology concerns the study of population dynamics, distribution, behaviour, and the interrelationships of animals and their environment. Because animals depend upon plants for food and shelter, animal ecology cannot be fully understood without a considerable background of plant ecology. This is particularly true in applied areas of ecology—wildlife and range management.

Both plant and animal ecology may be approached as the study of the interrelations of an individual organism with its environment, called autecology, or as the study of groups of organisms, called synecology.

Autecology and synecology

Autecology, in many ways the classical study of ecology, is experimental and inductive. Because it is usually concerned with the relationship of an organism to one or more variables such as humidity, light, salinity, or nutrient levels, it is easily quantified and lends itself to experimental design both in the field and the laboratory. It has borrowed techniques from chemistry, physics, and physiology. Autecology has contributed at least two important concepts: the constancy of interaction between an organism and its environment, and the genetic adaptability of local populations to local environmental conditions.

Synecology, on the other hand, is philosophical and deductive. It is largely descriptive and not easily quantified and contains a bewildering array of terminology. Only recently, since the advent of the electronic and atomic ages, has synecology developed the tools to study complex systems and enter an experimental phase. Important concepts developed by synecology are those concerned with nutrient cycling, energy budgets, and ecosystem development. Synecology has strong ties with pedology, geology, meteorology, and cultural anthropology.

Synecology may be subdivided according to environmental types, as terrestrial or aquatic. Terrestrial ecology, which may be further subdivided into forest, grassland, arctic, and desert ecology, concerns such aspects of terrestrial ecosystems as microclimate, soil chemistry, soil fauna, hydrologic cycles, ecogenetics, and productivity. Terrestrial ecosystems are more influenced by organisms and subject to much wider environmental fluctuations than are aquatic ecosystems. Aquatic ecosystems are affected more by the condition of the water and resist such environmental variables as temperature. Because the physical environment is so important in controlling aquatic ecosystems, considerable attention is paid to the chemical and physical characteristics of the ecosystem, such as the currents and the chemical composition of the water. By convention, aquatic ecology, called limnology, is limited to freshwater stream ecology and lake ecology. The former concerns life in flowing waters; the latter, life in relatively still water. Marine ecology deals with life in the open sea and in estuaries.

Other ecological approaches concern specialized areas. The study of the geographic distribution of plants and animals is ecological plant and animal geography. The study of population growth, mortality, natality, competition, and predator–prey relations is population ecology. The study of the genetics and ecology of local races and distinct species is ecological genetics. The study of the behavioral responses of animals to their environment, and of social interactions as they affect population dynamics, is behavioral ecology. Investigations of interactions between the physical environment and the organism fall under ecoclimatology and physiological ecology. The study of groups of organisms is community ecology (though it is difficult to separate it from studies of bioenergetics, biogeochemical cycles, and trophic-dynamic aspects of the community or ecosystem ecology). That part of ecosystem ecology concerned with the analysis and understanding of the structure and function of ecosystems by the use of applied mathematics, mathematical models, and computer programs is systems ecology. Systems ecology, concentrating on input and output analysis, has stimulated the rapid development of applied ecology, concerned with the application of ecological principles to the management of natural resources, agricultural production, and problems of environmental pollution.

Systems ecology

Methods in ecology. Because ecologists work with living systems possessing numerous variables, the techniques used by physicists and chemists, mathematicians and engineers, require modification; they are not easily applied nor are the results as precise as those obtained in other sciences. It is relatively simple, for example, for a physicist to measure gain and loss of heat from metals or other inanimate objects, which possess certain constants of conductivity, expansion, surface features, and the like. To determine the heat exchange between an animal and its environment, however, a physiological ecologist is confronted with an array of almost unquantifiable variables and has the formidable task of both gathering the numerous data and analyzing them. Ecological measurements probably never will be as precise or as subject to the same ease of analysis as measurements in physics, chemistry, or certain quantifiable areas of biology.

In spite of these problems, various aspects of the environment can be determined by physical and chemical means, ranging from simple chemical identifications and physical measurements to the use of sophisticated mechanical apparatus. The development of biostatistics and proper experimental design, and the improvements in methods of sampling, permit a quantified statistical approach to

the study of ecology. Because of the extreme difficulties of controlling environmental variables in the field, studies involving the use of experimental design are largely confined to the laboratory and to controlled field experiments designed to test the effects of only one variable or several variables. The use of statistical procedures, and the application of computer science to mathematical models based on data obtained from the field, are providing new insights into population interactions and ecosystem function. Mathematical programming is becoming increasingly important in applied ecology, especially in the management of natural resources and agricultural problems having an ecological basis.

Controlled environmental chambers enable experimenters to maintain plants and animals under known conditions of light, temperature, humidity, and daylength so that the effects of each variable (or combination of variables) on the organism can be studied. Biotelemetry and other electronic tracking equipment, products of the space age, permit the rapid and nondestructive sampling of plant and animal populations. Such tools enable ecologists to follow from a distance the movements and behaviour of a free-ranging animal by radio signals beamed from a sender attached to the organism. Radioisotopes are used for tracing the pathways of nutrients through ecosystems, for determining the time and extent of transfer of energy and nutrients through the different components of the ecosystem, and for the determination of food chains. The use of laboratory microcosms—aquatic and soil microecosystems, consisting of biotic and nonbiotic material from natural ecosystems, held under conditions similar to those found in the field—are useful in determining rates of nutrient cycling, ecosystem development, and other functional aspects of ecosystems. Microcosms enable the ecologist to duplicate experiments and to perform experimental manipulation on them.

(R.L.Sm.)

The study of types of living organisms

ZOOLOGY

Zoology, the study of animals, includes both the inquiry into individual animals and their constituent parts, even to the molecular level, and the inquiry into animal populations, entire faunas, and the relationships of animals to each other, to plants, and to the nonliving environment. Though this wide range of studies results in some isolation of specialties within zoology, the conceptual integration in the contemporary study of living things that has occurred in recent years emphasizes the structural and functional unity of life rather than its diversity.

Historical background. Prehistoric man's survival as a hunter defined his relation to other animals, which were a source of food and danger. As man's cultural heritage developed, animals were variously incorporated into man's folklore and philosophical awareness as fellow living creatures. Domestication of animals forced man to take a systematic and measured view of animal life, especially after urbanization necessitated a constant and large supply of animal products.

The role of the Greeks

Study of animal life by the ancient Greeks became more rational, if not yet scientific, in the modern sense, after the cause of disease—until then thought to be demons—was postulated by Hippocrates to result from a lack of harmonious functioning of body parts. The systematic study of animals was encouraged by Aristotle's extensive descriptions of living things, his work reflecting the Greek concept of order in nature and attributing to nature an idealized rigidity.

In Roman times Pliny brought together in 37 volumes a treatise, *Historia naturalis,* that was an encyclopaedic compilation of both myth and fact regarding celestial bodies, geography, animals and plants, metals, and stone. Volumes VII to XI concern zoology; volume VIII, which deals with the land animals, begins with the largest one, the elephant. Although Pliny's approach was naïve, his scholarly effort had a profound and lasting influence as an authoritative work.

Zoology continued in the Aristotelian tradition for many centuries in the Mediterranean region and by the Middle Ages, in Europe, it had accumulated considerable folklore, superstition, and moral symbolisms, which were added to otherwise objective information about animals. Gradually, much of this misinformation was sifted out: naturalists became more critical as they compared directly observed animal life in Europe with that described in ancient texts. The use of the printing press in the 15th century made possible an accurate transmission of information. Moreover, mechanistic views of life processes (*i.e.,* that physical processes depending on cause and effect can apply to animate forms) provided a hopeful method for analyzing animal functions; for example, the mechanics of hydraulic systems were part of William Harvey's argument for the circulation of the blood—although Harvey remained thoroughly Aristotelian in outlook. In the 18th century, zoology passed through reforms provided by both the system of nomenclature of Carolus Linnaeus and the comprehensive works on natural history by Georges-Louis Leclerc de Buffon; to these were added the contributions to comparative anatomy by Georges Cuvier in the early 19th century.

Physiological functions, such as digestion, excretion, and respiration, were easily observed in many animals, though they were not as critically analyzed as was blood circulation.

Following the introduction of the word cell in the 17th century and microscopic observation of these structures throughout the 18th century, the cell was incisively defined as the common structural unit of living things in 1839 by two Germans: Matthias Schleiden and Theodor Schwann. In the meanwhile, as the science of chemistry developed, it was inevitably extended to an analysis of animate systems. In the middle of the 18th century the French physicist René Antoine Ferchault de Réaumer demonstrated that the fermenting action of stomach juices is a chemical process. And in the mid-19th century the French physician and physiologist Claude Bernard drew upon both the cell theory and knowledge of chemistry to develop the concept of the stability of the internal bodily environment, now called homeostasis.

The cell concept influenced many biological disciplines, including that of embryology, in which cells are important in determining the way in which a fertilized egg develops into a new organism. The unfolding of these events—called epigenesis by Harvey—was described by various workers, notably the German-trained comparative embryologist Karl von Baer, who was the first to observe a mammalian egg within an ovary. Another German-trained embryologist, Christian Heinrich Pander, introduced in 1817 the concept of germ, or primordial, tissue layers into embryology.

In the latter part of the 19th century, improved microscopy and better staining techniques using aniline dyes, such as hematoxylin, provided further impetus to the study of internal cellular structure.

Darwin and the theory of evolution

By this time Darwin had made necessary a complete revision of man's view of nature with his theory that biological changes in species occur through the process of natural selection. The theory of evolution—that organisms are continuously evolving into highly adapted forms—required the rejection of the static view that all species are especially created and upset the Linnaean concept of species types. Darwin recognized that the principles of heredity must be known to understand how evolution works; but, even though the concept of hereditary factors had by then been formulated by Mendel, Darwin never heard of his work, which was essentially lost until its rediscovery in 1900.

Genetics has developed in the 20th century and now is essential to many diverse biological disciplines. The discovery of the gene as a controlling hereditary factor for all forms of life has been a major accomplishment of modern biology. There has also emerged clearer understanding of the interaction of organisms with their environment. Such ecological studies help not only to show the interdependence of the three great groups of organisms—plants, as producers; animals, as consumers; and fungi and many bacteria, as decomposers—but they also provide informa-

tion essential to man's control of the environment and, ultimately, to his survival on Earth. Closely related to this study of ecology are inquiries into animal behaviour, or ethology. Such studies are often cross disciplinary in that ecology, physiology, genetics, development, and evolution are combined as man attempts to understand why an organism behaves as it does. This approach now receives substantial attention because it seems to provide useful insight into man's biological heritage—that is, the historical origin of man from nonhuman forms.

The emergence of animal biology has had two particular effects on classical zoology. First, and somewhat paradoxically, there has been a reduced emphasis on zoology as a distinct subject of scientific study; for example, workers think of themselves as geneticists, ecologists, or physiologists who study animal rather than plant material. They often choose a problem congenial to their intellectual tastes, regarding the organism used as important only to the extent that it provides favourable experimental material. Current emphasis is, therefore, slanted toward the solution of general biological problems; contemporary zoology thus is to a great extent the sum total of that work done by biologists pursuing research on animal material.

Second, there is an increasing emphasis on a conceptual approach to the life sciences. This has resulted from the concepts that emerged in the late 19th and early 20th centuries: the cell theory; natural selection and evolution; the constancy of the internal environment; the basic similarity of genetic material in all living organisms; and the flow of matter and energy through ecosystems. The lives of microbes, plants, and animals now are approached using theoretical models as guides rather than by following the often restricted empiricism of earlier times. This is particularly true in molecular studies, in which the integration of biology with chemistry allows the techniques and quantitative emphases of the physical sciences to be used effectively to analyze living systems.

Areas of study. Although it is still useful to recognize many disciplines in animal biology—*e.g.,* anatomy or morphology; biochemistry and molecular biology; cell biology; developmental studies (embryology); ecology; ethology; evolution; genetics; physiology; and systematics—the research frontiers occur as often at the interfaces of two or more of these areas as within any given one.

Anatomy or morphology. Descriptions of external form and internal organization are among the earliest records available regarding the systematic study of animals. Aristotle was an indefatigable collector and dissector of animals. He found differing degrees of structural complexity, which he described with regard to ways of living, habits, and body parts. Although Aristotle had no formal system of classification, it is apparent that he viewed animals as arranged from the simplest to the most complex in an ascending series. Since man was even more complex than animals and, moreover, possessed a rational faculty, he therefore occupied the highest position and a special category. This hierarchical perception of the animate world proved to be useful in every century to the present, except that in the modern view there is no such "scale of nature," and there is change in time by evolution from the simple to the complex.

Aristotle's attempt at classification

After the time of Aristotle, Mediterranean science was centred at Alexandria, where the study of anatomy, particularly the central nervous system, flourished and, in fact, first became recognized as a discipline. Galen studied anatomy at Alexandria in the 2nd century and later dissected many animals. Much later, the contributions of the Renaissance anatomist Andreas Vesalius, though made in the context of medicine, as were those of Galen, stimulated to a great extent the rise of comparative anatomy. During the latter part of the 15th century and throughout the 16th century, there was a strong tradition in anatomy; important similarities were observed in the anatomy of different animals, and many illustrated books were published to record these observations.

But anatomy remained a purely descriptive science until the advent of functional considerations in which the correlation between structure and function was consciously investigated; as by French biologists Buffon and Cuvier. Cuvier cogently argued that a trained naturalist could deduce from one suitably chosen part of an animal's body the complete set of adaptations that characterized the organism. Because it was obvious that organisms with similar parts pursue similar habits, they were placed together in a system of classification. Cuvier pursued this viewpoint, which he called the theory of correlations, in a somewhat dogmatic manner and placed himself in opposition to the romantic natural philosophers, such as the German intellectual Johann Wolfgang von Goethe, who saw a tendency to ideal types in animal form. The tension between these schools of thought—adaptation as the consequence of necessary bodily functions and adaptation as an expression of a perfecting principle in nature—runs as a leitmotiv through much of biology, with overtones extending into the early 20th century.

The twin concepts of homology (similarity of origin) and analogy (similarity of appearance), in relation to structure, are the creation of the 19th-century British anatomist Richard Owen. Although they antedate the Darwinian view of evolution, the anatomical data on which they were based became, largely as a result of the work of the German comparative anatomist Carl Gegenbaur, important evidence in favour of evolutionary change, despite Owen's steady unwillingness to accept the view of diversification of life from a common origin.

In summary, anatomy moved from a purely descriptive phase as an adjunct to classificatory studies, into a partnership with studies of function and became, in the 19th century, a major contributor to the concept of evolution.

Taxonomy or systematics. Not until the work of Carolus Linnaeus did the variety of life receive a widely accepted systematic treatment. Linnaeus strove for a "natural method of arrangement," one that is now recognizable as an intuitive grasp of homologous relationships, reflecting evolutionary descent from a common ancestor; however, the natural method of arrangement sought by Linnaeus was more akin to the tenets of idealized morphology because he wanted to define a "type" form as epitomizing a species.

The binomial nomenclature of Linnaeus

It was in the nomenclatorial aspect of classification that Linnaeus created a revolutionary advance with the introduction of a Latin binomial system: each species received a Latin name, which was not influenced by local names and which invoked the authority of Latin as a language common to the learned people of that day. The Latin name has two parts. The first word in the Latin name for the dog, *Canis familiaris,* for example, indicates the larger category, or genus, to which dogs belong; the second word is the name of the species within the genus. In addition to species and genera, Linnaeus also recognized other classificatory groups, or taxa (singular taxon), which are still used; namely, order, class, and kingdom, to which have been added family (between genus and order) and phylum (between class and kingdom). Each of these can be divided further by the appropriate prefix of sub- or super-, as in subfamily or superclass. Linnaeus' great work, the *Systema naturae,* went through 12 editions during his lifetime; the 13th, and final, edition appeared posthumously. Although his treatment of the diversity of living things has been expanded in detail, revised in terms of taxonomic categories, and corrected in the light of continuing work—for example, Linnaeus treated whales as fish—it still sets the style and method, even to the use of Latin names, for contemporary nomenclatorial work.

Linnaeus sought a natural method of arrangement, but he actually defined types of species on the basis of idealized morphology. The greatest change from Linnaeus' outlook is reflected in the phrase "the new systematics," which was introduced in the 20th century and through which an explicit effort is made to have taxonomic schemes reflect evolutionary history. The basic unit of classification, the species, is also the basic unit of evolution—*i.e.,* a population of actually or potentially interbreeding individuals. Such a population shares, through interbreeding, its genetic resources. In so doing, it creates the gene pool—its total genetic material—that determines the biological resources of the species and on which natural selection continuously acts. This approach has guided work on

classifying animals away from somewhat arbitrary categorization of new species to that of recreating evolutionary history (phylogeny) and incorporating it in the system of classification. Modern taxonomists or systematists, therefore, are among the foremost students of evolution.

Physiology. The practical consequences of physiology have always been an unavoidable human concern, in both medicine and animal husbandry. Inevitably, from Hippocrates to the present, practical knowledge of human bodily function has accumulated along with that of domestic animals and plants. This knowledge has been expanded, especially since the early 1800s, by experimental work on animals in general, a study known as comparative physiology. The experimental dimension had wide applications following Harvey's demonstration of the circulation of blood. From then on, medical physiology developed rapidly; notable texts appeared, such as Albrecht von Haller's eight-volume work *Elementa Physiologiae Corporis Humani* (*Elements of Human Physiology*), which had a medical emphasis. Toward the end of the 18th century the influence of chemistry on physiology became pronounced through Antoine Lavoisier's brilliant analysis of respiration as a form of combustion. This French chemist not only determined that oxygen was consumed by living systems but also opened the way to further inquiry into the energetics of living systems. His studies further strengthened the mechanistic view, which holds that the same natural laws govern both the inanimate and the animate realms.

Physiological principles achieved new levels of sophistication and comprehensiveness with Bernard's concept of constancy of the internal environment, the point being that only under certain constantly maintained conditions is there optimal bodily function. His rational and incisive insights were augmented by concurrent developments in Germany, where Johannes Müller explored the comparative aspects of animal function and anatomy, and Justus von Liebig and Carl Ludwig applied chemical and physical methods, respectively, to the solution of physiological problems. As a result, many useful techniques were advanced—*e.g.*, means for precise measurement of muscular action and changes in blood pressure and means for defining the nature of body fluids.

The study of organ systems

By this time the organ systems—circulatory, digestive, endocrine, excretory, integumentary, muscular, nervous, reproductive, respiratory, and skeletal—had been defined, both anatomically and functionally, and research efforts were focussed on understanding these systems in cellular and chemical terms, an emphasis that continues to the present and has resulted in specialties in cell physiology and physiological chemistry. General categories of research now deal with the transportation of materials across membranes; the metabolism of cells, including synthesis and breakdown of molecules; and the regulation of these processes.

Interest has also increased in the most complex of physiological systems, the nervous system. Much comparative work has been done by utilizing animals with structures especially amenable to various experimental techniques; for example, the large nerves in squids have been extensively studied in terms of the transmission of nerve impulses, and insect and crustacean eyes have yielded significant information on patterns of sensory inputs. Most of this work is closely associated with studies on animal orientation and behaviour. Although the contemporary physiologist often studies functional problems at the molecular and cellular levels, he is also aware of the need to integrate cellular studies into the many-faceted functions of the total organism.

Embryology, or developmental studies. Embryonic growth and differentiation of parts have been major biological problems since ancient times. A 17th-century explanation of development assumed that the adult existed as a miniature—a homunculus—in the microscopic material that initiates the embryo. But in 1759 the German physician Caspar Friedrick Wolff firmly introduced into biology the interpretation that undifferentiated materials gradually become specialized, in an orderly way, into adult structures. Although this epigenetic process is now accepted as characterizing the general nature of development in both plants and animals, many questions remain to be solved. The French physician Marie François Xavier Bichat declared in 1801 that differentiating parts consist of various components called tissues; with the subsequent statement of the cell theory, tissues were resolved into their cellular constituents. The idea of epigenetic change and the identification of structural components made possible a new interpretation of differentiation. It was demonstrated that the egg gives rise to three essential germ layers out of which specialized organs, with their tissues, subsequently emerge. Then, following his own discovery of the mammalian ovum, von Baer in 1828 usefully applied this information when he surveyed the development of various members of the vertebrate groups. At this point, embryology, as it is now recognized, emerged as a distinct subject.

The concept of cellular organization had an effect on embryology that continues to the present day. In the 19th century, cellular mechanisms were considered essentially to be the basis for growth, differentiation, and morphogenesis, or molding of parts. The distribution of the newly formed cells of the rapidly dividing zygote (fertilized egg) was precisely followed to provide detailed accounts not only of the time and mode of germ layer formation but also of the contribution of these layers to the differentiation of tissues and organs. Such descriptive information provided the background for experimental work aimed at elucidating the role of chromosomes and other cellular constituents in differentiation. About 1895, before the formulation of the chromosomal theory of heredity, Theodor Boveri demonstrated that chromosomes show continuity from one cell generation to the next. In fact, biologists soon concluded that in all cells arising from a fertilized egg, half the chromosomes are of maternal and half of paternal origin. The discovery of the constant transmission of the original chromosomal endowment to all cells of the body served to deepen the mystery surrounding the factors that determine cellular differentiation.

The present view is that differential activity of genes is the basis for cellular and tissue differentiation; that is, although the cells of a multicellular body contain the same genetic information, different genes are active in different cells. The result is the formation of various gene products, which regulate the functional and structural differentiation of cells. The actual mechanism involved in the inactivation of certain genes and the activation of others, however, has not yet been established. That cells can move extensively throughout the embryo and selectively adhere to other cells, thus starting tissue aggregations, also contributes to development as does the fate of cells—*i.e.*, certain ones continue to multiply, others stop, and some die.

Research methods in embryology now exploit many experimental situations: both unicellular and multicellular forms; regeneration (replacement of lost parts) and normal development; and growth of tissues outside and inside the host. Hence, the processes of development can be studied with material other than embryos; and the study of embryology has become incorporated into the more inclusive subdiscipline of developmental biology.

Evolutionism. Darwin was not the first to speculate that organisms can change from generation to generation and so evolve, but he was the first to propose a mechanism by which the changes are accumulated. He proposed that heritable variations occur in conjunction with a never-ending competition for survival and that the variations favouring survival are automatically preserved. In time, therefore, the continued accumulation of variations results in the emergence of new forms. Because the variations that are preserved relate to survival, the survivors are highly adapted to their environment. To this process Darwin gave the apt name natural selection.

Natural selection

Many of Darwin's predecessors, notably Jean-Baptiste Lamarck, were willing to accept the idea of species variation, even though to do so meant denying the doctrine of special creation and the static-type species of Linnaeus. But they argued that some idealized perfecting principle, expressed through the habits of an organism, was the basis of variation. The contrast between the romanticism of Lamarck and the objective analysis of Darwin clearly

reveals the type of revolution provoked by the concept of natural selection. Although mechanistic explanations had long been available to biologists—forming, for example, part of Harvey's explanation of blood circulation—they did not pervade the total structure of biological thinking until the advent of Darwinism.

There were two immediate consequences of Darwin's viewpoints. One has involved a reappraisal of all subject areas of biology; reinterpretations of morphology and embryology are good examples. The comparative anatomy of the British anatomist Owen became a cornerstone of the evidence for evolution, and German anatomists provided the basis for the comment that evolutionary thinking was born in England but gained its home in Germany. The reinterpretation of morphology carried over into the study of fossil forms, as paleontologists sought and found evidence of gradual change in their study of fossils. But some workers, although accepting evolution in principle, could not easily interpret the changes in terms of natural selection. The German paleontologist Otto Schindewolf, for example, found in shelled mollusks called ammonites evidence of progressive complexity and subsequent simplification of forms. The American paleontologist George Gaylord Simpson, however, has been a consistent interpreter of vertebrate fossils by Darwinian selection. Embryology was seen in an evolutionary light when the German zoologist Ernst Haeckel proposed that the epigenetic sequence of embryonic development (ontogeny) repeated its evolutionary history (phylogeny). Thus, the presence of gill clefts in the mammalian embryo and also in less highly evolved vertebrates can be understood as a remnant of a common ancestor.

The other consequence of Darwinism—to make more explicit the origin and nature of heritable variations and the action of natural selection on them—depended on the emergence of the following: genetics and the elucidation of the rules of Mendelian inheritance; the concept of the gene as the unit of inheritance; and the nature of gene mutation. The development of these ideas provided the basis for the genetics of natural populations.

The subject of population genetics began with the Mendelian laws of inheritance and now takes into account selection, mutation, migration (movement into and out of a given population), breeding patterns, and population size. These factors affect the genetic makeup of a group of organisms that either interbreed or have the potential to do so; *i.e.*, a species. Accurate appraisal of these factors allows precise predictions regarding the content of a given gene pool over significant periods of evolutionary time. From work involving population genetics has come the realization, eloquently documented by two contemporary American evolutionists, Theodosius Dobzhansky and Ernst Mayer, that the species is the basic unit of evolution. The process of speciation occurs as a gene pool breaks up to form isolated gene pools. When selection pressures similar to those of the original gene pool persist in the new gene pools, similar functions and the similar structures on which they depend also persist. When selection pressures differ, however, differences arise. Thus, the process of speciation through natural selection preserves the evolutionary history of a species. The record may be discerned not only in the gross, or macroscopic, anatomy of organisms but also in their cellular structure and molecular organization. Significant work now is carried out, for example, on the homologies of the nucleic acids and proteins of different species.

The process of speciation

Genetics. The problem of heredity had been the subject of careful study before its definitive analysis by Mendel. As with Darwin's predecessors, those of Mendel tended to idealize and interpret all inherited traits as being transmitted through the blood or as determined by various "humors" or other vague entities in animal organisms. When studying plants, Mendel was able to free himself of anthropomorphic and holistic explanations. By studying seven carefully defined pairs of characteristics—*e.g.*, tall and short plants; red and white flowers, etc.—as they were transmitted through as many as three successive generations, he was able to establish patterns of inheritance that apply to all sexually reproducing forms. Darwin, who was searching for an explanation of inheritance, apparently never saw Mendel's work, which was published in 1866 in the obscure journal of his local natural history society; it was simultaneously rediscovered in 1900 by three different European geneticists.

Further progress in genetics was made early in the 20th century, when it was realized that heredity factors are found on chromosomes. The term gene was coined for these factors. Studies by the American geneticist Thomas Hunt Morgan on the fruit fly (*Drosophila*), moved animal genetics to the forefront of genetic research. The work of Morgan and his students established such major concepts as the linear array of genes on chromosomes; the exchange of parts between chromosomes; and the interaction of genes in determining traits, including sexual differences. In 1927 one of Morgan's former students, Hermann Muller, used X-rays to induce the mutations (changes in genes) in the fruit fly, thereby opening the door to major studies on the nature of variation.

Meanwhile, other organisms were being used for genetic studies, most notably fungi and bacteria. The results of this work provided insights into animal genetics just as principles initially obtained from animal genetics provided insight into botanical and microbial forms. Work continues not only on the genetics of humans, domestic animals, and plants but also on the control of development through the orderly regulation of gene action in different cells and tissues.

Cellular and molecular biology. Although the cell was recognized as the basic unit of life early in the 19th century, its most exciting period of inquiry has probably occurred since the 1940s. The new techniques developed since that time, notably the perfection of the electron microscope and the tools of biochemistry, have changed the cytological studies of the 19th and early 20th centuries from a largely descriptive inquiry, dependent on the light microscope, into a dynamic, molecularly oriented inquiry into fundamental life processes.

The so-called cell theory, which was enunciated about 1838, was never actually a theory. As Edmund Beecher Wilson, the noted American cytologist, stated in his great work, *The Cell,*

> By force of habit we still continue to speak of the cell 'theory' but it is a theory only in name. In substance it is a comprehensive general statement of fact and as such stands today beside the evolution theory among the foundationstones of modern biology.

More precisely, the cell doctrine was an inductive generalization based on the microscopial examination of certain plant and animal species.

Rudolf Virchow, a German medical officer specializing in cellular pathology, first expressed the fundamental dictum regarding cells in his phrase *omnis cellula e cellula* (all cells from cells). For cellular reproduction is the ultimate basis of the continuity of life; the cell is not only the basic structural unit of life but also the basic physiological and reproductive unit. All areas of biology were affected by the new perspective afforded by the principle of cellular organization. Especially in conjunction with embryology was the study of the cell most prominent in animal biology. The continuity of cellular generations by reproduction also had implications for genetics. It is little wonder, then, that the full title of Wilson's survey of cytology at the turn of the century was *The Cell: Its Role in Development and Heredity.*

Cells as the units of life

The study of the cell nucleus, its chromosomes, and their behaviour served as the basis for understanding the regular distribution of genetic material during both sexual and asexual reproduction. This orderly behaviour of the nucleus made it appear to dominate the life of the cell, for by contrast the components of the rest of the cell appeared to be randomly distributed.

The biochemical study of life had helped in the characterization of the major molecules of living systems—proteins, nucleic acids, fats, and carbohydrates—and in the understanding of metabolic processes. That nucleic acids are a distinctive feature of the nucleus was recognized after their discovery by the Swiss biochemist Johann Friedrich Miescher in 1869. In 1944 a group of American bacteri-

ologists, led by Oswald T. Avery, published work on the causative agent of pneumonia in mice (a bacterium) that culminated in the demonstration that deoxyribonucleic acid (DNA) is the chemical basis of heredity. Discrete segments of DNA correspond to genes, or Mendel's hereditary factors. Proteins were discovered to be especially important for their role in determining cell structure and in controlling chemical reactions.

The advent of techniques for isolating and characterizing proteins and nucleic acids now allows a molecular approach to essentially all biological problems—from the appearance of new gene products in normal development or under pathological conditions to a monitoring of changes in and between nerve cells during the transmission of nerve impulses.

Ecology. The harmony that Linnaeus found in nature, which redounded to the glory and wisdom of a Judaeo-Christian god, was the 18th-century counterpart of the balanced interaction now studied by ecologists. Linnaeus recognized that plants are adapted to the regions in which they grow, that insects play a role in flower pollination, and that certain birds prey on insects and are in turn eaten by other birds. This realization implies, in contemporary terms, the flow of matter and energy in a definable direction through any natural assemblage of plants, animals, and microorganisms. Such an assemblage, termed an ecosystem, starts with the plants, which are designated as producers because they maintain and reproduce themselves at the expense of energy from sunlight and inorganic materials taken from the nonliving environment around them (earth, air, and water). Animals are called consumers because they ingest plant material or other animals that feed on plants, using the energy stored in this food to sustain themselves. Lastly, the organisms known as decomposers, mostly fungi and bacteria, break down plant and animal material and return it to the environment in a form that can be used again by plants in a constantly renewed cycle.

Components of an ecosystem

The term ecology, first formulated by Haeckel in the latter part of the 19th century as "oecology" (from the Greek word for house, *oikos*), referred to the dwelling place of organisms in nature. In the 1890s various European and U.S. scientists laid the foundations for modern work through studies of natural ecosystems and the populations of organisms contained within them.

Animal ecology, the study of consumers and their interactions with the environment, is very complex; attempts to study it usually focus on one particular aspect. Some studies, for example, involve the challenge of the environment to individuals with special adaptations (*e.g.,* water conservation in desert animals); others may involve the role of one species in its ecosystem or the ecosystem itself. Food-chain sequences have been determined for various ecosystems, and the efficiency of the transfer of energy and matter within them has been calculated so that their capacity is known; that is, productivity in terms of numbers of organisms or weight of living matter at a specific level in the food chain can be accurately determined (see ECOSYSTEMS; BIOSPHERE).

In spite of advances in understanding animal ecology, this subject area of zoology does not yet have the major unifying theoretical principles found in genetics (gene theory) or evolution (natural selection).

Ethology. The study of animal behaviour (ethology) is largely a 20th-century phenomenon and is exclusively a zoological discipline. Only animals have nervous systems with their implications for perception, coordination, orientation, learning, and memory. Not until the end of the 19th century did animal behaviour become free from anthropocentric interests and assume an importance in its own right. The British behaviorist C. Lloyd Morgan was probably most influential with his emphasis on parsimonious explanations—*i.e.,* that the explanation "which stands lower in the psychological scale" must be invoked first. This principle is exemplified in the American Herbert Spencer Jennings' pioneering work in 1906 on *The Behavior of Lower Organisms.*

The study of animal behaviour now includes many diverse topics, ranging from swimming patterns of protozoans to socialization and communication among the great apes. Many disparate hypotheses have been proposed in an attempt to explain the variety of behavioral patterns found in animals. They focus on the mechanisms that stimulate courtship in reproductive behaviour of such diverse groups as spiders, crabs, and domestic fowl; and on whole life histories, starting from the special attachment of newly born ducks and goats to their actual mothers or to surrogate (substitute) mothers. The latter phenomenon, called imprinting, has been intensively studied by the Austrian ethologist Konrad Lorenz. Physiologically oriented behaviour now receives much attention; studies range from work on conditioned reflexes to the orientation of crustaceans and the location and communication of food among bees; such diversity of material is one measure of the somewhat diffuse but exciting current state of these studies.

General trends. Zoology has become animal biology—that is, the life sciences display a new unity, one that is founded on the common basis of all life; on the gene pool-species organization of organisms; and on the obligatory interacting of the components of ecosystems. Even as regards the specialized features of animals—involving physiology, development, or behaviour—the current emphasis is on elucidating the broad biological principles that identify animals as one aspect of nature. Zoology has thus given up its exclusive emphasis on animals—an emphasis maintained from Aristotle's time well into the 19th century—in favour of a broader view of life. The successes in applying physical and chemical ideas and techniques to life processes have not only unified the life sciences but have also created bridges to other sciences in a way only dimly foreseen by earlier workers. The practical and theoretical consequences of this trend have just begun to be realized.

Zoology as animal biology

Methods in zoology. Because the study of animals may be concentrated on widely different topics, such as ecosystems and their constituent populations, organisms, cells, and chemical reactions, specific techniques are needed for each kind of investigation. The emphasis on the molecular basis of genetics, development, physiology, behaviour, and ecology has placed increasing importance on those techniques involving cells and their many components. Microscopy, therefore, is a necessary technique in zoology, as are certain physicochemical methods for isolating and characterizing molecules. Computer technology also has a special role in the analysis of animal life. These newer techniques are used in addition to the many classical ones—measurement and experimentation at the tissue, organ, organ system, and organismic levels.

Microscopy. In addition to continuous improvements in the techniques of staining cells, so that their components can be seen clearly, the light used in microscopy can now be manipulated to make visible certain structures in living cells that are otherwise undetectable. The ability to observe living cells is an advantage of light microscopes over electron microscopes; the latter require the cells to be in an environment that kills them. The particular advantage of the electron microscope, however, is its great powers of magnification. Theoretically, it can resolve single atoms; in biology, however, magnifications of lesser magnitude are most useful in determining the nature of structures lying between whole cells and their constituent molecules.

Separation and purification techniques. The characterization of components of cellular systems is necessary for biochemical studies. The specific molecular composition of cellular organelles, for example, affects their shape and density (mass per unit volume); as a result, cellular components settle at different rates (and thus can be separated) when they are spun in a centrifuge.

Other methods of purification rely on other physical properties. Molecules vary in their affinity for the positive or negative pole of an electrical field. Migration to or away from these poles, therefore, occurs at different rates for different molecules and allows their separation; the process is called electrophoresis. The separation of molecules by liquid solvents exploits the fact that the molecules differ in their solubility, and hence they migrate to various degrees as a solvent flows past them. This process, known as chromatography because of the colour used to identify

the position of the migrating materials, yields samples of extraordinarily high purity.

Radioactive tracers. Radioactive compounds are especially useful in biochemical studies involving metabolic pathways of synthesis and degradation. Radioactive compounds are incorporated into cells in the same way as their nonradioactive counterparts. These compounds provide information on the sites of specific metabolic activities within cells and insights into the fates of these compounds in both organisms and the ecosystem.

Computers. Computers process information using their own general language, which is able to complete calculations as complex and diverse as statistical analyses and determinations of enzymatically controlled reaction rates. Computers with access to extensive data files can select information associated with a specific problem and display it to aid the researcher in formulating possible solutions. They help perform routine examinations such as scanning chromosome preparations in order to identify abnormalities in number or shape. Test organisms can be electronically monitored with computers, so that adjustments can be made during experiments; this procedure improves the quality of the data and allows experimental situations to be fully exploited. Computer simulation is important in analyzing complex problems; as many as 100 variables, for example, are involved in the management of salmon fisheries. Simulation makes possible the development of models that approach the complexities of conditions in nature, a procedure of great value in studying wildlife management and related ecological problems.

Importance of computers in research

Applied zoology. Animal-related industries produce food (meats and dairy products), hides, furs, wool, organic fertilizers, and miscellaneous chemical byproducts. There has been a dramatic increase in the productivity of animal husbandry since the 1870s, largely as a consequence of selective breeding and improved animal nutrition. The purpose of selective breeding is to develop livestock whose desirable traits have strong heritable components and can therefore be propagated. Heritable components are distinguished from environmental factors by determining the coefficient of heritability, which is defined as the ratio of variance in a gene-controlled character to total variance.

Another aspect of food production is the control of pests. The serious side effects of some chemical pesticides make extremely important the development of effective and safe control mechanisms. Animal food resources include commercial fishing. The development of shellfish resources and fisheries management (*e.g.*, growth of fish in rice paddies in Asia) are important aspects of this industry. (E.D.H.)

BOTANY

Botany is the study of plants. Plants were of paramount importance to early man; he depended upon them as sources of food, shelter, clothing, medicine, ornament, tools, and magic. Today it is known that, in addition to their practical and economic values, green plants are indispensable to all life on Earth: through the process of photosynthesis, plants transform energy from the sun into the chemical energy of food, which makes all life possible. A second unique and important capacity of green plants is the formation and release of oxygen as a by-product of photosynthesis. The oxygen of the atmosphere, so absolutely essential to many forms of life, represents the accumulation of over 3,500,000,000 years of photosynthesis by green plants.

Although the many steps in the process of photosynthesis have become fully understood only in recent years, even in prehistoric times man somehow recognized intuitively that some important relation existed between the sun and plants. Such recognition is suggested by the fact that, in primitive tribes and early civilizations, worship of the sun was often combined with the worship of plants.

Earliest man, like the other anthropoid mammals (*e.g.*, apes, monkeys), depended totally upon the natural resources of his environment, which, until he developed methods for hunting, consisted almost completely of plants. The behaviour of pre-Stone Age man can be inferred by studying the botany of aboriginal peoples in various parts of the world. Isolated tribal groups in South America, Africa, and New Guinea, for example, have extensive knowledge about plants and distinguish hundreds of kinds according to their utility, as edible, poisonous, or otherwise important in their culture. They have developed surprisingly sophisticated systems of nomenclature and classification, which approximate the binomial system (*i.e.*, generic and specific names) found in modern biology. The urge to recognize different kinds of plants and to give them names thus seems to be as old as the human race.

In time plants were not only collected by primitive man but also grown by him. This domestication resulted not only in the development of agriculture but also in a greater stability of human populations that had previously been nomadic. From the settling down of agricultural peoples in places where they could depend upon adequate food supplies came the first villages and the earliest civilizations.

Because of the long preoccupation of man with plants, a large body of folklore, general information, and actual scientific data has accumulated, which has become the basis for the science of botany.

The basis of botany

Historical background. Theophrastus, a Greek philosopher who studied first with Plato and then became a disciple of Aristotle, is credited with founding botany. Only two of an estimated 200 botanical treatises written by him are known to science: originally written in Greek about 300 BC, they have survived in the form of Latin manuscripts, *De causis plantarum* and *De historia plantarum.* His basic concepts of morphology, classification, and the natural history of plants, accepted without question for many centuries, are now of interest primarily because of Theophrastus' independent and philosophical viewpoint.

Pedanius Dioscorides, a Greek botanist of the 1st century AD, was the most important botanical writer after Theophrastus. In his major work, an herbal in Greek, he described some 600 kinds of plants, with comments on their habit of growth and form as well as on their medicinal properties. Unlike Theophrastus, who classified plants as trees, shrubs, and herbs, Dioscorides grouped his plants under three headings: as aromatic, culinary, and medicinal. His herbal, unique in that it was the first treatment of medicinal plants to be illustrated, remained for about 15 centuries the last word on medical botany in Europe.

From the 2nd century BC to the 1st century AD, a succession of Roman writers—Cato, Varro, Virgil, and Columella—prepared Latin manuscripts on farming, gardening, and fruit growing but showed little evidence of the spirit of scientific inquiry for its own sake that was so characteristic of Theophrastus. In the 1st century AD, Pliny the Elder, though no more original than his Roman predecessors, seemed more industrious as a compiler. His *Historia naturalis*—an encyclopaedia of 37 volumes, compiled from some 2,000 works representing 146 Roman and 327 Greek authors—has 16 volumes devoted to plants. Although uncritical and containing much misinformation, this work contains much information otherwise unavailable, since most of the volumes to which he referred have been destroyed.

The printing press revolutionized the availability of all types of literature, including that of plants. In the 15th and 16th centuries, many herbals were published with the purpose of describing plants useful in medicine. Written by physicians and medically oriented botanists, the earliest herbals were based largely on the work of Dioscorides and to a lesser extent on Theophrastus, but gradually they became the product of original observation. The increasing objectivity and originality of herbals through the decades is clearly reflected in the improved quality of the woodcuts prepared to illustrate these books.

The significance of herbals

In 1552 an illustrated manuscript on Mexican plants, written in Aztec, was translated into Latin by Badianus; other similar manuscripts known to have existed seem to have disappeared. Whereas herbals in China date back much further than those in Europe, they have become known only recently and so have contributed little to the progress of Western botany.

The invention of the optical lens during the 16th century and the development of the compound microscope about 1590 opened an era of rich discovery about plants; prior to that time, all observations by necessity had been made

with the unaided eye. The botanists of the 17th century turned away from the earlier emphasis on medical botany and began to describe all plants, including the many new ones that were being introduced in large numbers from Asia, Africa, and America. Among the most prominent botanists of this era was Gaspard Bauhin, who for the first time developed, in a tentative way, many botanical concepts still held as valid. In 1665 Robert Hooke published, under the title *Micrographia,* the results of his microscopic observations on several plant tissues. He is remembered as the coiner of the word cell, referring to the cavities he observed in thin slices of cork; his observation that living cells contain sap and other materials too often has been forgotten. In the following decade, Nehemiah Grew and Marcello Malpighi founded plant anatomy; in 1671 they communicated the results of microscopic studies simultaneously to the Royal Society of London, and both later published major treatises.

The founding of plant anatomy

Experimental plant physiology began with the brilliant work of Stephen Hales, who published his observations on the movements of water in plants under the title *Vegetable Staticks* (1727). His conclusions on the mechanics of water transpiration in plants are still valid, as is his discovery—at the time a startling one—that air contributes something to the materials produced by plants. In 1774, Joseph Priestley showed that plants exposed to sunlight give off oxygen, and Jan Ingenhousz demonstrated, in 1779, that plants in the dark give off carbon dioxide. In 1804 Nicolas de Saussure demonstrated convincingly that plants in sunlight absorb water and carbon dioxide and increase in weight, as had been reported by Hales nearly a century earlier.

The widespread use of the microscope by plant morphologists provided a turning point in the 18th century—botany became largely a laboratory science. Until the invention of simple lenses and the compound microscope, the recognition and classification of plants were, for the most part, based on such large morphological aspects of the plant as size, shape, and external structure of leaves, roots, and stems. Such information was also supplemented by observations on more subjective qualities of plants, such as edibility and medicinal uses.

In 1753 Linnaeus published his master work, *Species Plantarum,* which contains careful descriptions of 6,000 species of plants from all of the parts of the world known at the time. In this work, which is still the basic reference work for modern plant taxonomy, Linnaeus established the practice of binomial nomenclature—that is, the denomination of each kind of plant by two words, the genus name and the specific name, as *Rosa canina,* the dog rose. Binomial nomenclature had been introduced much earlier by some of the herbalists, but it was not generally accepted; most botanists continued to use cumbersome formal descriptions, consisting of many words, to name a plant. Linnaeus for the first time put the contemporary knowledge of plants into an orderly system, with full acknowledgment to past authors, and produced a nomenclatural methodology so useful that it has not been greatly improved upon. Linnaeus also introduced a "sexual system" of plants, by which the numbers of flower parts—especially stamens, which produce male sex cells, and styles, which are prolongations of plant ovaries that receive pollen grains—became useful tools for easy identification of plants. This simple system, though effective, had many imperfections. Other classification systems, in which as many characters as possible were considered in order to determine the degree of relationship, were developed by other botanists; indeed, some appeared before the time of Linnaeus. The application of the concepts of Charles Darwin (on evolution) and Gregor Mendel (on genetics) to plant taxonomy has provided insights into the process of evolution and the production of new species.

Binomial nomenclature

Systematic botany now uses information and techniques from all the subdisciplines of botany, incorporating them into one body of knowledge. Phytogeography (the biogeography of plants), plant ecology, population genetics, and various techniques applicable to cells—cytotaxonomy and cytogenetics—have contributed greatly to the current status of systematic botany and have to some degree become part of it. More recently, phytochemistry, computerized statistics, and fine-structure morphology have been added to the activities of systematic botany.

The 20th century has seen an enormous increase in the rate of growth of research in botany and the results derived therefrom. The combination of more botanists, better facilities, and new technologies, all with the benefit of experience from the past, has resulted in a series of new discoveries, new concepts, and new fields of botanical endeavour. Some important examples are mentioned below.

New and more precise information is being accumulated concerning the process of photosynthesis, especially with reference to energy-transfer mechanisms.

The discovery of the pigment phytochrome, which constitutes a previously unknown light-detecting system in plants, has greatly increased knowledge of the influence of both internal and external environment on the germination of seeds and the time of flowering.

The discovery of phytochrome

Several types of plant hormones (internal regulatory substances) have been discovered—among them auxin, gibberellin, and kinetin—whose interactions provide a new concept of the way in which the plant functions as a unit.

The discovery that plants need certain trace elements usually found in the soil has made it possible to cultivate areas lacking some essential element by adding it to the deficient soil.

The development of genetical methods for the control of plant heredity has made possible the generation of improved and enormously productive crop plants.

The development of radioactive-carbon dating of plant materials as old as 50,000 years is useful to the paleobotanist, the ecologist, the archaeologist, and especially to the climatologist, who now has a better basis on which to predict climates of future centuries.

The discovery of alga-like and bacteria-like fossils in Precambrian rocks has pushed the estimated origin of plants on Earth to 3,500,000,000 years ago.

The isolation of antibiotic substances from fungi and bacteria-like organisms has provided control over many bacterial diseases and has contributed biochemical information of basic scientific importance as well.

Areas of study. For convenience, but not on any mutually exclusive basis, several major areas or approaches are recognized commonly as disciplines of botany; these are morphology, physiology, ecology, and systematics.

Morphology. Morphology deals with the structure and form of plants and includes such subdivisions as: cytology, the study of the cell; histology, the study of tissues; anatomy, the study of the organization of tissues into the organs of the plant; reproductive morphology, the study of life cycles; and experimental morphology, or morphogenesis, the study of development.

Physiology. Physiology deals with the functions of plants. Its development as a subdiscipline has been closely interwoven with the development of other aspects of botany, especially morphology. In fact, structure and function are sometimes so closely related that it is impossible to consider one independently of the other. The study of function is indispensable for the interpretation of the incredibly diverse nature of plant structures. In other words, around the functions of the plant, structure and form have evolved. Physiology also blends imperceptibly into the fields of biochemistry and biophysics, as the research methods of these fields are used to solve problems in plant physiology.

The close relationship of plant form and function

Ecology. Ecology deals with the mutual relationships and interactions between organisms and their physical environment. The physical factors of the atmosphere, the climate, and the soil affect the physiological functions of the plant in all its manifestations, so that, to a large degree, plant ecology is a phase of plant physiology under natural and uncontrolled conditions; in fact, it has been called "outdoor physiology." Plants are intensely sensitive to the forces of the environment, and both their association into communities and their geographical distribution are determined largely by the character of climate and soil. Moreover, the pressures of the environment and of organisms upon each other are potent forces, which lead to new species and the continuing evolution of larger groups.

Systematics. Systematics deals with the identification and ranking of all plants; it includes classification and nomenclature (naming) and enables the botanist to comprehend the broad range of plant diversity and evolution.

Other subdisciplines. In addition to the major subdisciplines, several specialized branches of botany have developed as a matter of custom or convenience. Among them are bacteriology, the study of bacteria; mycology, the study of fungi; algology or phycology, the study of algae; bryology, the study of mosses and liverworts; pteridology, the study of ferns and their relatives; and paleobotany, the study of fossil plants. Palynology is the study of modern and fossil pollen and spores, with particular reference to their identification; plant pathology deals with the diseases of plants; economic botany deals with plants of practical use to man; and ethnobotany covers the use of plants by aboriginal peoples, now and in the distant past.

Botany also relates to other scientific disciplines in many ways, especially to zoology, medicine, microbiology, agriculture, chemistry, forestry, and horticulture, and specialized areas of botanical information may relate closely to such humanistic fields as art, literature, history, religion, archaeology, sociology, and psychology.

Fundamentally, botany remains a pure science, including any research into the life of plants and limited only by man's technical means of satisfying his curiosity. It has often been considered an important part of a liberal education, not only because it is necessary for an understanding of agriculture, horticulture, forestry, pharmacology, and other applied arts and sciences, but also because an understanding of plant life is related to life in general.

The aesthetic influence of plants

Because man has always been dependent upon plants and surrounded by them, he has woven them into his designs, into the ornamentation of his life, even into his religious symbolism. A Persian carpet and a bedspread from a New England loom both employ conventional designs derived from the forms of flowers. Medieval painters and great masters of the Renaissance represented various revered figures surrounded by roses, lilies, violets, and other flowers, which symbolized chastity, martyrdom, humility, and other Christian attributes.

Methods in botany. *Morphological aspects.* The invention of the compound microscope provided a valuable and durable instrument for the investigation of the inner structure of plants. Early plant morphologists, especially those studying cell structure, were handicapped as much by the lack of adequate knowledge of how to prepare specimens as they were by the imperfect microscopes of the time. A revolution in the effectiveness of microscopy occurred in the second half of the 19th century with the introduction of techniques for fixing cells and for staining their component parts. Before the development of these techniques, the cell, viewed with the microscope, appeared as a minute container with a dense portion called the nucleus. The discovery that parts of the cell respond to certain stains made observation easier. The development of techniques for preparing tissues of plants for microscopic examination was continued in the 1870s and 1880s and resulted in the gradual refinement of the field of nuclear cytology, or karyology. Chromosomes were recognized as constant structures in the life cycle of cells, and the nature and meaning of meiosis, a type of cell division in which the daughter cells have half the number of chromosomes of the parent, was discovered; without this discovery, the significance of Mendel's laws of heredity might have gone unrecognized. Vital stains, dyes that can be used on living material, were first used in 1886 and have been greatly refined since then.

Improvement of the methodology of morphology has not been particularly rapid, even though satisfactory techniques for histology, anatomy, and cytology have been developed. The embedding of material in paraffin wax, the development of the rotary microtome for slicing very thin sections of tissue for microscope viewing, and the development of stain techniques are refinements of previously known methods. The invention of the phase microscope made possible the study of unfixed and unstained living material—hopefully nearer its natural state. The development of the electron microscope, however, has provided the plant morphologist with a new dimension of magnification of the structure of plant cells and tissues. The fine structure of the cell and of its components, such as mitochondria and the Golgi apparatus, have come under intensive study. Knowledge of the fine structure of plant cells has enabled investigators to determine the sites of important biochemical activities, especially those involved in the transfer of energy during photosynthesis and respiration. The scanning electron microscope, a relatively recent development, provides a three-dimensional image of surface structures at very great magnifications.

Fine structure a clue to biochemical activities

For experimental research on the morphogenesis of plants, isolated organs in their embryonic stage, clumps of cells, or even individual cells are grown. One of the most interesting techniques developed thus far permits the growing of plant tissue of higher plants as single cells; aeration and continuous agitation keep the cells suspended in the liquid culture medium.

Physiological aspects. Plant physiology and plant biochemistry are the most technical areas of botany; most major advances in physiology also reflect the development of either a new technique or the dramatic refinement of an earlier one to give a new degree of precision. Fortunately, the methodology of measurement has been vastly improved in recent decades, largely through the development of various electronic devices. The phytotron at the California Institute of Technology represents the first serious attempt to control the environment of living plants on a relatively large scale; much important information has been gained concerning the effects on plants of day length and night length and the effects on growth, flowering, and fruiting of varying night temperatures. Critical measurements of other plant functions have also been obtained.

Certain complex biochemical processes, such as photosynthesis and respiration, have been studied stepwise by immobilizing the process through the use of extreme cold or biochemical inhibitors and by analyzing the enzymatic activity of specific cell contents after spinning cells at very high speeds in a centrifuge. The pathways of energy transfer from molecule to molecule during photosynthesis and respiration have been determined by biophysical methods, especially those utilizing radioactive isotopes.

An investigation of the natural metabolic products of plants requires, in general, certain standard biochemical techniques—*e.g.*, gas and paper chromatography, electrophoresis, and various kinds of spectroscopy, including infrared, ultraviolet, and nuclear magnetic resonance. Useful information on the structure of the extremely large cellulose molecule has been provided by X-ray crystallography.

Ecological aspects. When plant ecology first emerged as a subscience of botany, it was largely descriptive; today, however, it has become a common meeting ground for all the plant sciences, as well as for other sciences. In addition, it has become much more quantitative. As a result, the tools and methods of plant ecologists are those available for measuring the intensity of the environmental factors that impinge on the plant and the reaction of the plant to these factors. The extent of the variability of many physical factors must be measured. The integration and reporting of such measurements, which cannot be regarded as constant, may therefore conceal some of the most dynamic and significant aspects of the environment and the responses of the plant to them. Because the physical environment is a complex of biological and physical components, it is measured by biophysical tools. The development of electronic measuring and recording devices has been crucial for a better understanding of the dynamics of the environment. Such devices, however, produce so much information that computer techniques must be used to reduce the data to meaningful results.

The ecologist, concerned primarily with measuring the effect of the external environment on a plant, adapts the methodology of the plant physiologist to field conditions.

The plant sociologist

The plant sociologist, on the other hand, is concerned with both the relation of different kinds of plants to each other and the nature and constitution of their association into natural communities. One widely used technique in this respect is to count the various kinds of plants within a standard area in order to determine such factors as the

percentage of ground cover, dominance of species, aggressiveness, and other characteristics of the community. In general, the plant sociologist has relatively few quantitative factors to measure and must therefore take a subjective and intuitive approach, which, nevertheless, gives extremely useful results and some degree of predictability.

Some ecologists are most concerned with the inner environment of the plant and the way in which it reacts to the external environment. This approach, which is essentially physiological and biochemical, is useful for determining energy flow in ecosystems. The physiological ecologist is also concerned with evaluating the adaptations that certain plants have made toward survival in a hostile environment.

In summary, the techniques and methodology of plant ecology are as diverse and as varied as the large number of sciences that are drawn upon by ecologists. Completely new techniques, although few, are important; among them are techniques for measuring the amount of radioactive carbon-14 in plant deposits up to 50,000 years old. The most important new method in plant ecology is the rapidly growing use of computer techniques for handling vast amounts of data. Furthermore, modern digital computers can be used to simulate simple ecosystems and to analyze real ones.

Taxonomic aspects. Experimental research under controlled conditions, made possible by botanical gardens and their ranges of greenhouses and controlled environmental chambers, has become an integral part of the methodology of modern plant taxonomy.

A second major tool of the taxonomist is the herbarium, a reference collection consisting of carefully selected and dried plants attached to paper sheets of a standard size and filed in a systematic way so that they may be easily retrieved for examination. Each specimen is a reference point representing the features of one plant of a certain species; it lasts indefinitely if properly cared for, and, if the species becomes extinct in nature—as hundreds have—it remains the only record of the plant's former existence. The library is also an essential reference resource for descriptions and illustrations of plants that may not be represented in a particular herbarium.

One of the earliest methods of the taxonomist, the study of living plants in the field, has benefitted greatly by fast and easy methods of transportation; botanists may carry on fieldwork in any part of the world and make detailed studies of the exact environmental conditions under which each species grows.

New approaches in systematic botany

During the present century, many new approaches have been applied to the elucidation of problems in systematic botany. The transmission electron microscope and the scanning electron microscope have added to the knowledge of plant morphology, upon which classical taxonomy so much depends.

Refined methods for cytological and genetical studies of plants have given the taxonomist new insights into the origin of the great diversity among plants, especially the mechanisms by which new species arise and by which they then maintain their individuality in nature. From such studies have arisen further methods and also the subdisciplines of cytotaxonomy, cytogenetics, and population genetics.

Phytochemistry, or the chemistry of plants, one of the early subdivisions of organic chemistry, has been of great importance in the identification of plant substances of medicinal importance. With the development of new phytochemical methods, new information has become available for use in conjunction with plant taxonomy; thus has arisen the modern field of chemotaxonomy, or biochemical systematics. Each species tends to differ to some degree from every other species, even in the same genus, in the biochemistry of its natural metabolic products. Sometimes the difference is subtle and difficult to determine; sometimes it is obvious and easily perceptible. With new analytical techniques, a large number of individual compounds from one plant can be identified quickly and with certainty. Such information is extremely useful in adding confirmatory or supplemental evidence of an objective and quantitative nature. An interesting by-product of chemical plant taxonomy has resulted in understanding better the restriction of certain insects to specific plants.

Computer techniques have recently been applied to plant taxonomy to develop a new field, numerical taxonomy, or taximetrics, by which relationships between plant species or those within groups of species are determined quantitatively and depicted graphically. Another method measures the degree of molecular similarity of deoxyribonucleic acid (DNA) molecules in different plants. By this procedure it should be possible to determine the natural taxonomic relationships among different plants and plant groups by determining the extent of the relationship of their DNA's: closely related plants will have more similarities in their DNA's than will unrelated ones. (W.C.St.)

The use of nucleic acids in taxonomy

MICROBIOLOGY

The 17th-century discovery that living forms exist that are invisible to the naked eye was a dramatic one in man's history, for, from the 13th century onward, it had been postulated that "invisible" organisms were responsible for decay and disease. The word microbe was coined in the latter quarter of the 19th century to describe these organisms, all of which were thought to be related. As microbiology eventually developed into a separate science, it was found that microbes comprise a very large group of extremely diverse organisms; thus, microbiology became subdivided into various disciplines—*e.g.*, bacteriology, protozoology, and virology. The diversity of microbes, or microorganisms as they are now commonly called, has meant that it is almost impossible for one person to be knowledgeable in all of the disciplines grouped under microbiology.

Microbiology involves the identification of microorganisms and the study of their structure and function. It encompasses the study of bacteria, rickettsiae, small fungi (*e.g.*, yeasts and molds), algae, and protozoans, as well as problematical forms of life such as viruses. Because of the difficulty of assigning plant or animal status to microorganisms—some are plantlike, others animal-like—they are sometimes considered a separate group called protists. Microbes can also be divided into procaryotes, which have a primitive and dispersed kind of nuclear material—*i.e.*, the blue-green algae, bacteria, and rickettsiae—and eucaryotes, which display a distinct nucleus bounded by a membrane. Such are small algae other than the blue-greens, yeasts and molds, and protozoans. (All higher organisms are eucaryotes.)

The biological status of microorganisms

Man's daily life is interwoven inextricably with microorganisms. They abound in the soil, in the seas, and in the air. Everywhere abundant, although usually unnoticed, microorganisms provide ample evidence of their presence, sometimes unfavourably, as when they cause decay of objects valued by man or generate disease, and sometimes favourably, as when they ferment alcohol to wine and beer, raise bread, flavour cheeses, and create other dairy products from milk. Microorganisms are of incalculable value in nature, causing the disintegration of animal and plant remains and converting them to gases and minerals that can be recycled in other organisms.

Historical background. Microbiology can be said to have begun with the development of the microscope. Although others may have seen microbes before him, Antonie van Leeuwenhoek, a Dutch draper whose hobby was lens grinding and microscope making, was the first to provide proper descriptions of his observations, which included protozoans from the guts of animals and bacteria from teeth scrapings. His descriptions and drawings were excellent because his lenses were of an exceptional quality. Leeuwenhoek conveyed his findings in a series of letters to the British Royal Society during the mid-1670s. Although his observations stimulated much interest, no one made a serious attempt either to repeat or to extend them. Leeuwenhoek's "animalcules," as he called them, thus remained mere oddities of nature to the scientists of his day, and enthusiasm for the study of microbes gained ground slowly. It was only later, during the 18th-century revival of a long-standing controversy about whether or not life can develop out of nonliving material, that the significance of microorganisms in the scheme of nature and in the health and welfare of man became evident.

The early Greeks believed that living things could originate from nonliving matter; the goddess Gea was credited with creating life from stones. Although Aristotle discarded this notion, he still held that animals could arise spontaneously from other unlike organisms or from soil. His influence regarding this concept of spontaneous generation was still felt as late as the 17th century. Toward the end of the 17th century, a chain of observations, experiments, and arguments began that dealt a deathblow to the idea that life could be generated from nonlife. It was an involved series of events, with the forces of personality and strong wills often obscuring the facts.

Although Francesco Redi, an Italian naturalist, disproved that higher forms of life could originate spontaneously, proponents of the concept claimed that microbes were different and did indeed arise in this way. Such illustrious names as John Needham, Lazzaro Spallanzani, Franz Schultze, and Theodor Schwann figured in the debates.

It remained for Louis Pasteur to settle the matter. He proved in a series of masterful experiments that only preexisting microbes could give rise to other microbes—at least under current earthly conditions (that life arose spontaneously from nonlife at some earlier time, under appropriate physical and chemical conditions, is an undisputed postulate of chemical evolution).

Microbes and disease

Regarding microbes and disease, Girolamo Fracastoro, an Italian scholar, advanced the notion as early as the mid-1500s that contagion is an infection that passes from one thing to another. The "thing" that is passed along eluded discovery until the late 1800s, when the work of many scientists, Pasteur foremost among them, determined the role of bacteria in fermentation and disease. Robert Koch, a German physician, defined the procedure for proving that a specific organism causes a specific disease.

The foundation of microbiology was securely laid during the period from about 1880 to 1900. The students of Pasteur, Koch, and others discovered in rapid succession a host of bacteria capable of causing specific diseases (pathogens) and elaborated an extensive armamentarium of techniques and laboratory procedures for revealing the ubiquity, diversity, and power of microbes.

All of these developments occurred in Europe. Not until the early 1900s did microbiology become established in America. Many of the microbiologists who worked in America at this time either had studied under Koch or at the Pasteur Institute, in Paris. All microbiologists of the early 20th century, however, were influenced by such men as Koch. Once established in America, microbiology flourished, especially with regard to such related disciplines as biochemistry and genetics.

Since the 1940s, microbiology has experienced an extremely productive period, during which many disease-causing microbes have been identified and methods to control them have been developed. Microorganisms have also been effectively utilized in industry; their activities have been channelled so that valuable products of commerce and agricultural benefits result.

The study of microorganisms has also advanced man's knowledge of living things. Microbes provide easy-to-work-with material for studying the complex processes of life; *e.g.*, metabolism. Correlated with the intensive probing into the functions of microbes have been numerous, and often unexpected, dividends that can be applied to solving existing problems. Knowledge of the basic metabolism of a pathogenic bacterium, for example, often leads to a means for controlling the pathogen. Nutritional requirements of bacteria thus may be of value in combatting an infection.

Areas of study. The study of bacteria, which were among the first objects of microbiological study, is called bacteriology. Various subdisciplines deal exclusively with particular microorganisms.

From another standpoint microbiology can be subdivided into theoretical, or pure, microbiology, and practical, or applied, microbiology. The latter can be further subdivided according to specialties, such as medical, industrial, agricultural, food, and dairy microbiology (see below *Applied microbiology*).

Interdisciplinary work. The science of microbiology has been influenced by, and in turn has influenced, other sciences. Microorganisms are no longer the exclusive concern of microbiologists. Biochemists, geneticists, cytologists, and molecular biologists have discovered the value of microbes as experimental tools in the study of such fundamental biological processes as metabolism, photosynthesis, enzyme action, gene action, and population dynamics. Microorganisms are well suited to such uses; they represent a vast range of metabolic types, and genetic changes are correlated with their rapid proliferation. In addition, they can subsist on relatively simple inorganic nutrients and, because they multiply rapidly, are available in extremely large numbers in a relatively short period of time. They are also relatively easy to maintain and handle under laboratory conditions.

Bacteriology. Modern and accurate knowledge of the forms of bacteria dates from the researches of the German botanist F.J. Cohn, the chief results of which were published at various periods between 1853 and 1892. Cohn's classification of bacteria, published in 1872 and extended in 1875, dominated the study of these organisms thereafter. While various observers added to the knowledge of the structure of bacteria, others laid the foundation of what is known about the relations of bacteria to fermentation and disease. When Pasteur showed in 1857 that lactic acid fermentation depends upon the presence of an organism, it was already known that fermentation and putrefaction are intimately connected with the presence of organisms in the air. In 1862 Pasteur placed beyond reasonable doubt the fact that production of ammonia by the fermentation of urea is caused by the action of a minute bacterium, named in 1874 *Micrococcus ureae.*

Isolation of disease-causing bacteria

After the introduction of bacteriological techniques (see below *The study of microbiology: Cultivation of microorganisms*) came the isolation of many bacteria. It was discovered in 1882, for example, that a bacillus (a rod-shaped bacterium) is the cause of glanders, a disease of horses. In 1883 Koch isolated the organism of Asiatic cholera, and the same year that of diphtheria was found. In 1885 the tetanus bacillus was observed in pus produced in mice and rabbits inoculated with soil; only in 1889, however, did the Japanese bacteriologist S. Kitasato discover the way in which the organisms could be cultivated (they grow only in the absence of oxygen). W.D. Miller, a U.S. dentist, studied the microorganisms of the human mouth in the 1880s, noting their possible relationship to the decay of teeth.

Pasteur and his associates found that animals vaccinated with a specially cultivated anthrax bacillus showed immunity to disease when reinoculated with the deadly wild form. These findings were destined to lead to a study of the principles of immunity, which underlie the prevention and treatment of disease by vaccines and immune serum. Questions relating to causes and to the nature of changes occurring both in the bacteria and in the host, as well as the development of immunity in the latter, continue to be subjects of great interest and importance.

While investigations on infectious diseases and immunity were under way, it became apparent that other activities of bacteria also are of importance to man. In 1878, only two years after Koch announced the discovery of the anthrax bacillus, the U.S. botanist T.J. Burrill discovered the bacterial cause of fire blight in pears, thereby establishing that certain plant diseases are caused by bacteria. The importance of some bacteria in soil and their contribution to soil fertility was recognized in the 1880s and 1890s. The significance of bacterial activities in many aspects of the dairy industry was recognized at about the same time.

The further application of bacterial activities to industrial processes, other than early studies on alcoholic and lactic acid fermentations, began later. Thus, within the span of a few decades the study of bacteria had progressed enormously, and by the early 20th century it was recognized that the activities of bacteria bear an intimate relation to many aspects of human activity.

The original interest of bacteriologists centred upon what bacteria do. This interest has broadened to include the study of the bacteria themselves: what they are, their relationships to each other, and their relationships to other organisms.

Protozoology and others. The name Protozoa is derived from the Greek words meaning "first animal," and, indeed, protozoans are sometimes considered the most simple of all animals. First observed by Leeuwenhoek in the 17th century, protozoans are almost as ubiquitous as bacteria. The study of protozoans at one time centred on the parasites that cause malaria and sleeping sickness, thereby stimulating research in tropical medicine. Improved laboratory techniques for cultivating this diverse group of organisms have made them valuable tools in many types of investigations—from physiological studies, such as cytoplasmic motion in *Amoeba,* to ecological studies involving their role in the food chain of the oceans.

The term fungi, Latin for "mushroom," applies not only to mushrooms but to all of the large and diverse group of plantlike organisms to which the mushrooms belong. The fungi of most interest in microbiology are the yeasts and molds. Many molds are studied because of their economic importance. *Mucor,* for example, not only causes spoilage of foods such as vegetables and fruits but also is used to manufacture some cheeses.

Lichens and slime molds

Some molds form unique relationships with other organisms; lichens, for instance, are composed of both algae and fungi. The unusual nature of lichens was not discovered until after the invention of the microscope. The name lichen appeared long before, however, in the writings of Theophrastus, a disciple of Aristotle, about the 4th century BC. At that time, and for many centuries thereafter, lichens were confused with mosses and the hepatics. Early students thought that the green cells, now known to be an alga, were specialized fungal structures. Not until 1867 were the green cells in lichens indentified as algae; at the same time it was announced that the fungal cells were parasitic upon them. The introduction of pure culture techniques permitted the lichen constituents to be isolated and grown apart from each other. In 1873 the German botanist H.A. De Bary suggested the name symbiosis to describe the mutual benefit between the components of lichens.

Once objects of interest for their curious form, slime molds—as their individual nutritional, environmental, and genetic characteristics have become better understood—have become objects useful in teaching and research. The creeping sheet (plasmodium), or vegetative stage of a slime mold, has the characteristics of a primitive animal and resembles a primitive amoeba. The reproductive stage (sporangium) has the characteristics of a mold. The double life of these organisms is reflected in the name Mycetozoa, meaning fungus animals, which was coined by De Bary in 1858.

Algae, a heterogeneous group of primitive organisms, are of great interest to all biologists because they are capable of photosynthesis and are of evolutionary interest as well. Intensive research is currently directed toward the small forms making up much of the free-floating microscopic life in water (phytoplankton) because they provide food for aquatic animals and thus are of great importance. Many freshwater algae produce undesirable tastes and odours, and studies have been aimed at eliminating them from domestic water supplies. Although algae occasionally form mats on water surfaces, sometimes causing suffocation of aquatic life, heavy growth of certain species also has been found to reduce water hardness and to remove the salts found in brackish water, thereby making the water more suitable for human consumption. Algae are also being studied for their potential value as a food source for man.

Virology and others. Viruses comprise a heterogeneous assemblage of self-reproducing agents, smaller than the microscopically visible bacteria, that multiply only within living susceptible cells. Certain diseases caused by viruses have been known since the late 1700s, when the British physician Edward Jenner developed a vaccine from material isolated from cowpox lesions. He did not see the causative agent. Pasteur, in the mid-1850s, developed an attenuated strain (*i.e.,* one made less virulent) of the virus that causes rabies, but he was not then aware of the viral nature of the disease. An associate of Pasteur, Charles Chamberland, discovered that bacteria would not pass through a porcelain filter but that the causative agent of rabies did. The term filter-passing, or filterable, virus was used to describe these agents.

The Russian bacteriologist D. Ivanovski in 1892 found that a filtrate of sap from tobacco plants infected with mosaic disease could be used to transfer infection to healthy plants; the Dutch microbiologist M.W. Beirjerinck later confirmed the work. The further finding in 1902 that the agent of foot-and-mouth disease of cattle is also filterable made clear the fact that the agents cause disease in animals as well as plants. By 1930 the term filterable had been dropped, and virus is routinely used by microbiologists as a name for these agents.

The study of viruses

Nobel Prize winner Max Theiler found in 1951 a method for attenuating virulent yellow fever virus; the technique has since been modified to produce vaccines against other viral diseases. Thus, viruses have been studied intensively in recent years not only because they provide important information about life processes but also because vaccines against them are constantly sought.

The study of rickettsiae, which resemble very small bacteria but grow only in susceptible cells, is intimately associated with those forms causing human disease. Elucidation of the microbes causing such diseases as epidemic typhus, Rocky Mountain spotted fever, and scrub typhus began in 1906 with the work of Howard Taylor Ricketts, for whom the organisms are named; he described the organism of spotted fever. In 1910 Ricketts and an associate described the organism of typhus fever, which in 1916, was named *Rickettsia prowazekii* in honour of Ricketts and Stanislas von Prowazek, both of whom died of the disease. Research continues as better methods for prevention, control, and treatment of rickettsial diseases are sought.

General trends. One of the more current studies involving microorganisms is their possible occurrence in outer space and on planets other than Earth. A branch of exobiology, space microbiology, includes the investigation of microbes as providers of food and oxygen in the closed environment of spaceships.

A less positive development of microbiology has been biological warfare—the selection and cultivation of microbes as weapons of war, to cause disease or injury to domestic plants, animals, and man.

Although most of the subdisciplines of microbiology are directed at the occurrence of microbes in specific environments, the study of gnotobiotics is concerned with the exclusion of microbes except those involved in any given experiment. Such germfree organisms are important research tools in investigating parasitic diseases, immunological processes, nutrition, stress, shock, and aging.

Methods in microbiology. The study of microbiology is channelled in two major directions. Pure microbiology—that is, the study of a particular group of microorganisms in order to learn about their morphology, physiology, taxonomy, occurrence, variation, heredity, and evolution—seeks to understand the nature of microbes. Applied microbiology, on the other hand, is motivated by the desire to exploit the effects of microorganisms that are of benefit to man and to control the activities of those that are harmful.

Bacteriology as a synonym for microbiology

The methods used to study bacteria have been widely adapted to the study of other microorganisms; indeed, the techniques employed in the microbiological sciences probably have more in common than do the origin and evolution of the organisms themselves. For this reason, the term microbiology is often considered to be synonymous with bacteriology. Similarly, the term microorganism no longer refers only to microscopic organisms. With the adoption of bacteriological methods for the study of many types of organisms, the meaning of the term microorganism has gradually been extended to include any organism that can grow and be studied using the methods originally developed for bacteria. The study of microbiology, therefore, encompasses many organisms that are not microscopic (*e.g.,* molds) but usually does not include some that are (*e.g.,* rotifers).

Microscopy. That invisibly small organisms exist was believed from early times. Lucretius, who expressed an atomic view of matter, wrote about AD 75 that even the plague must be caused by a kind of atom. He considered

the atoms, or seeds, as lifeless, however. The writings of Lucretius influenced 16th-century scientists, but the means for actually observing microorganisms—a magnifying lens—was not developed until the 17th century by Galileo.

The convex lens of Leeuwenhoek allowed a greatly enlarged image to be seen, enabling him to see protozoans, filamentous fungi, and yeasts. In 1676, using his simple lens, which was capable of magnifying 280 times, he first saw bacteria. A few years before Leeuwenhoek's observation of bacteria, Robert Hooke, the father of the cell theory, had observed filamentous fungi (1667) through a compound microscope—one with two lenses, which gives a larger image than that given by a simple lens.

By 1786 various microorganisms, including bacteria, had been viewed through the compound microscope. The compound microscope of today is widely used; it differs from those of the 18th century mainly in the addition of a device, called a condenser, for lighting the object being viewed with a wide cone of light. A condenser is necessary to provide sufficient light when large magnifications are used.

Types of light microscopy used in microbiology

Of the types of microscopy now used in microbiology, most utilize light microscopes; magnification is obtained by a system of optical lenses. The types of light microscopes include the commonly used bright field, in which the background is brightly lighted, the objects studied are dark, and the power of magnification is about 1,000; the dark field, in which the background is dark and the objects studied are bright, a phenomenon particularly useful for examining unstained organisms suspended in fluid; the ultraviolet, the greatest advantage of which is that magnifications two to three times greater than those obtained with the light microscope can be achieved because ultraviolet light has a shorter wavelength than visible light; and phase contrast, in which controlled illumination is obtained by the use of special equipment that enables the refraction, or bending, of light passing from one material to another to be seen, thereby revealing differences in cells not discernible with other microscopic methods. Fluorescence microscopy utilizes the ultraviolet microscope; a chemical substance with the property to absorb ultraviolet waves and emit visible ones is used with a mixture of microorganisms, some of which take up the substance, and thus can be distinguished from those that do not. Still another type of microscopy, electron microscopy, allows very great magnification; waves of electrons (negatively charged particles), rather than light, and magnetic fields, rather than lenses, are used to achieve an image. Although electron microscopy has the advantage of great magnification, it also has several limitations, most important of which is that the material must be dry. This not only eliminates the possibility of studying living specimens but also raises the possibility that the drying process may alter the characteristics of the specimen.

Cultivation of microorganisms. Microorganisms growing on a nutrient medium are referred to as a culture. Early experiments (1776) by Spallanzani, begun as the result of the previously mentioned controversy over whether or not life could develop out of lifeless matter, laid the foundation for the technique of sterile culture. This involves first freeing a suitable medium of microorganisms by heating and, second, keeping the medium sterile—*i.e.,* keeping microorganisms out of the medium. In his experiments, Spallanzani boiled various kinds of seeds in a flask and stoppered it. After a few days, many organisms could be found in the flask; Spallanzani distinguished the larger ones, which were destroyed by boiling for one-half minute, and microbes, which survived boiling and developed even after the flask had been sealed. Eventually, he discovered that, after boiling sealed flasks for as long as 45 minutes, no microorganisms developed.

The objection then voiced to Spallanzani's work—that the quality of the air in the flasks had, by heating, been rendered unable to support life—was overcome later (1836) when air was passed into the flask after having been slowly drawn through solutions containing sulfuric acid. The meat extract in the flask remained uncontaminated because the microorganisms in the air were killed by its passage through the solutions. In 1853 it was discovered that flasks did not have to be sealed after boiling but could be closed with a plug of cotton, which effectively filtered the incoming air; this procedure is still used. Because many microorganisms can produce bodies (spores) that are extremely heat-resistant, mediums for the growth of microorganisms are usually sterilized by heating under pressure to 120° C (250° F) for 15 or 20 minutes.

Simplification of the method for sealing flasks

The Scottish surgeon Joseph Lister contributed to culture methods in microbiology with the introduction in 1878 of the dilution method. During studies probably concerning the souring of milk, he used sterile water to dilute a small amount of milk, then diluted some of this mixture with more sterile water. He continued the dilution process until a sample of the milk and sterile water mixture no longer caused the milk to sour. The last dilution that resulted in souring thus contained a minimum number of organisms, and Lister considered it a pure culture—that is, containing only one species of organism. The dilution method has become an essential part of pure-culture methods.

The dilution method was modified in 1896 by a Danish microbiologist, Emil Hansen, who studied yeasts. He added one drop of a yeast culture to the first of a series of small drops of sterile water, then removed a drop of the culture and added it to the second drop of water. Eventually, he obtained a drop containing just one yeast cell.

Koch also made a significant contribution to culture techniques with his development of a simple method for obtaining pure cultures of bacteria. He added a solidifying agent (gelatin) to a nutrient medium. After the medium had been heat sterilized and partially cooled, he added some microorganisms. By cooling the solution further and spreading the organisms before the gelatin solidified, he was able to isolate the microorganisms from each other, with the result that each organism gave rise to a separate colony, or crop of cells.

The introduction of agar-agar

In 1883 a woman in Koch's laboratory, Frau Hesse, further improved the technique by substituting agar-agar for the gelatin. Unlike gelatin, agar-agar can be liquefied by only a few microorganisms and does not provide a food source, thus allowing better control of the nutrient content of the medium. Silica gel also has been used as a solidifying agent.

One final advance in culture technique was developed in the 1890s by M.W. Beirjerinck and by the Russian microbiologist Sergei Winogradsky. They selected a medium that favours the growth of one organism over another. If an organism, one from the soil, for example, is capable of utilizing a certain substance (*e.g.,* a specific sugar), growth of the organism can be induced by using the substance in the medium. Only the organisms capable of attacking the substance will grow in great numbers. Successive transfers of a relatively small number of organisms (inoculum) to a fresh medium eventually will result in a culture strongly enriched with the organism that utilizes the desired substance.

After a species of microorganism has been obtained as a pure culture, it is necessary to maintain it alive and as a pure culture. Microbiology laboratories in schools, universities, and industry usually maintain collections of pure cultures of the particular species they use. Various organizations throughout the world maintain pure cultures of microorganisms; such collections are important in that they make available pure cultures of species when needed.

The Kral Collection in Prague, established in 1900, was the first known culture collection. The names and locations of a few of the other collections are: the American Type Culture Collection (Rockville, Maryland), the Japanese Type Culture Collection (Tokyo), and the (British) National Collection of Type Cultures (London). Other collections also exist worldwide to serve particular needs. A section on culture collections was established in 1966 by the International Association of Microbiological Societies in order to promote the exchange of information among culture collections.

Whenever a microbiologist proposes a new species, he provides one or more of the national culture collections with a pure culture of the species.

Staining techniques. Until the latter third of the 19th

century, microorganisms were observed only in the natural state; the similarity of their refractive index to water and their small size, however, made them difficult to see when viewed through a microscope. Thus, the discovery of a method that would allow them to be seen more easily was an important development.

By 1875 the German pathologist Carl Weigert, using techniques developed more than a decade earlier for staining animal tissues with dyes, had found that dead ("fixed") bacteria would become heavily stained with the dye picrocarmine. Other dyes were then introduced—*e.g.,* methylene blue, fuchsin, and crystal violet. Hans Christian Gram, a Danish bacteriologist, discovered in 1884 a simple procedure for placing bacteria into either a gram-negative class or a gram-positive class, depending on whether or not the organisms retained crystal violet when treated in a specific way. Although the gram stain is used mostly with bacteria, other microorganisms also show a reaction. Yeasts and actinomycetes, for example, are gram-positive; rickettsiae are gram-negative. Stained preparations now are used particularly to observe structural features of microorganisms.

Other methods. Twentieth-century developments in microbiological methods have depended largely upon the techniques of other disciplines—namely, biochemistry, physiology, organic chemistry, and physics.

The gas exchange of respiring microorganisms, called the respiratory quotient (the ratio of carbon dioxide produced to oxygen consumed), is measured by following the pressure change of respiring organisms in a closed vessel. Other processes—*e.g.,* fermentation—can also be studied using this method.

The transfer of hydrogen ions (positively charged atoms) in microorganisms has been studied by adding to the organisms or extracts prepared from them a substance (*e.g.,* the dye methylene blue) the colour of which changes when hydrogenated. Many other dyes behave in this way.

Techniques for breaking apart microorganisms

The techniques used to break up microorganisms include such biochemical techniques as alternate freezing and thawing, prolonged grinding, vigorous shaking with glass beads, and exposure to supersonic vibration; certain microorganisms (*e.g.,* yeast) are especially difficult to disrupt because of their resistant cell walls.

The apparatus of modern physics has greatly aided the advance of virology. Around 1930 the steady development of new technical methods for studying the physical, chemical, and biological properties of microorganisms completely changed the outlook of biologists toward viruses. An important advance in this new approach was the development of graded filter membranes of known pore size. This created for the first time the opportunity to assess the actual size of virus particles.

Another important advance of the 1930s was the growth of the virus of fowl pox in the tissues of a developing chick embryo. Methods for the growth in the laboratory of rickettsiae are similar in that living host cells are necessary. In 1949, tissue culture methods, long used in half-hearted fashion for the cultivation of viruses, were shown to be suitable for the growth of poliomyelitis viruses. This not only opened the way to immunization against polio but supplied a method by which many new types of virus could be isolated.

PHYSICAL ANTHROPOLOGY

Physical anthropology is that branch of natural science that is concerned with the origin and evolution of man. Both the course that human evolution has taken and the causes or processes that have brought it about are of equal concern. In order to interpret the diversity within races and between races, physical anthropologists have had to search for the meaning of these differences by including as objects of study past races of fossil man as well as such nonhuman primates as anthropoid apes, monkeys, tarsiers, and lemurs. Much light has been thrown upon man's relation to other primates and upon the nature of the transformation of his skeleton in the course of evolution from early man to modern man, a span of at least 2,000,000 years. Discoveries of the South African man apes, the first in 1924, and of other forms in East Africa beginning in 1959 have revealed unanticipated data concerning the diverse combinations of traits that can coexist and the singularly illuminating fact that erect posture preceded the great expansion of the brain in human evolution.

Many of the processes responsible for the differentiation of man into different races, although he still remained a single species, *Homo sapiens,* are known: selection, genetic drift, migration, and mutation. Objective methods of isolating various kinds of traits and dealing mathematically with their frequencies, as well as their functional or phylogenetic significance, make it possible to understand the composition of human populations and to formulate hypotheses concerning their future. The resulting information accumulated by physical anthropologists makes available facts about the groups inhabiting the world as well as the individuals composing these groups. Thus, it is possible for a person to learn about his own genetic constitution with reference to traits ranging from blood types to the fissural patterns of his molar teeth and to know the frequency and distribution of these traits in the various races about the world. He is also able to secure some estimate of the probability with which his children will inherit these features.

Historical background. Inasmuch as the history of physical anthropology is, in large part, a history of man's attempt to determine his place in nature, to compare himself with other primates, and to interpret the physical differences, more is to be gained from examining the kinds of problems that have been studied and the nature of the evidence that has been used than from reviewing the succession of great names. Twentieth-century familiarity with primates tends to obscure the fact that the precise distinction between man and the apes, based on evidence secured from actual dissection, had not been made before 1699. At that time Edward Tyson published the comparative anatomy of a chimpanzee and correctly deduced that the chimpanzee was not a human being. Succeeding Tyson is a distinguished series of men who launched into the formidable task of describing and classifying human beings as well as the other primates. Georges-Louis Leclerc de Buffon, Immanuel Kant, Johann Friedrich Blumenbach, Jean-Baptiste Lamarck, and Georges Cuvier in the 18th century all made notable contribution. Of these, Blumenbach (1752–1840) is recognized as the father of physical anthropology. Although many early classifications of human races failed to distinguish between physical and cultural characteristics and leaned heavily upon the concept of primordial types with a corresponding minimization of variability, it is noteworthy that Blumenbach as early as 1775 observed that the "innumerable varieties of mankind run into each other by insensible degrees." He also initiated use of the term Causasian to describe members of the white race, basing the choice of his term upon the race of men of Georgia on the southern slope of the Caucasus Mountains who at the time enjoyed a remarkable reputation for beauty. Blumenbach anticipated 20th-century analyses of locomotion in his critical observation that the chimpanzee was essentially quadrupedal in spite of occasionally erect posture, and he cited as evidence the receding heel bones and the elongated pelvis.

Tyson's comparative anatomy

Great chain of being and the "missing link." A basic concept that constituted the organizing assumption of primate taxonomy, from its inception to well after the formal appearance of a theory of evolution in 1859, is the idea of the "great chain of being" or the hierarchical arrangement of nature. This great explanatory theory made necessary a correlative concept, that of the famous "missing link." The missing link proved to be of less value to scientists than to P.T. Barnum, who entertained the public with specimens of every conceivable link between groups of animals, including for good measure a mermaid. One useful result of this belief in unilinear gradation was the search for previously unknown forms that would complete man's knowledge of the chain of being. With this inclusive charter, the scientists of western Europe were ordained to take an inventory of nature and to determine the appropriate place of each newly described form. In practice, it of course supplied the theory for understanding not only man's place in nature but that of all other organisms and

was extended to cultures as well. In the absence of genetics and a concept of culture, personality traits as well as skin colour were used for classification. The period between Tyson and the later study by Thomas Huxley, *Man's Place in Nature* (1863), is filled with works devoted to the positioning of the anthropoid apes, monkeys, and newly discovered "races." While some persons investigated the anthropoids, others were impressed with the possibility that various groups of unknown aborigines might be the missing link or links. Thus, the "savage Hottentot" or the "stupid native of Nova Zembla" (Novaya Zemlya) was pressed into service to fill the gap between anthropoid apes and man.

Darwin's theory of evolution. Publication of *On the Origin of Species* by Charles Darwin in 1859 provided the outstanding theoretical contribution of the 19th century. The essential concept for the evolution of man was that of natural selection, though many more decades were to pass before its implications were either appreciated or employed. Darwin not only showed that evolution took place in the different organisms but he also provided an explanation that was sound in its major outline and that subsequently was given more precise meaning. The idea that nature selects those forms which are better adapted to a particular geographic zone and way of life laid the basis for understanding the adaptive radiations of the primates. Anthropologists were slow to appreciate the implications of adaptation and continued to measure and to observe traits that were felt to be nonadaptive. Linnaean species, the hypothetical prototypes of which individual specimens were copies in varying degrees of perfection, continued to be described. The weight of attention remained on classification rather than on processes of evolution, and research continued to focus on description without a corresponding concern for understanding the functional significance of traits or the role that they might play in the adaptation of a species to its ecological zone.

Natural selection

The idea of evolution was slowly accepted, accompanied by many rancorous disputes. The chief accomplishment of the period between 1859 and 1900 lay in the recognition of a considerable time depth for man. The system of arranging the primates remained unilinear, but it was extended far back into the geological past. Both fossil races of man and contemporary races were arranged according to their assumed degree of morphological primitiveness. The findings of extinct forms of animals by paleontologists, the excavation of older cultures by archaeologists, and the discovery of Neanderthal skulls all contributed to questioning the scriptural authorities who credited the world with being only approximately 6,000 years old. Combined researches of geologists, paleontologists, archaeologists, and physical anthropologists finally established the antiquity of man as well as of other forms of organic life; and thus the evolutionary sequence became of foremost interest. The first recognized discovery of the extinct race of Neanderthal man was made in 1856 in northwestern Germany and was disputed by many persons, including Rudolf Virchow. An earlier discovery of the Neanderthal skull in Gibraltar in 1848 had been largely ignored. The continuing search for more primitive specimens of fossil man was richly rewarded when in 1891–92 Eugène Dubois discovered the Java ape man, now classified as *Pithecanthropus erectus*, considered one of the earliest known races of man.

New approach. By the end of the 19th century several useful classifications of all the races of the world had been completed. Many of the differences between the races were quite well known. Similarly, differences between man and the anthropoid apes had been inventoried and also between modern man and fossil man. A great deal of attention was paid to the number of differences, and it was assumed that the degree of relatedness was adequately indicated by the degree of morphological similarity.

The year 1900 may be taken as a turning point in the development of a new conceptual approach, though no single date can fully comprehend a shift in such an exceedingly complex continuum. Two major events took place which were to have far-reaching consequences: (1) the rediscovery of Gregor Mendel's two genetic principles by several investigators and (2) Karl Landsteiner's discovery of the ABO blood groups. Mendel had formulated the basic principles of heredity in 1865, but these had passed unnoticed. The inheritance of the blood groups was not at first appreciated, but within 10 years became a focus of research and thus a building stone in modern racial studies. Many new concepts came into use with the increasing influence of genetic theory, such as the breeding population, genetic equilibrium, genetic drift, and the gene frequency method of differentiating populations or races. Attention was immediately directed to the processes of change, those ways in which gene frequencies are modified. At the same time increasing attention was paid to the significance of traits and to the ways in which they were interrelated. Experimental studies were devised to demonstrate the functional significance of differences in morphology. The concept of straight-line evolution (orthogenesis) was qualified, and examples of reversals or major shifts in the direction of evolution were objectively examined. The concept of the missing link went into discard.

Effect of Mendel and Landsteiner

It was recognized that a change in way of life which introduced a species to a new "adaptive plateau" and brought about many radical changes in its structure obviated the need for a belief in so-called missing links. The techniques of anthropometry were reinforced by many other techniques, such as blood typing, so that the physical anthropologists had a much more versatile set of research tools.

Statistical methods. Spanning both the pre-evolutionary and the evolutionary periods of physical anthropology was the development of statistics and their application to the measurement of man, included under biometry. As early as 1835, L.A.J. Quételet (1796–1874) had applied the statistical concept of the normal probability curve to human beings. He demonstrated that most of the statures in a population will cluster about a mean and that the frequency of both taller and shorter statures will diminish in frequency at opposite ends of the curve. Sir Francis Galton (1822–1911) succeeded in constructing new measures of variability, correlation, and regression. Galton also investigated certain aspects of human heredity. Karl Pearson (1857–1936) classified the types of distribution, perfected the coefficient of correlation, developed chi-square, and was a founder of the journal *Biometrika*. Not the least of his many contributions consisted in separating the sample from the population it represented and indicating the necessary restrictions for interpreting a population in the light of information drawn from the sample. Sir Ronald Fisher, continuing the development of modern statistical theory initiated by Pearson, made extensive contributions in the design of experiments and to the field of genetics. One of the trends in the anthropological use of statistics is toward those kinds that preserve the description of the organism as a whole. Analysis of variance, multiple regressions, discriminate functions, generalized distance, and factor analysis all share the criterion of treating different variates as a single coherent vector. Unfortunately, the finest statistical treatment will not compensate for poorly selected units of observation or measurement.

The measurement of man

Areas of study. Insight into the field of physical anthropology may be obtained by reviewing the general fields of investigation in which particular researchers are actively engaged.

Human ecology. The relationship of the human organism to its environment, or human ecology, is a bridge between the biological and social sciences. Problems of population, size, and stability are important in many ways. A very immediate aspect is the varying rate of change that may occur in populations of different sizes. Theoretically, a small population is more susceptible to chance fluctuations than a large one. Both the natural environment and the economy of a particular society affect the population size.

Studies of evolution. Human evolution is another area that serves as a focus for research. The essential problems are not only the complete description of fossil forms but the evaluation of the significance of particular traits. Old concepts of orthogenesis, or evolution in a straight line, have had to be abandoned, and radiant and parallel evolution have come to the fore. It is clear that the course

of human evolution may swing about so that what were formerly considered relatively modern forms may actually precede morphologically earlier ones, as in the case of the more modern kind of Neanderthal man who apparently preceded the more primitive or early form. Fossil man of considerable antiquity has been found in Europe, Asia, and Africa. No area in the world is without skeletal remains, so that researchers may everywhere engage in skeletal studies.

Primatology. Primatology is an area in which man's place in nature is no longer the major point of focus. Researchers are concerned with experimental, medical, and ecological aspects of primates, and it is from these studies that much of the functional significance of various bone and muscle complexes has been derived. The different kinds of adaptations that groups of monkeys have made to life in the trees or to life on the ground have had many consequences in the proportions of their limbs or in the elaboration of various groups of muscles. Similarly, the mode of progression through trees by the use of the arms, known as brachiation, is accompanied by the elongation of the arms, reduced size of the thumb, and many other structural arrangements in the anthropoid apes. Thus the primates provide a natural laboratory of many kinds of experiments in physical adaptation to fundamentally different ways of life or adaptive zones.

Genetics. Genetics is a fourth major area of research. The study of inherited traits in individuals and the behaviour of the genes responsible for these traits in populations is essential to understanding human variability. Although the blood groups have provided the bulk of the data, many other traits are being analyzed and theories of their inheritance tested. As a result of research in this area it is possible to describe races in terms of gene frequencies and to calculate amounts of race mixture.

Growth studies. Growth studies, both of human beings and other primates, engage the attention of many physical anthropologists in medical and dental schools as well as in independent clinics and universities. Methods of assessing rates of growth, skeletal age compared with chronological age, and the genetic, endocrinological, and nutritional factors are some of the aspects involved in these studies. The relation between growth and socioeconomic status and other cultural features receives considerable attention. Dentitions have been frequently studied, often because the emergence of the teeth is of great practical importance and serves as an index of development. As a result of research in this area it is possible to describe races in terms of gene frequencies and to calculate amounts of race mixture.

Anthropometry. Measurement, or anthropometry, has been a mainstay of anthropological research for well over a century. More attention is being given to the judicious selection of measurements than to the simple techniques with which calipers are handled. Many traditional measurements cannot be analyzed, and the addition of multivariate statistics does not make them any more intelligible. Statistical considerations are especially important in genetic and anthropometric research, and part of the history of statistics is identical with the history of the development of these two research interests. Quite certainly the end results of growth can only be studied by measurement. Newer growth studies have followed children through their morphological and biochemical changes with the aim of discerning why children grow.

Applications of physical anthropology

Many of the applications of physical anthropology lie in the field of measurement. Thus the problems involved in providing military and civilian clothing for large numbers of people depend on measurement and statistical treatment. Substantial savings have been made possible by measuring the people of a particular area and adjusting the clothing tariffs to these known distributions of body sizes. There is sufficient variation within most countries so that the geographical variations in size of the body and intermembral proportions are of practical importance. Human constitution is another area of research interest. Several descriptive systems exist for classifying persons who vary from a lateral or relatively squat body build through a relatively muscular to a linear body build. The components of body build, the different tissues and dimensions, were being studied in the late 20th century by means of factor analysis and comparisons of siblings and twins, and their actual mode of inheritance and response to environmental conditions were slowly being specified. The various areas of interest referred to above are, of course, not mutually exclusive; ideally they should all form part of the training of researchers in the origin and evolution of man. (Ed.)

BIBLIOGRAPHY. The following publications deal with fundamental biological concepts as well as with the structure and functions of living things: JEFFREY J.W. BAKER, *Cell* (1966), a paperback that presents the structure of the cell as seen by the electron microscope; and, with GARLAND E. ALLEN, *Matter, Energy, and Life,* 2nd ed. (1970), a fundamental and easily understood basic chemistry of living systems, and *The Study of Biology,* 2nd ed. (1971), a general college-level survey of biological principles; GEORGE W. and MURIEL BEADLE, *The Language of Life* (1966), the chemistry of the gene explained in an easily understood manner; PETER R. BELL and C.L.F. WOODCOCK, *The Diversity of Green Plants* (1968), a compact survey of the multiformity of green plants; ERNEST BOREK, *The Atoms Within Us* (1961), a nontechnical paperback explaining the chemical processes involved in living matter; A.J. CARLSON, V. JOHNSON, and H.M. CAVERT, *The Machinery of the Body,* 5th ed. rev. (1961), metabolism, digestion, and other human physiological processes clearly explained in this excellent text; RACHEL CARSON, *The Sea Around Us,* rev. ed. (1966), a beautifully written book that discusses food chains and ecological problems of the sea in an almost poetic fashion; T. DOBZHANSKY, *Evolution, Genetics, and Man* (1955), an excellent textbook on the relationship of the evolutionary theory to man; L.C. DUNN, *Heredity and Evolution in Human Populations,* rev. ed. (1967), a nontechnical and interesting explanation of genetics and the evolution of man; JAMES D. EBERT and IAN M. SUSSEX, *Interacting Systems in Development,* 2nd ed. (1970), one of an excellent series of books, each developed around a different biological concept; PAUL R. and ANNE H. EHRLICH, *Population, Resources, Environment: Issues in Human Ecology* (1970), a general textbook of ecological problems written for the layman; LOREN EISELEY, *The Immense Journey* (1957), a well-written and most interesting story of man's development as an organism; DONALD KENNEDY, *The Living Cell* (1965), a collection of articles from *Scientific American* with good photographs of cell structure; H.R. MAHLER and E.H. CORDES, *Biological Chemistry,* 2nd ed. (1971), a standard college-level text of biochemistry; A.I. OPARIN, *The Chemical Origin of Life* (1964; originally published in Russian, 1936), an updated version of Oparin's theory of the origin of life; E.P. ODUM, *Fundamentals of Ecology,* 3rd ed. (1971), a standard college-level text on ecology with an excellent discussion of the cycles in nature; JAMES A. PETERS (ed.), *Classic Papers in Genetics* (1959), original papers by outstanding geneticists; A.E. ROMER, *Man and the Vertebrates* (1941; rev. as *The Vertebrate Story,* 1959), a comparative study of the anatomy of man and other vertebrates with emphasis on evolutionary significance; M.W. STRICKBERGER, *Genetics* (1968), a standard college text with a clearly written section on population genetics; N. TINBERGEN, *Social Behavior in Animals,* 2nd ed. (1965), an introductory, well-written book about animal behaviour; FRITS WENT *et al., The Plants* (1963), presents the structure, evolution, and function of plants in a nontechnical and well-illustrated fashion.

History of biology: Publications concerned with the history and philosophy of biology include: ISAAC ASIMOV, *A Short History of Biology* (1964), a book written for the layman, emphasizing accomplishments of the 19th and 20th centuries; SIR GAVIN DE BEER, *Charles Darwin* (1963), a biography of Darwin by an eminent biologist; M. BERGER, *Famous Men of Modern Biology* (1968), brief biographies of 19th- and 20th-century biologists; F.S. BODENHEIMER, *The History of Biology: An Introduction* (1958), contains a short history, but also a source reader from 130 authors, beginning with the Egyptians; MORDECAI L. GABRIEL and SEYMOUR FOGEL (eds.), *Great Experiments in Biology* (1955), a presentation of scientific writings in the original, from Robert Hooke to the 20th century; ELDON J. GARDNER, *History of Biology,* 2nd ed. (1965), a college-level text, and *History of Life Science* (1960), an outline of biological history written for the biology student; PHILIP GOLDSTEIN, *Triumphs of Biology* (1965), a survey of selected aspects of biology, written for college level; URLESS LANHAM, *Origins of Modern Biology* (1968), a survey of biology from Greece to the present; RUTH MOORE, *The Coil of Life* (1961), a clearly written description of the development of molecular biology for the general reader; CHARLES SINGER, *A History of Biology,* rev. ed. (1950), a highly readable classic that surveys the historical development of biological problems; M.J. SIRKS and CONWAY ZIRKLE, *The Evolution of Biology* (1964), a college-level survey of the history of biology from prehistory to the present; HENRY OSBORN TAYLOR, *Greek Biology and Medicine* (1922, reprinted

1963), a sketch for the layman of the scientific method of Greece and Rome.

Morphology: B.I. BALINSKY, *An Introduction to Embryology,* 3rd ed. (1970), an excellent general college-level text that includes details of organogenesis in each of the vertebrate types, plus a synthesis of the experimental analysis of cellular differentiation; R.D. BARNES, *Invertebrate Zoology,* 2nd ed. (1968), an excellent modern treatment of invertebrate morphology; E.D.P. DE ROBERTIS, W.W. NOWINSKI, and F.A. SAEZ, *Cell Biology,* 5th ed. (1970), an excellent text describing the cytological features of animal and plant cells; KATHERINE ESAU, *Plant Anatomy,* 2nd ed. (1965), a beautifully illustrated text with an excellent treatment of plant anatomy; DON FAWCETT, *The Cell* (1966), a collection of superb electron micrographs of several kinds of cells and cell organelles, with a brief description of each; LIBBIE H. HYMAN, *The Invertebrates,* 6 vol. (1940–67), the definitive treatise of the invertebrates, covering morphology, development, and evolution; A.S. ROMER, *The Vertebrate Body,* 4th ed. (1969), a standard college-level text of comparative anatomy of the vertebrates; C.A. VILLEE, W.F. WALKER, and R.D. BARNES, *General Zoology,* 4th ed. (1973), a standard college-level text of vertebrate and invertebrate morphology and its relation to function.

Physiology: K.J. FRANKLIN, *A Short History of Physiology,* 2nd ed. (1949), through the 19th century; J.F. FULTON (ed.), *Selected Readings in the History of Physiology* (1930), excerpts from classical publications; T.S. HALL, *Ideas of Life and Matter* (1969), on concepts of physiology and related fields; K.E. ROTHSCHUH, *Geschichte der Physiologie* (1953), a detailed account in German, with emphasis on German physiologists; J. BARCROFT, *Features in the Architecture of Physiological Function* (1934), broad interpretations of physiology; W.B. CANNON, *The Wisdom of the Body* (1932), on regulatory physiology; A.V. HILL, *Adventures in Biophysics* (1931), studies on the biophysics of muscle; M. FOSTER, *Lectures on the History of Physiology During the 16th, 17th, and 18th Centuries* (1901, reissued 1970), a classic by one of the founders of modern physiology; C.S. SHERRINGTON, *Man on his Nature,* 2nd ed. (1951), a summary of and reflections on classical work in neurophysiology; E.F. ADOLPH, *Origins of Physiological Regulation* (1968), a broad survey of regulatory physiology; AMERICAN PHYSIOLOGICAL SOCIETY, *Handbook of Physiology* (1959–), a series of volumes summarizing in detail the status of research in specialized areas of physiology; H. DAVSON, *A Textbook of General Physiology,* 3rd ed. (1964), a revision of a classical English text by SIR WILLIAM BAYLISS; A.C. GIESE, *Cell Physiology,* 3rd ed. (1968), a standard text; B.T. SCHEER, (ed.), *Comparative Physiology* (1968), a collection of recent research papers.

Taxonomy: GEOFFREY C. AINSWORTH, and P.H.A. SNEATH (eds.), *Microbial Classification* (1962), a symposium of the Society of General Microbiology, with discussion of the principles of classification and their application to microbial groups; GUY R. BISBY, *An Introduction to the Taxonomy and Nomenclature of Fungi,* 2nd ed. (1953); ARTHUR J. CAIN (ed.), *Function and Taxonomic Importance* (1959), a symposium of the Systematics Association on the relevance of function to classification, especially to convergence; OLOV HEDBERG (ed.), *Systematics of Today* (1958), symposium held at the University of Uppsala to commemorate the 250th anniversary of Linnaeus's birth; JOHN G. HAWKES (ed.), *Symposium on Chemotaxonomy and Serotaxonomy* (1968); WILLI HENNIG, *Phylogenetic Systematics* (1966), a statement of phylogenetic "principles"; VERNON H. HEYWOOD and J. MCNEILL (eds.), *Phenetic and Phylogenetic Classification* (1964); ROBERT R. SOKAL and P.H.A. SNEATH, *Principles of Numerical Taxonomy* (1963), an important introduction to numerical taxonomy. General texts include RICHARD E. BLACKWELDER, *Taxonomy* (1967); VERNON H. HEYWOOD and PETER H. DAVIS, *Principles of Angiosperm Taxonomy* (1963); JULIAN S. HUXLEY (ed.), *The New Systematics* (1940); ERNST MAYR, *Principles of Systematic Zoology* (1969); GEORGE G. SIMPSON, *Principles of Animal Taxonomy* (1961). Also consult INTERNATIONAL ASSOCIATION OF MICROBIOLOGICAL SOCIETIES, *International Code of Nomenclature of Bacteria and Viruses* (1958); JOSEPH LANJOUW (ed.), *International Code of Botanical Nomenclature* (1966); INTERNATIONAL COMMISSION ON ZOOLOGICAL NOMENCLATURE, *International Code of Zoological Nomenclature* (1961).

Biophysics: LUIGI GALVANI, *De viribus electricitatis in motu musculari commentarius* (1791; Eng. trans., *Commentary on the Effect of Electricity on Muscular Motion,* 1953), Galvani's own account of the discovery of animal electricity; H. BENCE JONES (ed.), *On Animal Electricity, Being an Abstract of the Discoveries of Emil Du Bois-Reymond* (1852), an English abstract of the work of this great 19th-century German scientist; A.V. HILL, "Why Biophysics?" *Science,* 124: 1233–1237 (1956), probably the best concise statement of the attributes of a biophysicist and the qualities needed to make important scientific contributions in the field; NOBEL FOUNDATION, *Les Prix Nobel en 1962* (1963), contains the lectures delivered by Crick, Wilkins, Watson, Kendrew, and Perutz when they received the Nobel Prize; R. OLBY, "Francis Crick, DNA and the Central Dogma," *Daedalus,* 99:938–987 (1970), a good historical account of the discovery of DNA and the subsequent events; J.D. WATSON, *The Double Helix* (1968), a fascinating first-person account of the discovery of the molecular conformation of DNA; A.L. HODGKIN, "The Ionic Basis of Nervous Conduction," *Science,* 145:1148–1153 (1964), his Nobel prize lecture; A.F. HUXLEY, "Excitation and Conduction in Nerve: Quantitative Analysis," *Science,* 145:1154–1159 (1964), his Nobel prize lecture; H.E. HUXLEY, "The Mechanism of Muscular Contraction," *Sci. Am.,* 213:18–27 (1965), a definitive semipopular exposition of the mechanism of muscular contraction; A.K. SOLOMON, "A Short History of the Foundation of the International Union for Pure and Applied Biophysics," *Q. Rev. Biophys.,* 1:107–124 (1968).

Biochemistry: JOSEPH NEEDHAM (ed.), *The Chemistry of Life: Eight Lectures on the History of Biochemistry* (1970), provides brief development of the important areas of photosynthesis, enzymes, microbiology, neurology, hormones, vitamins, and other topics. Two selected biographical treatments that couple human interest with development of biochemical areas are FRITZ A. LIPMANN, *Wanderings of a Biochemist* (1971); and DAVID KEILIN, *The History of Cell Respiration and Cytochrome* (1966). Two monographs that describe areas in the forefront of biochemistry are JAMES D. WATSON, *Molecular Biology of the Gene,* 2nd ed. (1970); and THOMAS H. JUKES, *Molecules and Evolution* (1966). Two attractively illustrated textbooks are ALBERT L. LEHNINGER, *Biochemistry: The Molecular Basis of Cell Structure and Function* (1970); and ROBERT W. MCGILVERY, *Biochemistry: A Functional Approach* (1970). Other standard and somewhat more comprehensive texts are ABRAHAM WHITE, PHILIP HANDLER, and EMIL L. SMITH, *Principles of Biochemistry,* 5th ed. (1973); and EDWARD S. WEST *et al., Textbook of Biochemistry,* 4th ed. (1966). A useful set of paperbacks is *Essays in Biochemistry,* published for the Biochemical Society (1965–). More comprehensive survey treatises are MARCEL FLORKIN and ELMER H. STOTZ (eds.), *Comprehensive Biochemistry* (1962–); and MARCEL FLORKIN and HOWARD S. MASON (eds.), *Comparative Biochemistry,* 7 vol. (1960–64).

Genetics: A.G. DEBUSK, *Molecular Genetics* (1968), a paperback providing concise coverage of this subject; L.C. DUNN, *A Short History of Genetics* (1965), a paperback outlining the major features of the development of genetics; I.M. LERNER, *Heredity, Evolution and Society* (1968), a discussion of the problems of society as related to genetic discoveries; J.A. PETERS (ed.), *Classic Papers in Genetics* (1959), reprints of reports of important genetic discoveries; CURT STERN, *Principles of Human Genetics,* 3rd ed. (1971), a classic work in the field, thorough coverage; M.W. STRICKBERGER, *Genetics* (1968), a college-level treatment of modern genetics; R.P. WAGNER and H.K. MITCHELL, *Genetics and Metabolism,* 2nd ed. (1964), a comprehensive account of the relationship between genes and cellular activities; A.M. WINCHESTER, *Genetics* (1971), an introductory genetics text, including all principles of the subject, but slanted toward human genetics, and *Heredity: An Introduction to Genetics,* 2nd ed. (1966), providing concise coverage of the subject in terms understandable to the general reader.

Eugenics: Classic works include C.P. BLACKER, *Eugenics: Galton and After* (1952); THEODOSIUS G. DOBZHANSKY, *Mankind Evolving: The Evolution of the Human Species* (1962); FRANCIS GALTON, *Hereditary Genius: An Enquiry into Its Laws and Consequences* (1869); THOMAS HUNT MORGAN, *The Theory of the Gene* (1926); and J. SUTTER, *L'Eugénique* (1950). See also R. FREEDMAN, P.K. WHELPTON, and A.A. CAMPBELL, *Family Planning, Sterility, and Population Growth* (1959); FREDERICK H. OSBORN, *Preface to Eugenics,* rev. ed. (1951); C.F. WESTOFF *et al., Family Growth in Metropolitan America* (1961); C.F. WESTOFF, R.G. POTTER, JR., and P.C. SAGI, *The Third Child: A Study in the Prediction of Fertility* (1963); P.K. WHELPTON, A.A. CAMPBELL, and J.E. PATTERSON, *Fertility and Family Planning in the United States* (1966); P.K. WHELPTON and CLYDE V. KISER (eds.), *Social and Psychological Factors Affecting Fertility,* 4 vol. (1946–54); and ROGER J. WILLIAMS, *Biochemical Individuality: The Basis for the Genetotrophic Concept* (1956).

Ecology: W.C. ALLEE *et al., Principles of Animal Ecology* (1949), a somewhat dated but still important source book in animal ecology, with an excellent early history of ecology; H.G. ANDREWARTHA, *Introduction to the Study of Animal Populations* (1961), a concise summary of an important and controversial earlier text, *The Distribution and Abundance of Animals,* by Andrewartha and L.C. BIRCH (1954); PETER FARB, *Ecology* (1963), an enjoyable introduction to ecology, heavily illustrated with outstanding photographs; E.B. FORD, *Ecological Genetics* (1964), a basic reference on adaptation and adjustments of wild populations to their environment; E.J. KORMONDY, *Concepts*

of Ecology (1969), a short, interesting introduction to major concepts of ecology, and (ed.), *Readings in Ecology* (1965), a collection of annotated research papers that illustrate many important concepts of ecology, and, ed. with R.S. LEISNER, *Ecology* (1971), a collection of essays; G.K. REID, *Ecology of Inland Waters and Estuaries* (1961), a good summary of ecological principles as applied to freshwater and estuarine ecosystems; P. SHEPARD and D. MCKINLEY (eds.), *The Subversive Science: Essays Toward an Ecology of Man* (1969), a collection of thoughtful and provocative papers written by contemporary students of the ecology of human populations; R.L. SMITH, *Ecology and Field Biology* (1966), a comprehensive survey of all aspects of ecology; T.R.E. SOUTHWOOD, *Ecological Methods* (1966), an advanced book on the study of populations, with emphasis on insects and quantitative analysis; G.M. VANDYNE (ed.), *The Ecosystem Concept in Resource Management* (1969), a technical review of the application of ecological theory to forest range, watershed, and wildlife management; K.E. WATT, *Ecology and Resource Management* (1968), a highly technical but important text relating mathematical models and computer programming to resource management; C.B. WILLIAMS, *Patterns in the Balance of Nature and Related Problems in Quantitative Ecology* (1964), an advanced treatise.

Zoology: P.R. and A.H. EHRLICH, *Population Resources Environment: Issues in Human Ecology* (1970), vigorous documentation of our misuse of the environment, P.P. GRASSE (ed.), *Traité de zoologie,* 23 vol. (1948–), in French, the most recent encyclopaedic survey of animal biology; P. HANDLER (ed.), *Biology and the Future of Man* (1970), a brilliant survey of the current status of biological knowledge and its relation to human life with an excellent chapter on computer technology; L.H. HYMAN, *The Invertebrates,* 6 vol. (1940–68), a technical but lucid survey of animal biology, exclusive of the vertebrates; W. LEY, *The Dawn of Zoology* (1968), a fascinating account of certain historical aspects of zoology; E. NORDENSKIOELD, *Biologins historia,* 3 vol. (1920–24; Eng. trans., *The History of Biology,* 1928), a basic source for tracing the major ideas and personalities in the life sciences, although occasionally presenting interpretations that have been replaced.

Botany: H.S. REED, *A Short History of the Plant Sciences* (1942), an interesting and thought-provoking account with many useful insights; K.V. THIMANN, *The Plant Sciences, Now and in the Coming Decade* (1966), a thoughtful and persuasive statement, well-written and readable; J. REYNOLDS GREEN, *A History of Botany, 1860–1900* (1909, reprinted 1967), a scholarly treatment designed to supplement the classic work of JULIUS VON SACHS, *History of Botany, 1530–1860,* trans. by HENRY E.F. GARNSEY, rev. by ISAAC BAYLEY BALFOUR (1890, reprinted 1967), arranged by botanical structure and function; AGNES ARBER, *Herbals, Their Origin and Evolution: A Chapter in the History of Botany, 1470–1670,* 2nd ed. rev. (1938, reprinted 1970), a readable and authoritative work; ELLISON HAWKS and G.S. BOULGER, *Pioneers of Plant Study* (1928, reprinted 1969), an interesting account of major historical figures.

Microbiology: G.C. AINSWORTH, and P.H.A. SNEATH (eds.), *Microbial Classification* (1962), a series of essays about the problems of classifying microorganisms, includes names and addresses of institutions maintaining culture collections; T.D. BROCK (ed. and trans.), *Milestones in Microbiology* (1961), a collection of original papers important in the history of microbiology; ALBERT DELAUNAY *et al.* (eds.), *The World of Microbes,* vol. 4 of the *Encyclopedia of the Life Sciences* (1965), a well-illustrated account of microbial activities in relation to man; C.L. DUDDINGTON, *Micro-organisms As Allies* (1961), an interesting account of the industrial uses of bacteria and fungi; R.K. JENNINGS and R.F. ACKER, *The Protistan Kingdom* (1970), an informative and entertaining account of protists and viruses; H.A. LECHEVALIER and MORRIS SOLOTOROVSKY, *Three Centuries of Microbiology* (1965), a reconstruction of the main aspects of the development of microbiology; JOHN POSTGATE, *Microbes and Man* (1969), an introduction to the importance of microorganisms to life on Earth; THEODOR ROSEBURY, *Life on Man* (1969), a witty and informative book.

Physical anthropology: P. TOPINARD, *Eléments d'anthropologie générale* (1885), is the oldest of the treaties; A.E. MOURANT, *The Distribution of the Human Blood Groups* (1954); W.C. BOYD, *Genetics and the Races of Man: An Introduction to Modern Physical Anthropology* (1950); R. KHERUMIAN, *Génétique et anthropologie des groupes sanguins* (1951); G. HEBERER, G. KURTH, and I. SCHWIDETZKY-ROESING, *Anthropologie* (1959; Eng. trans., *Anthropology A to Z* (1963); ASHLEY MONTAGU, *An Introduction to Physical Anthropology,* 3rd ed. (1960); J. COMAS, *Manual of Physical Anthropology,* rev. ed. (1960); C.S. COON, *The Living Races of Man* (1965); R. MARTIN and K. SALLER, *Lehrbuch der Anthropologie,* 4 vol (1957–66); and H.V. VALLOIS, *Les Races humaines,* 7th ed. (1967).

(E.R.G./C.A.V./B.T.S./A.J.Ca./A.K.S./E.H.St./B.V./A.M.W./F.H.O./R.L.Sm./E.D.H./W.C.St./Ed.)

The Biosphere

Before the coming of life, the Earth was a bleak place, a rocky globe with shallow seas and a thin band of gases—largely carbon dioxide, carbon monoxide, molecular nitrogen, hydrogen sulfide, and water vapour. It was a hostile and barren planet. This strictly inorganic state of the Earth is called the geosphere; it consists of the lithosphere (the rock and soil), the hydrosphere (the water), and the atmosphere (the air). Energy from the Sun relentlessly bombarded the surface of the primitive Earth and in time—millions of years—chemical and physical actions produced the first evidence of life, formless, jelly-like blobs that could collect energy from the environment and produce more of their own kind. This generation of life in the thin outer layer of the geosphere established what is called the biosphere, the "zone of life," an energy-diverting skin that uses the matter of the Earth to make living substance.

The biosphere is a system characterized by the continuous cycling of matter and an accompanying flow of solar energy in which certain large molecules and cells are self-reproducing. Water is a major predisposing factor, for all life depends on it. The elements carbon, hydrogen, nitrogen, oxygen, phosphorus, and sulfur, when combined as proteins, lipids, carbohydrates, and nucleic acids, provide the building blocks, the fuel, and the direction for the creation of life. Energy flow is required to maintain the structure of organisms by the formation and splitting of phosphate bonds. Organisms are cellular in nature and always contain some sort of enclosing membrane structure, and all have nucleic acids that store and transmit genetic information.

All life on Earth depends ultimately upon green plants, as well as upon water. Plants utilize sunlight in a process called photosynthesis to produce the food upon which animals feed and to provide, as a by-product, oxygen, which most animals require for respiration. At first, the oceans and the lands were teeming with large numbers of a few kinds of simple single-celled organisms, but slowly plants and animals of increasing complexity evolved. Interrelationships developed so that certain plants grew in association with certain other plants, and animals associated with the plants and with one another to form communities of organisms, including those of forests, grasslands, deserts, dunes, bogs, rivers, and lakes. Living (biotic) communities and their nonliving (abiotic) environment are inseparably interrelated and constantly interact upon each other. For convenience, any segment of the landscape that includes the biotic and abiotic components is called an ecosystem. A lake is an ecosystem when considered in totality as water, nutrients, climate, and all of the life contained within it. A given forest, meadow, or river is likewise an ecosystem. One ecosystem grades into another along zones termed ecotones, where a mixture of plant and animal species from the two ecosystems occurs. A forest considered as an ecosystem is not simply a stand of trees but is a complex of soil, air, and water, of climate and minerals, of bacteria, viruses, fungi, grasses, herbs, and trees, of insects, reptiles, amphibians, birds, and mammals.

Stated another way, the abiotic, or nonliving, portion of each ecosystem in the biosphere includes the flow of energy, nutrients, water, and gases and the concentrations of organic and inorganic substances in the environment. The biotic, or living, portion includes three general categories of organisms based on their methods of acquiring energy: the primary producers, largely green plants; the consumers, which include all the animals; and the decomposers, which include the microorganisms that break down the remains of plants and animals into simpler components for recycling in the biosphere. Aquatic ecosystems are those involving marine environments and freshwater environments on the land. Terrestrial ecosystems are those based on major vegetational types, such as forest, grassland, desert, and tundra. Particular kinds of animals are associated with each such plant province.

Ecosystems may be further subdivided into smaller biotic units called communities. Examples are the organisms in a stand of pine trees, on a coral reef, in a cave, a valley, a lake, or a stream. The major consideration in the community is the living component, the organisms; the abiotic factors of the environment are excluded.

A community is a collection of species populations. In a stand of pines, there may be many species of insects, of birds, of mammals, each a separate breeding unit but each dependent on the others for its continued existence. A species, furthermore, is composed of individuals, single functioning units identifiable as organisms. Beyond this level, the units of the biosphere are those of the organism: organ systems composed of organs, organs of tissues, tissues of cells, cells of molecules, and molecules of atomic elements and energy. The progression, therefore, proceeding upward from atoms and energy, is toward fewer units, larger and more complex in pattern, at each successive level (see Figure 1).

This article focuses on the makeup of the biosphere and examines the relationships between its principal components, including man. The characteristics and dynamics of biological populations and communities are dealt with, as are the interactions that constitute the primary stabilizing links among the constituent organisms. Due attention is also given to the distribution patterns of these biotic units and to the processes that produced such patterns. The major aquatic and terrestrial ecosystems of the Earth are treated in some detail. Other points include energy transformations and transfers within the biosphere and the cyclic flow of materials needed for life. For the development, methodology, and applications of the study of interrelations of organisms with their environment and each other, see BIOLOGICAL SCIENCES, THE: *The study of biological structure and function: Ecology.* Further treatment of the various aquatic and terrestrial environments is provided in OCEANS; LAKES; RIVERS; and CONTINENTAL LANDFORMS. For a discussion of the origin of life on Earth and the varieties of and commonalities among organisms, see LIFE.

For coverage of related topics in the *Macropædia* and *Micropædia,* see the *Propædia,* sections 351, 352, 354, and 355, and the *Index.* (D.M.G./Ed.)

The article is divided into the following sections:

THE ZONE OF LIFE: AN OVERVIEW

Energy flow and material cycling

THE FLOW OF ENERGY IN THE BIOSPHERE

Energy and organization. A unique characteristic of life is that it is an organized system capable of creating more order from less order. This seems contrary to the general trend of the universe, in which there is a tendency to move toward maximum disorder (entropy) as expressed in the second law of thermodynamics. Life results from the steady-state, or balanced, situation in which there is a flow of energy from the Sun to the Earth and then to the cosmic cold of outer space. The first law of thermodynamics, the conservation of energy, states that energy is neither created nor destroyed but remains constant for the universe. It does not indicate how energy is transformed but only that all the energy within the system must be accounted for.

The constancy of energy in the universe

In order for sunlight to organize life on Earth, it must irradiate the surface with electromagnetic frequencies harmonious with the peculiar chemical bonds of organic molecules. Whereas the short-wave ultraviolet rays, gamma rays, and X rays are destructive to organic molecules, and the long-wave infrared radiation is absorbed and dissipated as heat, the near-ultraviolet and visible wavelengths interact well with matter to stimulate the formation of bonds and the arrangement of organic molecules.

The single most important photochemical reaction in the world is photosynthesis, the union of carbon dioxide and water in plants through the interaction of sunlight and chlorophyll molecules. In the photosynthetic process, light energy is absorbed by chlorophyll to convert carbon dioxide and water into carbohydrate and oxygen. This photochemical event is a stepwise process by which electrons are energized within the chlorophyll molecule and raised to higher energy levels with the formation of carbohydrate. In the process, which is complex, high-energy phosphate bonds are formed, and adenosine triphosphate (ATP) results. The end products within green plants are carbohydrates and energy-rich compounds that become food for plant-eating (herbivorous) organisms and a succession of animal-eating (carnivorous) organisms. The food web of ecosystems is provisioned in this way.

The universal fuel within all organisms is ATP, in which high-energy phosphate bonds store energy and release it when the bonds are broken. The complex reactions within a cell that lead to the formation of ATP combine at least one phosphorus atom per adenosine molecule. Without phosphorus no life could exist, since the entire linkage between photosynthesis and cellular activity would be missing. One process by which the energy stored in ATP is utilized is in the transformation of glucose sugar to sucrose, or ordinary sugar.

Phosphorous

Efficiency of utilization of solar energy. The intensity, as well as the composition, of the sunlight irradiating the Earth's surface is important to life. The tremendous amount of solar energy incident upon the Earth's outer atmosphere each day is distributed unevenly over the world, with the greatest amount of solar radiation reaching the desert regions and the least amount reaching the polar regions, where the slanted rays of the sunlight through the atmosphere are long. Temperate regions of the world, where food crops are grown, receive an intermediate amount of solar radiation, but a substantial fraction of this sunlight is received during the winter, when temperatures are too low for maximum plant growth. When intense winter cold grips the Arctic landscape, there is little light and no plant growth; only higher members of the food chain, birds and mammals, move about the land. In the tropics, temperatures are warm and constant throughout the year; the land is irradiated by sunlight more evenly throughout the seasons and plant growth is abundant. Although desert regions receive by far the greatest amount of sunlight, the lack of water limits plant productivity.

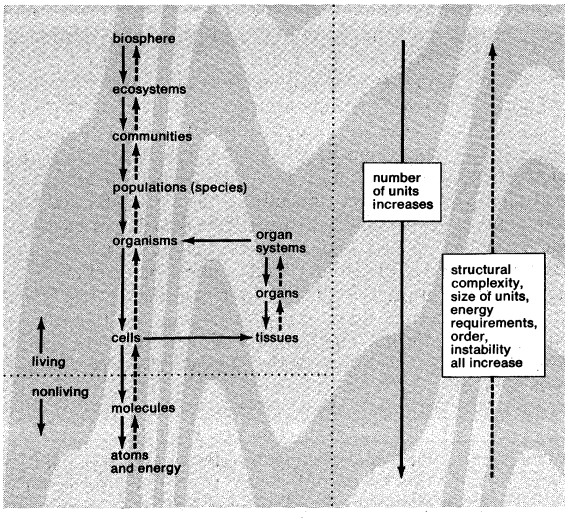


Figure 1: Hierarchy of levels in the biosphere.

The conversion of energy into matter

Only about 25 percent of the sunlight reaching the ground is in wavelengths useful for photosynthesis, and only a fraction of useful light is available to green plants. An understanding of the performance of the biosphere requires an appreciation of the efficiency of the primary production process. The efficiency of energy conversion from incident sunlight into organic matter (net primary productivity) seldom is as large as 3 percent and usually is 1 percent or less. Estimates for a cornfield, for example, show that only slightly more than 1 percent of the total sunlight striking it ends up as part of the corn plants; about 44 percent is used in the evaporation of water from soil and plants, 54 percent is reflected or dissipated as heat, and a fraction of a percent is consumed by the respiration of plants and animals in the field. This is not to say that all these other processes are not important, for indeed they are necessary to the proper functioning of the biosphere, but only a little more than 1 percent of the total incident energy ends up in new plant material. Other estimates of the net productivity of various vegetation types are:

	percent productivity
Tropical rain forest	1.4
Perennial grass–herb field	1.05
Coniferous forest	0.9
Deciduous temperate forest	0.4
Temperate lake	0.3
Desert	0.03

Animals grazing on the primary production convert only about 10 to 15 percent of the energy stored in plant material into animal tissue, and humans eating the animals convert only about 1 percent of this energy stored in animal tissue into human tissue. The total conversion process from sunlight to plants to animals to humans operates at an overall efficiency of about 0.001 percent. For a discussion of productivity in terms of the function of biological communities, see below *Biological communities, biomes, and ecosystems: Community function.*

Energy balance of organisms. Organisms utilize heat energy as well as light. All plants and animals must remain in reasonable energy balance. A plant leaf has a temperature that depends upon the amount of radiation absorbed, the flow of air over its surface, and its rate of transpiration, or gas exchange. The temperature a plant leaf assumes for any set of environmental conditions is of great importance to the plant, since all chemical reactions, including photosynthesis, are influenced by temperatures. A plant is coupled to its environment through the exchange of energy, gases, and nutrients.

A warm-blooded (homeothermic) animal remains in energy balance with approximately a constant body temperature. It moves about in its environment in such a manner that the amount of radiation it absorbs, the amount of energy it exchanges by convectional heating or cooling, and the quantity of energy it consumes by respirational water loss and evaporative cooling are such that its body temperature remains within a narrow range. A cold-blooded (poikilothermic) animal, on the other hand, has much less control over its body temperature. It nevertheless seeks the combination of radiation, air temperature, wind speed, and humidity that will result in an energy flow that is compatible with its body-temperature limits. Each and every animal in the world, depending upon its coloration, body size, shape, quantity of thermal insulation (fat, fur, or feathers), metabolic and water loss rates, and preferred body-temperature range and limits, has a well-defined climate space within which it must live in order to survive. Each species is restricted to its particular climate space as defined by its physiology and general body characteristics. Many animals will live in a somewhat restricted climate space, but none can live beyond its maximum climate space limits. The animals of the world are therefore distributed accordingly (see below *The distribution of organisms*).

Climate space

THE CYCLING OF MATTER IN THE BIOSPHERE

The general pattern of chemical cycles in nature. All life depends upon the cycling of matter (nutrients and water), as well as upon the flow of energy. Clearly, no organism can grow, propagate, and continue its kind for millions of generations without replenishing the elements that support it. Like birth and growth, death and decay are rules of the living landscape. Even the purely physical world has its cyclic processes of evaporation and condensation of water and the rise and erosion of rocks. In contrast to the unidirectional flow of energy through the ecosystem, whereby sunlight is absorbed by plants and heat is emitted to space at every conversion of energy from the eaten to the eater, any living and most nonliving entities emerge from the surface of the Earth only to return to their original point of origin in one form or another. Such a movement is called a biogeochemical cycle, in reference to the biological and geologic phases of chemical substances, impelled by the Sun's energy (see Figure 3).

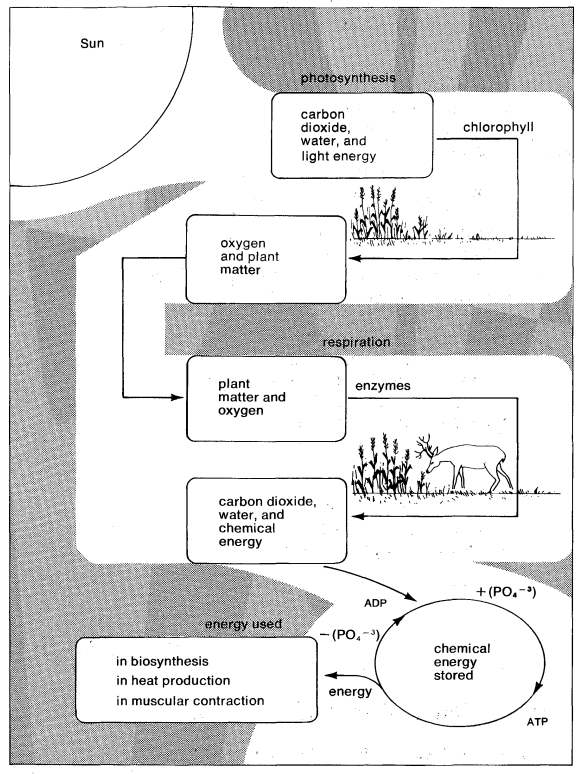


Figure 2: Energy relationships in the biosphere.
From *Biological Science: An Inquiry into Life*, 2nd ed. (1968); Harcourt Brace Jovanovich, Inc., New York; by permission of the Biological Sciences Curriculum Study

Elementary composition of organisms

All organisms contain water and utilize carbon, hydrogen, nitrogen, oxygen, phosphorus, and sulfur in order to form carbohydrates, proteins, fats, lipids, and other structural materials. Most elements are obtained by plants through molecules such as water, H_2O; and carbon dioxide, CO_2; and compounds such as nitrate, NO_3^-; ammonia, NH_3; sulfate, SO_4^{2-}; and hydrogen sulfide, H_2S. Oxygen is taken in by animals through respiration, bound up in organic products, and converted to water, in which form it is eventually returned to the environment. Plants utilize soil and water and, through photosynthesis, return oxygen to the air. Phosphorus, which is so desperately needed by organisms that its abundance or lack often limits populations, does not cycle easily in the ecosystem. Eutrophication is a condition of lakes that results from phosphorus enrichment of the water. Much phosphorus is lost from soils by erosion, and the sediment deposits in the bottom of the ocean keep phosphorus out of circulation for extended periods. Nitrogen, the most abundant of atmospheric constituents, is a relatively inert gas that most organisms cannot use directly but that is essential to life. Fortunately, certain species of bacteria and some blue-green algae are able to utilize gaseous nitrogen that diffuses into the soil from the atmosphere. These soil organisms convert nitrogen (N_2) into ammonia (NH_3). Some plants utilize the ammonia directly, but most depend upon the oxidation of ammonia to nitrite (NO_2^-) and then to nitrate (NO_3^-) by other soil bacteria before it is absorbed by plant roots. This process is known as nitrification.

Mineral elements such as calcium, sulfur, magnesium, potassium, and boron are taken up in solution from soils by the roots of plants and are returned to the soil by decomposers, organisms that promote breakdown and decay of the dead litter of forests or fields. Although rainwater slowly washes out, or leaches, some minerals from an ecosystem, it also brings in with it some nutrients during the precipitation process. As rainwater runs down mountains and hills it gradually picks up nutrients and makes them available to green valleys below. When humans cut the forest and expose the soils to sun and rain, they accelerate the leaching process, often with disastrous consequences. In order to maintain high agricultural productivity it becomes necessary to apply increasing quantities of synthetic fertilizers to replace those nutrients washed away.

The carbon and oxygen cycles. The total biosphere contains approximately 20 quadrillion (2×10^{16}) tons of carbon, mostly in the form of inorganic carbonates in the rocks and oceans, and in organic fossil fuel deposits such as coal, oil, and natural gas. The atmosphere contains 7×10^{11} tons of carbon in the form of carbon dioxide, and the green plants of the world contain 4.5×10^{11} tons as carbohydrates and other organic compounds. The exchange of carbon dioxide with the atmosphere by means of photosynthesis and respiration in plants results in a net annual productivity of the land of about 2.5×10^{10} tons per year and of the oceans of about 2×10^{10} tons per year. During the daytime, photosynthesis often produces a 12 percent drop in atmospheric carbon dioxide in the vicinity of plants, but at night, respiration by soil bacteria, plants, and animals often produces a 25 percent increase in carbon dioxide concentration near the ground.

Fossil fuels combustion, deforestation, and decay accounted for a CO_2 atmospheric concentration of about 340 parts per million in the late 20th century. Much of this output may cycle into the oceans, and it is possible that a considerable fraction is bound up in new plant growth. But the increasing concentration of atmospheric carbon dioxide appears to be sufficient to accelerate the so-called greenhouse effect, the means by which the atmosphere retains sun-heat. This acceleration poses the threat of global warming, which would disturb weather patterns and diminish the polar ice caps, thus causing a potentially devastating rise in the sea level. These possibilities have made the greenhouse effect and its causes major public policy concerns in the late 20th century.

Greenhouse effect

Carbon dioxide is relatively soluble in water. It is not surprising, therefore, that the oceans play a significant role in the global distribution of carbon dioxide and that rainfall sometimes contains about 0.3 cubic centimetre of CO_2 per litre of water. Carbon dioxide combines with water to form carbonic acid (H_2CO_3), which dissociates into hydrogen ions (H^+) and bicarbonate ions (HCO_3^-); the bicarbonate ions, in turn, dissociate into hydrogen and carbonate ions (CO_3^{2-}). These reactions can be expressed:

$$CO_2 + H_2O \rightleftharpoons H_2CO_3 \rightleftharpoons H^+ + HCO_3^- \rightleftharpoons 2H^+ + CO_3^{2-}.$$

These reactions are readily reversible, the direction depending upon the concentration of the components. The amount of carbon present as bicarbonate or carbonate depends upon the alkalinity or acidity (pH) of the water. If the water is alkaline, more carbon is present as carbonate than if the water is acid. High amounts of dissolved carbon dioxide in water produce high plant productivity (*e.g.*, algal blooms), but often this results in higher respiration, depleted oxygen, and subsequent fish kills. Acidic waters usually have low productivity.

Another gas of enormous significance to life is oxygen, which is cycled between the lithosphere, the atmosphere, and the biosphere. Plants are primarily responsible for the presence of atmospheric oxygen through the photosynthetic process. Oxygen in metabolism and in the production of energy-rich phosphorus bonds provides the power for all higher forms of life. Although oxygen is utilized within cell constituents such as mitochondria for the release of energy and the synthesis of ATP, other cellular bodies called peroxisomes appear to protect the cell from too much oxygen, which would result in destruction of the cell. Hence, in the oxygen cycle some organisms utilize the gas, some must be protected against it, and some generate it, all at the same time.

The nitrogen cycle. The atmosphere contains nearly 80 percent nitrogen. Ironically, most green plants are unable to use free nitrogen and must have it converted to soluble compounds of nitrogen, such as ammonia (NH_3), nitrite (NO_2), or nitrate (NO_3), which can then be taken up by their roots and the nitrogen converted into amino acids and plant proteins. The Earth's primitive atmosphere apparently contained ammonia, so the necessity

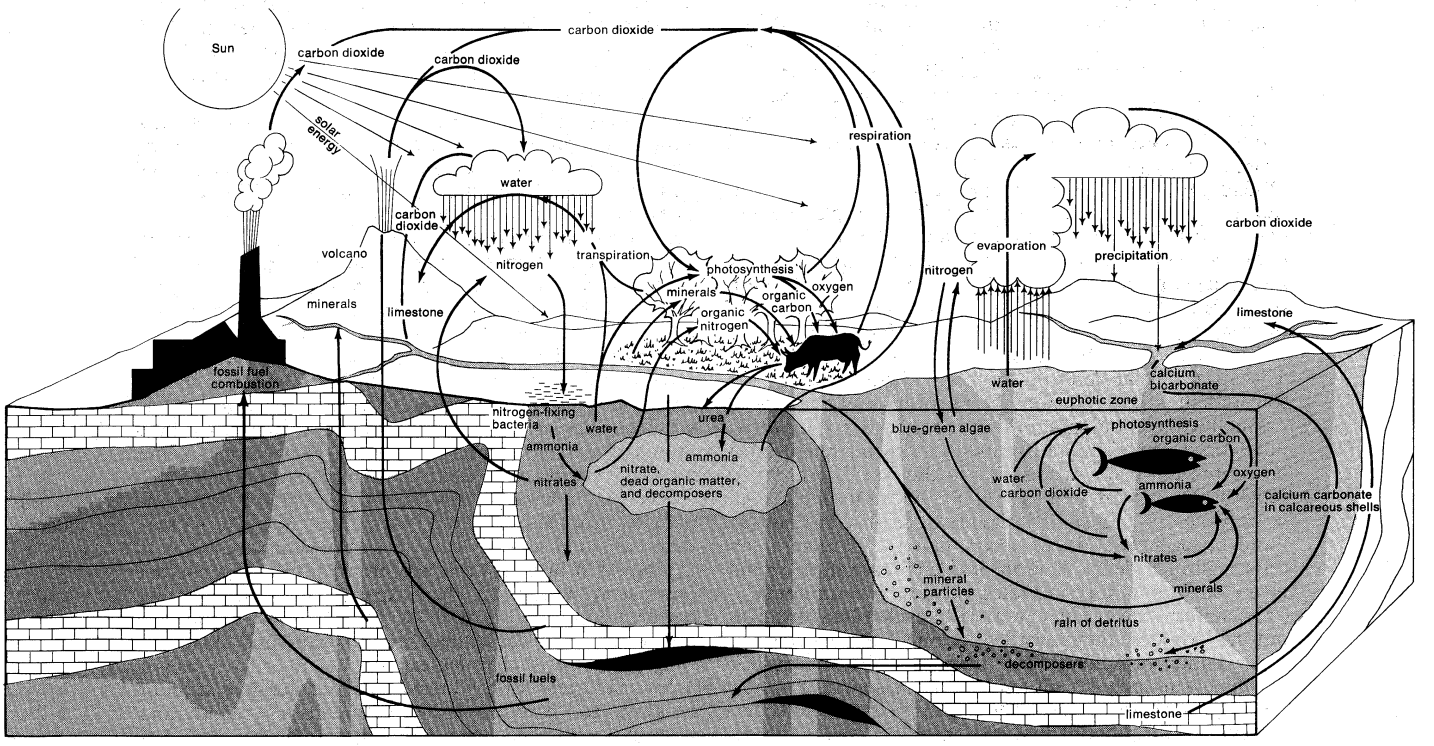


Figure 3: Major cycles of the biosphere.

for conversion of nitrogen into soluble products (nitrogen fixation) did not arise until more recently. Nitrogen fixation, or nitrification, is performed by a few species of microorganisms. The reverse process—by which soluble compounds of nitrogen are reduced to molecular nitrogen (N_2)—is called denitrification and is accomplished by other microorganisms. Microorganisms that decompose the remains of dead plants and animals reduce amino acids containing nitrogen to ammonium ions and other products. This process is known as ammonification. The rate at which nitrogen fixation removes nitrogen from the atmosphere is almost balanced by the rate at which denitrification returns nitrogen to the atmosphere. (The large-scale and widespread use of fertilizers by humans may be upsetting this balance, however.)

The indispensability of microorganisms in the movement of nitrogen

Nitrogen is present in proteins and nucleic acids. It is a versatile element in living processes because of its ability to form many different kinds of compounds and to release energy when moving from one compound to another. Nitrogen is cycled in nature from reduced inorganic compounds to oxidized compounds by atmospheric oxygen, with the release of energy; when oxygen is unavailable, oxidized nitrogen compounds can in turn oxidize organic compounds. Two main types of bacteria and algae participate in nitrogen fixation. One lives in a mutualistic partnership (symbiosis) with higher plants; the other is a free-living form that derives energy directly from sunlight and indirectly from plant materials. Rhizobia, the most abundant of the root-nodule symbiotic bacteria, are found on the roots of legumes, alders, and buckthorns. Legumes, which include peas, soybeans, and alfalfa, are often used in agriculture to increase the productivity of the land. Symbiotic nitrogen fixers normally contain certain trace elements such as molybdenum and cobalt. When these are absent, the bacteria do not function effectively, and the rate of nitrogen fixation is reduced.

Symbiotic microorganisms occur only in terrestrial ecosystems as far as is known. Among the nonsymbiotic nitrogen fixers are aerobic soil bacteria such as azotobacters, which supply fixed nitrogen in grasslands and in marine and other ecosystems where symbiotic microorganisms are absent. Blue-green algae (cyanophyceans) are an important source of fixed nitrogen in rice paddies and in aquatic systems. One free-living nitrogen-fixing bacterial genus that lives only without oxygen *(i.e.,* anaerobically) is *Clostridium.*

Autotrophs

A group of microorganisms known as autotrophs functions without an organic source of energy by oxidizing various forms of soluble nitrogen compounds. Nitrosomonads oxidize ammonium to nitrite with the release of water and energy. Nitrobacters, in turn, oxidize nitrites to nitrates with the release of energy. Nitrates are highly soluble in water and are easily taken up by roots and assimilated by plants.

Pseudomonads and some fungi use nitrate as an oxygen source; they draw on the energy available in glucose and phosphate to convert the nitrate to nitrous oxide and nitric oxide. Some denitrifiers reverse the process only partway by reducing nitrate to nitrite or ammonia. Denitrification proceeds best under anaerobic conditions, since whenever oxygen is present it is more efficient for organisms to use it than to use the oxygen bound up in nitrate ions. This suggests that denitrification would occur best in the soil and in deeper portions of the waters of the world. From the enormous amount of nitrogen fixed annually, however, about 92 million tons, it is obvious there must be enormous anaerobic reserves in the world in order to return this amount of nitrogen to the atmosphere each year. It is entirely likely that the use of fertilizers is causing more nitrogen fixation to occur than the biosphere can return through denitrification, in which case there would be a gradual buildup of nitrates, nitrites, and ammonia.

The sulfur cycle. For materials to cycle easily throughout the biosphere they must be not only water-soluble but also volatile. In addition to carbon and nitrogen, sulfur is a highly mobile constituent of the biosphere. Furthermore, all proteins incorporate sulfur in their molecular structure to form bonds that give them well-defined three-dimensional shapes. Just as carbon and nitrogen are reduced during the formation of proteins, so also is sulfur. In distinction to carbon, however, which requires green plants and sunlight for reduction, nitrogen and sulfur reductions are most often accomplished anaerobically by microorganisms. Most of this activity occurs in oxygen-deficient soils, bogs, and swamps.

When an organism dies and decays, much of its incorporated sulfur is returned to a mineralized state by bacteria and fungi, but some is reduced under anaerobic conditions directly to sulfides, such as the evil-smelling hydrogen sulfide (H_2S), some of which escapes to the atmosphere. Burning of fossil fuels sends massive quantities of sulfur dioxide into the atmosphere as a pollutant. When the sulfur dioxide combines with water in rainfall it forms dilute sulfuric acid (H_2SO_4), a form of acid rain. Inorganic sulfate (SO_4), formed by the decomposition of organic matter, is readily soluble in water and serves as a further source of sulfur to plants and animals. Sulfate is reduced under anaerobic conditions to elemental sulfur or to sulfides by bacteria. Large quantities of hydrogen sulfide occur in the anaerobic (deeper) portions of aquatic ecosystems. The anaerobic bacteria utilize the sulfate in metabolic oxidation much as bacteria denitrify nitrate and nitrite.

Acid rain

There are sulfur-utilizing bacteria that play the same role with sulfur compounds as nitrifying bacteria do with nitrogen compounds. These include the green and purple photosynthetic bacteria; the green bacteria oxidize sulfide to sulfur and the purple bacteria generate sulfate.

Sulfur precipitates in lake waters as sulfides in the presence of iron under anaerobic conditions. Some sulfide is insolubly bound with iron in the bottom muds of lakes, bogs, swamps, etc., where other elemental metals, such as copper, cobalt, cadmium, and zinc, also become bound as precipitates.

The water cycle. Water is a ubiquitous and unique substance necessary for life on Earth. Water in liquid and vapour states in the atmosphere regulates and ameliorates the climates of the Earth. Strictly speaking, water conforms to a gaseous cycle, but because of its great importance in the biosphere it is here discussed separately. Life began in the oceans, evolved in the waters of the world, and spread upon the land; yet life has always remained dependent upon the availability of water. The relative distribution and availability of water on the land determines the vegetative character of the landscape. Water erodes and sculpts the rocky surface of the Earth, transports nutrients and sediments, and forms the lakes, swamps, rivers, and seas. Sun-heat evaporates water from land and sea into the atmosphere, where it is transported with the global circulation of air and precipitated onto the surface as rain or snow.

The topography of continents and islands affects very strongly the precipitation pattern of the Earth. Windward mountain slopes are wet, as warm moist air rising cools and condenses its moisture as rain or snow. The leeward sides of mountains are dry, since the moisture has already been wrung from the air on the other side of the divide, and the air moving down the leeward face warms, thereby retaining whatever moisture it still has. Semiarid regions with sparse vegetation often occur to the leeside of mountain ranges. Cold polar ice caps have relatively low precipitation since very cold air can contain little moisture. Deserts generally occur where the air is stable. The trade winds move toward the equator from cooler latitudes, picking up moisture and heat as they go; hence the coastlines of southern California, Mexico, and Chile are relatively dry, and the equatorial regions are wet.

Properties of water

Water has amazing properties, particularly when compared with most other forms of matter known in nature, that make it chemically and physically suitable for life. It is a liquid at ordinary temperatures, contracts on cooling down to 4° C (39° F), then expands on further cooling to the freezing point at 0° C (32° F). Solid water, or ice, is thus less dense than liquid water and floats on water. The fact that water expands on freezing makes it a powerful agent for the breakdown and fragmentation of rock into soil particles. Water warms up less rapidly and in turn cools more slowly than other substances. Lakes and oceans, therefore, have a different temperature from

the adjoining landmasses, with a seasonal lag. Water tends to hold dissolved material in solution and also has the greatest surface tension of any known liquid. Moist air is less dense than dry air and rises above it, contributing to the weather dynamics of the atmosphere.

Taken over the world as a whole, the horizontal distribution of moisture must always add up to zero—the amount of water falling as precipitation is equal to the amount of water vapour taken up by evaporation. For any single region of the world, however, there may be a water surplus or a water deficit, differences that are theoretically balanced by ocean currents or river runoff. The Northern Hemisphere, for example, has an excess precipitation over the Southern Hemisphere, and the difference is redistributed by means of ocean currents. The drier the continent, the smaller the fraction of the annual precipitation that runs off. Australia and Africa, as a whole, are relatively dry continents in which large areas are covered by deserts and more than 75 percent of the annual precipitation is lost through evaporation.

In order for evaporation to occur, two conditions must exist. There must be energy available, which is derived largely from sunlight, and there must be a water vapour gradient from a moist surface to air that is less moist. In other words, heat and dryness aid evaporation.

Interrupting this cycle of simple precipitation and evaporation are the organisms of the world, which divert water for their own use. Water relationships are most dramatic in land situations. Soils hold considerable amounts of water. On the average, worldwide, this may be a reserve of about 10 centimetres (4 inches) in depth. In some parts of the world, such as North Carolina, where soils are thin, the total amount of water held in the soil may not exceed 5 centimetres (2 inches), but in the deep volcanic soils of east Africa as much as 50 centimetres (20 inches) is held in the soil.

Plants take up water largely through their roots and evaporate water through the leaves in a process known as transpiration. A green plant may transpire 3 to 5 millimetres (about 1/8 to 1/5 inch) of water a day, depending upon the amount of energy available. Maximum rates in sunny, hot, irrigated regions may reach 8 millimetres (1/3 inch) a day.

Transpiration ratio

An important ratio for plants, known as the transpiration ratio, is the amount of water used to the weight of accumulation of dry matter in the plants. For corn (maize) it is 317 grams of water per gram of dry weight, for cotton 568, deciduous trees 825, and evergreen trees 140. Taken over a whole growing season, although it is highly variable among crops, one can estimate that a production of 20 fresh-weight tons of crop will require 2,000 tons of water from the soil. Of the 20 tons of crop only about three tons consist of water molecules utilized in photosynthesis, the remainder being evaporated as transpiration.

In addition to their physiological need for water, large numbers of terrestrial animals use bodies of water as sanctuaries. The eggs of many insects are laid in water, larvae live in water, and the adults emerge from ponds or rivers to spend their adult lives on the land. Amphibians occupy a narrow zone between land and water, in which they can move from the harsher, more variable climate of land into the milder, more comfortable conditions of water. All land animals give off moisture when they exhale, and some animals pant or sweat to evaporate moisture and thus cool themselves.

The sedimentary cycles of essential minerals. While many elements give rise to gaseous compounds that are significant parts of biogeochemical cycles, some elements in the biosphere cycle primarily through water transport and sedimentation in bodies of water. For any soluble but nonvolatile compound a natural cycle is only possible when life is involved, since otherwise the compounds wash from the land to the rivers and oceans, where they remain in the sediments. Green plants pick up the compounds in the nutrient-rich soil water and convert them into plant cells where they may be passed to animals in the food chain; they eventually return to the soil through death and decay of plants and animals. Some compounds soluble in water are carried into the atmosphere by water evaporation and returned to the surface in rain.

The elements calcium, potassium, silicon, and magnesium each have important biological roles in the biosphere. Magnesium, for instance, is an essential element in the chlorophyll molecule, and calcium and silicon help form the hard parts of shells, bones, and teeth of animals. Iron, manganese, and sodium are present in organisms in minute, or trace, amounts, but nevertheless are vitally important. As mentioned earlier, phosphorus is probably the most important element among those that form nonvolatile compounds. A deficiency of phosphorus is most often responsible for poor crop production. Free phosphorus as such is not found in the atmosphere. If it were not for green plants absorbing its salts from the soil and transporting it to the leaves, phosphorus would not rise above the surface of the ground. So it is also for trace elements such as iron and manganese. Some trace elements such as vanadium, cobalt, nickel, and molybdenum are found primarily in aquatic plants, since they accumulate in bottom sediments.

The problem of phosphate enrichment

Phosphorus is soluble in acidic waters; under special environmental conditions it is bound up as calcium phosphate or iron (ferric) phosphate. Phosphorus is extremely scarce in the lithosphere, concentration normally being 1,050 parts per million. But great supplies of phosphorus are found in bird guano, the excrement of fish-eating gulls, cormorants, pelicans, and penguins, on the islands and in the ocean sediments off the coast of Peru. Artificial, or chemical, fertilizers are made from phosphate rocks and marine phosphates. Great quantities of these phosphates are used in detergents and wash into lakes and streams. Such phosphate enrichment creates rapid and excessive growth of algae, especially the blue-green algae. Increased decay and respiration result in oxygen depletion of the water and the suffocation of more sensitive species of fish, usually the game fish. This enrichment process, called eutrophication, usually results in a simpler animal and plant community, a shortening of the food web, and a less stable ecosystem. (D.M.G./Ed.)

The distribution of organisms

The subject of distribution embraces the individual organism as a unit of study as well as the highest taxonomic unit. As far as scale is concerned, the range is equally great, from a drop of water to the largest ocean. Not surprisingly, the study of distribution encompasses several overlapping disciplines. Ecology is concerned with the numbers and distribution of organisms on a local scale, and biogeography is concerned with distribution on a regional or even worldwide scale. Other disciplines contribute much to the study of distribution and are in turn helped by it. The two most prominent of these are systematics and the study of evolution. Systematics, the study of the relationships between organisms, in the widest sense embraces classifying and naming organisms, information essential to biology in general and to the study of distribution in particular. Systematists, in turn, use distribution information in reaching decisions on the relationships between organisms. The study of evolution has benefited considerably from distribution information, as the best-known pioneers in the field, Charles Darwin and Alfred Russel Wallace, freely acknowledged. In return, a knowledge of the processes of natural selection has made large contributions to the understanding of the reasons for changes in distribution with time.

PATTERNS AND PROCESSES OF DISTRIBUTION

It is convenient to distinguish two closely related modes of distribution that are, in fact, inseparable in nature: (1) the local arrangement—on a small scale—of members of a population to ensure their survival by maximizing the use of the environment, often in accordance with the time of day, week, month, or year, and (2) the large-scale geographic spread—on a regional, or even global, scale—of a species over a long span of time. Obviously, the distribution of organisms on a local scale eventually affects the geographic arrangement of the species.

Local distribution. Just as the body of an organism has

a structure, so does a population of organisms. The structure that results from the distributions of organisms is termed a pattern. Patterns may be characterized either by the forces that produced them (physical, social, or other) or, as in this discussion, by the ability of the organisms themselves to adjust their distribution.

Spatial and temporal arrangements. In a homogeneous environment—one that varies little from place to place—all parts of the environment may be occupied or visited by organisms. Protozoans may be distributed throughout a small pool of water, and all available space on a rock in the intertidal zone of a shore may be covered by mussels or barnacles. In a heterogeneous environment, the total volume occupied or visited by a population may contain parts that are not exploited. In the pelagic, or open water, realm of an ocean, minute plantlike organisms may occur patchily, and the animals that feed upon them may be distributed in the same manner. These extremes of distribution are difficult to characterize because of the problems of measurement and representation in three dimensions. If the third dimension is not occupied, or unimportant enough to ignore, then distribution in a single plane, such as a flat piece of ground, is easy to characterize.

The three patterns of dispersion

There are three patterns of internal distribution (or dispersion) of a population in a single plane (Figure 4). Organisms are distributed at random when individuals are free to choose a place to settle. But individuals often influence each other. When they attract each other, an aggregation, or clump, tends to form; when they repel each other, an overdispersion, or uniform spacing of the individuals, results. In the soil or in the leaf litter of a forest floor, almost all arthropod species are distributed in a clumped or aggregated fashion, with only a few species randomly distributed. Territorial animals, such as birds, and some plants, such as certain cacti, repel each other at a certain distance and thereby assume a uniform pattern of distribution.

The importance of scale is illustrated in Figure 4. It shows that, on a small scale, members of a population are uniformly distributed, but on a large scale they are aggregated in clumps. This happens, for example, when many flocks of birds occur in a large area but retain their discreteness. A certain minimum distance is maintained between individuals in each flock by mutual avoidance, yielding uniform spacing within the flock. But the flocks themselves may be randomly distributed or clumped.

For many species, distribution is a function of time. The members of a population make regular movements, so that their distribution is different from one hour to the next; one day, week, or month to the next. Short-term shifts in distribution are best exemplified by plankton, small, often microscopic, organisms that float in the surface waters of lakes and seas. They move vertically in response to daily changes in light intensity, so that they are at greatest depth at midday and are at the surface or close to it at dawn and dusk.

More striking, perhaps, are the regular annual movements of some animals, which bring them to one region at the time of breeding and to another region in the nonreproductive season. Such movements are known as migrations (see below *Biological populations: Changes in population characteristics: Movements*).

Environmental influences. Organisms cannot exist outside a certain range of environmental conditions. There is an upper temperature limit to the activities of protoplasm, beyond which proteins are denatured, and a lower limit, which for many organisms corresponds with the temperature at which water freezes. Many organisms, however, are capable of existing, at least for a short time, under inclement conditions. They can survive long enough to move out of the area (by migrating) or to become physiologically inactive and, thus, protected against adverse conditions (by hibernating, estivating, or encysting). Some "cold-blooded" animals have the ability to adjust or acclimate physiologically by altering their metabolic activities in such a way as to operate best under the prevailing temperatures. Physiological adjustments of this sort are possible only when the environmental conditions change relatively slowly and by a relatively small amount.

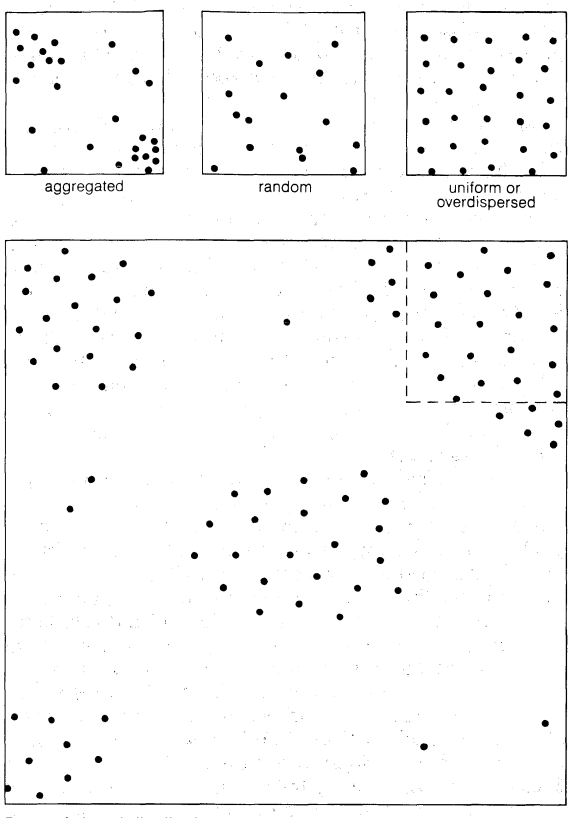


Figure 4: *Local distribution of organisms.*
(Top) The three basic dispersion patterns. (Bottom) The relativity of dispersion to scale. On a local scale (inset) the individuals are uniformly distributed (overdispersed), but, on a larger scale, they are aggregated.

Within this framework, many processes operate in the establishment of a pattern of distribution of a species on a local scale. Perhaps the easiest to envisage is a distribution determined largely by the characteristics of reproduction. In many plants, and some lower animals, the reproductive products (seeds, spores, etc.) remain close to the parents because their dispersal capabilities are poorly developed, thus creating an aggregated distribution. In contrast are the intertidal organisms, many of which, although themselves fixed firmly to the substrate, shed their reproductive products into the water at high tide. The larvae develop in the pelagic realm of the sea and are carried to shore at the time they are metamorphosing to the adult stage.

Directional versus random forces

The distribution pattern of adults that results from this combination of dispersal and settlement is determined by two kinds of processes, which are termed vectorial and stochastic. The vectorial process refers to directional dispersal caused by environmental motion, as in wind and water currents. The stochastic process refers to essentially random forces whose operation does not allow a prediction to be made as to where organisms will be carried and will settle.

As was mentioned earlier, the pattern of distribution may be determined also by the influence of one organism upon another. The social process that gives rise to the pattern is one of signaling. The messages are either "come here," leading to aggregation, or "go away," leading to uniform spacing. Alternatively, these two patterns, as well as a random one, may be produced by each organism responding individually to environmental features such as nutrients, which are themselves distributed in random, clumped, or uniform fashion.

Interactions with other organisms. Finally, the distribution of organisms of a species may be determined by interactive processes with other species. Heavy grazing by animals in a certain area often leads to a sparseness of plant species in the grazed area. Similarly, competitor species may influence each other, the one causing the other to be either rare or absent in one part of a local area. Throughout Europe, North Africa, and some of

the Canary Islands, for example, the chaffinch (*Fringilla coelebs*) occurs in both deciduous and coniferous forests, but on two of the Canary Islands, Tenerife and Gran Canaria, it occurs only in deciduous forests, a closely related species, the blue chaffinch (*Fringilla teydea*), occupying the coniferous areas. The distribution of the chaffinch on Tenerife and Gran Canaria appears to be influenced, in a restrictive way, by the blue chaffinch, with which it could compete for space and food. This type of influence, while not easy to recognize, is probably exercised commonly between species that live in the same habitat as well (see also below *Biotic interactions*).

Differences in the way in which the three basic patterns of distribution—random, clumped, or uniformly spaced—are set up have been emphasized, but a single mechanism could be responsible for any one. Consider plants, for example. If the spatial pattern of germinating seeds is random, if they thrive only in space not occupied by other root systems, and if they grow to a size that depends upon the space available, then the resulting distribution will depend solely upon the density, or numbers of seeds. Going from low to medium to high density, the distribution changes from clumped to random to uniformly spaced. Thus, the factors that govern the abundance of a species within a local area indirectly determine the distribution of the species in that area. It is possible that the factors that govern the abundance of a species also limit the distribution of that species. The processes leading to the establishment of a distribution range (area or volume), as opposed to the pattern of distribution within that range, are discussed in the next section.

Large-scale distribution. More is known about distribution on a large scale than on a local scale. The reasons are partly practical and partly historical; it is easier to collect a few isolated specimens of plants or animals over a wide area than to study in detail their distribution in a clearly circumscribed area such as a field or a grove. In the past taxonomists and students of evolution have spent much energy trying to determine the extent of geographic variation in morphological characters of organisms, for which purpose collected specimens are sufficient.

The shape and extent of spread. Most species occupy areas of irregular shape. The irregularities are often determined by obvious environmental features. The irregular pattern of distribution in North America of the pika (*Ochotona princeps*), a small rabbitlike animal, for example, corresponds with the irregular pattern of distribution of its habitat—rock slides and talus slopes at high altitude in the Rocky Mountains. When the distribution of certain habitat features is essentially linear, so too is the distribution of the species; the emergent vegetation at the edge of a lake and the intertidal mollusks in a narrow band along a rocky seacoast are cases in point.

Continuous versus discontinuous distribution

Viewed on a local scale, the distribution of organisms may be continuous, or without interruption, but when viewed on an intermediate regional scale, the distribution may be discontinuous, or spotty, because of unsuitable habitat intervening between areas of suitable habitat. Viewed on a large scale, a continental or worldwide scale, the distribution may appear continuous once again, as the narrow bands of unsuitable habitat that produce the discontinuities slip out of focus.

Discontinuities on a worldwide scale are evident for some species. Horseshoe crabs (family Limulidae) are found on the Atlantic coast of North America from the state of Maine to Mexico. These crabs are not found on the west coast of that continent, however, but on Asian coasts, from India to Japan, thousands of kilometres away.

Few species—of which man is one—have a worldwide distribution. Of those that do, several are associated closely with man and have achieved their present distribution as a result of this relationship. Nevertheless, there are cosmopolitan species that have achieved worldwide distribution by their own efforts. Such are the short-eared owl (*Asio flammeus*) and the osprey (*Pandion haliaetus*), two birds, both predators, conspicuous on all continents excluding Antarctica.

Restricted distribution

It is much more common to find species with restricted distributions. In addition to the numerous species confined to single, small islands, there are continental species found in extremely small areas on large landmasses. The dusky seaside sparrow (*Ammospiza nigrescens*), for example, occurs only in the salt marshes around Merritt Island, Fla., and on the adjacent peninsula.

Another form of restriction is indicated by the distribution of small carrion (*Catops*) beetles (Figure 5): restriction to a climatic zone, in this case the north temperate zone. The booby (*Sula leucogaster*), a seabird, has a pantropical distribution, breeding only—but extensively—within the tropical zones. Two closely related species of the marine wormlike priapulids are restricted to the polar regions, one in the Arctic, the other in the Antarctic.

Adapted from Jeannel, *La Genese des faunes terrestres* (1942); Presses Universitaires de France, Paris

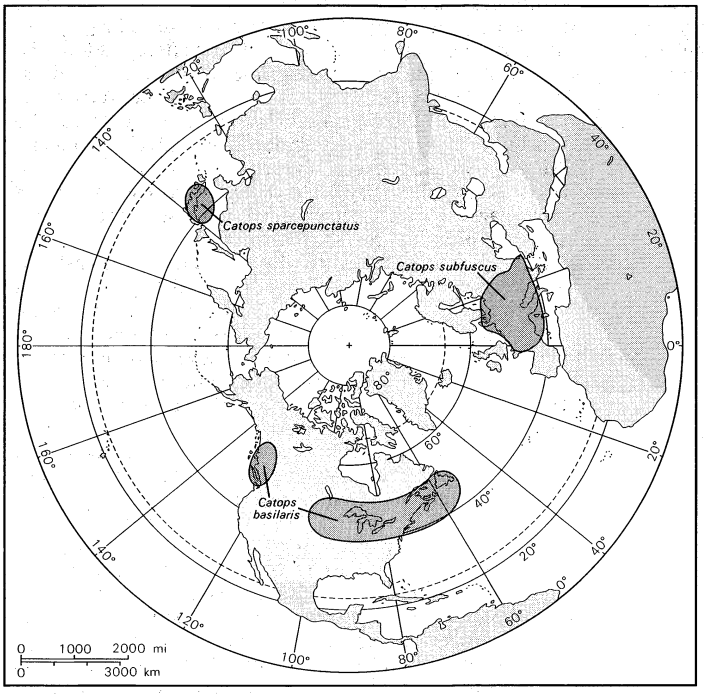


Figure 5: Discontinuous distribution of three beetles of the genus *Catops* in the north temperate zone.

In all these instances, and in their variations, the restriction is the result of the inability of the organisms to exist beyond a range of certain environmental conditions. This is contrasted with restriction caused by the inability of organisms to travel over barriers, which has led to regional differentiation of faunas and floras.

The factors limiting spread. Although more is known about the gross distribution of species than about their distribution within local areas, little has been established with certainty concerning the determination of the limits of an area occupied. In some cases the reasons for limitation of geographic range are obvious: the distribution of terrestrial organisms stops at the edge of sea, lake, or river; freshwater organisms occupy the length of a river, but extend no farther than its estuary. But within the major environments the reasons for the observed distributions are not always so easy to discern.

At the outer edge of a distribution range the continued existence of organisms depends upon either reproduction or immigration from more central regions. The requirements of the individual for existence and reproduction must be met. One obvious way in which the distribution is limited is by insufficiency of the supply, or rate of supply, of these requirements. Any number of environmental factors, acting singly or in combination, can act in this limiting way.

The northernmost limit of distribution of the wood frog (*Rana sylvatica*), in North America, is almost certainly set by a combination of temperature and time. Complete development of the tadpole stage in one summer is the rule. At their northern limit, these frogs have just enough time in the short Arctic summer, at the prevailing low temperatures, to complete development. Temperature has

also been implicated as a factor in limiting the southern distribution of introduced pheasants (*Phasianus*) in North America. If the environmental temperature is high before incubation of the eggs, the eggs will not hatch. The southern limit to the distribution of the pheasant in North America corresponds approximately with those spring temperatures that are critically high.

In Scandinavia, the distribution of a carabid beetle (*Bembidion aeneum*) corresponds well with the distribution of soil that was lightly salted when the land was covered by the sea about 9,000 years ago. The beetle appears to depend upon the salt, either directly or, perhaps more likely, indirectly, through plants or other animals that require the salt and that themselves serve as food for the beetle.

Relationship between plant and animal distribution

In general, gross climatic and soil conditions determine the distribution of plants, singly and as assemblages; animals, being dependent upon plants for energy, follow the plant distribution. There are some highly specialized associations between plants and animals, as between some pollinators and the plants they pollinate; the edges of the distributions of the two interdependent species coincide.

Many animal species, however, are not so much limited in distribution by a single plant species as by an assemblage of plant species, such as those constituting a savannah or a tidal marsh, for example.

Factors limiting the distribution of a species are easier to recognize when they are discrete rather than graded, as is temperature. The side-blotched lizard (*Uta stansburiana*), for example, which occurs in the deserts of southwestern North America, breeds only in the abandoned nests of the pack rat (*Neotoma micropus*); the number of such nests thus sets an upper limit to the number and the distribution of breeding pairs of lizards.

The Australian ecologists H.G. Andrewartha and L.C. Birch have suggested that the factor or factors that limit the abundance of an animal species within its distribution range also limit the range itself. Exceptions to this are easy to find, particularly where the limit of distribution is set, at least in part, by a physical barrier, such as a mountain range. Nevertheless, the statement is likely to be widely true, especially for arthropods, and it is easy to appreciate when the species fluctuates in numbers markedly from one year to the next. A fine example is provided by the cutworm moth (*Porosagrotis orthogonia*), a pest species in North America, whose larvae feed on the roots of wheat. In Montana, for instance, in years of great abundance the species occurs throughout most of the state, but in years of scarcity it is either rare or absent from some parts.

Not only are distributional limits at one point set by an interaction of factors, but different boundaries are set by different factors. The saguaro cactus (*Cereus giganteus*) occurs in the Sonoran Desert of the United States and Mexico, but is not found either at the highest or lowest altitude within its area of distribution. The upper limit appears to be set by low winter temperatures, and the lower limit is set by a variety of factors, most important among which seems to be the fineness of the soil; this in turn probably affects the plants in a complex way through its properties of stability and of retention of nutrients and water.

One species can prevent the spread of another by virtue of competitive superiority in the range that it occupies, thereby setting the limit to the distribution of the other species. The distribution of two species of western pocket gophers (*Thomomys*) in North America is representative. The northern pocket gopher (*T. talpoides*) occurs in northern states and the valley pocket gopher (*T. bottae*) in southern ones. An outline of their distributions shows that there is little overlap and that their boundaries are often complementary (Figure 6). No conspicuous environmental feature coincides with the highly distinctive boundaries, so it appears that the species have set them themselves by interacting.

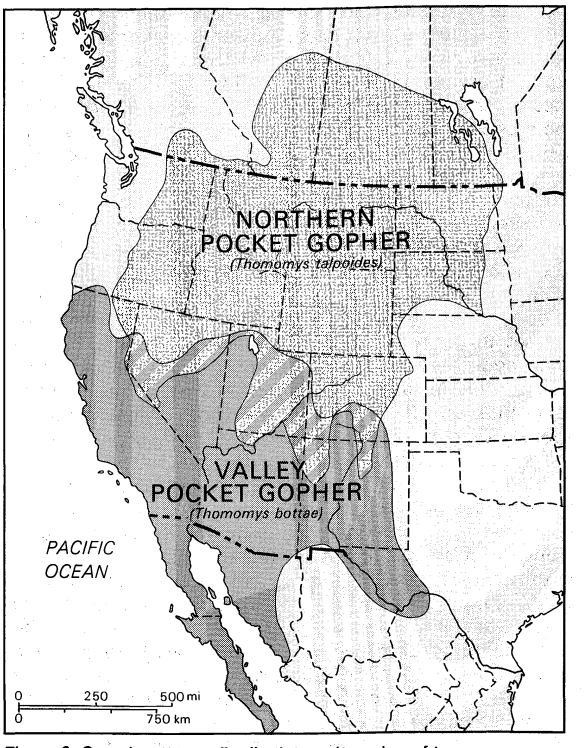


Figure 6: Complementary distribution and overlap of two species of pocket gopher in North America.
Adapted from W.H. Burt and R.P. Grossenheider, *A Field Guide to the Mammals* (1964); Houghton Mifflin Company, Boston

THE NATURE OF DISPERSAL

The requirements of an organism are a place to live, food, and, for sexually reproducing organisms, a mate. The requirement of space is not easily met in a densely populated area. Animals, of course, possess the ability to seek out unoccupied space, but for plants there is a strong element of chance in the finding of suitable space. In both, however, the dispersal characteristics of the organisms are of fundamental significance and, as such, are strongly under the influence of natural selection. These characteristics partly determine whether a species is widely or narrowly distributed and whether it can establish itself in isolated areas. Dispersal characteristics also influence the evolutionary process. Since the evolution of two populations from a single ancestral population usually requires spatial isolation, the dispersal frequency of members of the populations determines whether the degree of isolation is sufficient for evolution to take place.

The role of natural selection

Mechanisms of dispersal can be considered both active and passive. Some seeds are actively dispersed by mechanical means, as when a seed pod dries out, splits open, and sends a shower of seeds to the ground in the vicinity of the parent plant. Generally, however, plants rely on passive means of dispersing, whereas both passive and active means are used widely by animals.

Active dispersal. The conditions under which active dispersal occurs are not uniform throughout the life of organisms. Dispersal usually occurs at the time of reproduction, when the individual seeks a mate or, in the case of many vertebrates, seeks a place in which it can attract a potential mate. The first patch of suitable terrain or water is chosen by the dispersing individual, and this may be close to the place where it was born. If all the space in the vicinity of home is already occupied by other individuals, however, then that individual will be forced to disperse more widely, to increase its chances of securing a mate and food. It is these far-dispersing individuals that are likely to significantly alter the distribution range of the species. The rule is almost certainly that most of these individuals perish as they encounter unsuitable conditions beyond the limit of distribution of the species. Still, some reach suitable pockets of habitat beyond and somewhat isolated from the original distribution range, and occasionally these give rise to new populations.

The most spectacular of dispersal movements, however, are the so-called irruptions of birds. These movements are not the result of a search for a place to breed, but are induced by a combination of poor food supply and possibly other adverse environmental conditions coupled with

unusually high density at the end of the breeding season. An example of such an irruption is observed in Great Britain, which is beyond the breeding range of Pallas's sand grouse (*Syrrhaptes paradoxus*) but which has been visited irregularly by sizable numbers of this species. As far as vertebrates are concerned, it seems that all individuals in a population have an approximately equal ability to disperse. Whether they do disperse, and how far, is dependent upon density in relation to the required resources of food and water and upon the availability of mates.

Somewhat distinctive dispersal abilities are exhibited by members of many insect populations. Perhaps the best-known examples are found among locusts, ants, and termites. In central and eastern Africa the desert locust (*Schistocerca gregaria*) occasionally swarms and consumes all green plant material in its path. The species actually occurs in two forms of quite different behaviour. The bright green form is solitary and sluggish, while the darker form, the one observed in the plagues, is highly mobile and gregarious. If the young produced by the green form are raised at high density with frequent physical contact between individuals, they develop into adults of the dark form and not the green form. This phenomenon, known as phase polymorphism, is admirably suited to the species. As the population increases and depletes its food supply, a developmental and behavioral mechanism comes into play, resulting in the production of individuals that are predisposed to dispersing.

Polymorphism: forms to fit the need for dispersal

The caste polymorphism of ants and termites serves a similar function. In these highly social insects, in which different members of the society are structurally and behaviorally specialized to perform different tasks, dispersal to new areas and the founding of new colonies is the task of the winged forms.

A third form of polymorphism exhibited by insects is life-history polymorphism, as exhibited among butterflies, dragonflies, and wasps. In all of these insects the wingless larvae are the feeding phase and disperse little, whereas the adults are the dispersal phase and feed little. In fact, the adults of some insects, such as mayflies, do not feed at all, but disperse, mate, lay eggs, and die.

There are exceptions to the statement that larval insects disperse little, which introduces a fourth kind of polymorphism, termed behavioral polymorphism. In a population of the larch budmoth (*Zeiraphera griseana*) in Switzerland, there are at least two types of larva that differ in structure and behaviour. One is intolerant of crowding; it disperses readily and is usually the first to appear on the principal food plant, larch trees (*Larix*). The other is tolerant of crowding and disperses little. The proportion of these two types in the population varies from time to time in accordance with the overall density. A balance of advantages and disadvantages allows the two forms to persist together without one replacing the other.

As indicated by the above example, natural selection does not always favour the individuals that disperse the most. If the chances of dispersing and finding a new, suitable, and unexploited environment are very low, selection will act against the strong dispersers. Such seems to be the case in highly restricted and well-isolated environments such as oceanic islands and mountaintops. In these environments the number of flightless birds and insects is striking and is much greater than in continental regions. Charles Darwin pointed out that the possession of large wings and well-developed powers of flight could be a disadvantage to animals on oceanic islands, particularly small ones, because an occasional strong wind might carry an already airborne animal away from the island and out to sea. The zoogeographer P.J. Darlington has emphasized that wings are also used in flight from predators, but predators are scarce or absent on these islands, so the need for dispersal powers within the island environment is not as great as within continental environments.

Passive dispersal. Organisms are dispersed passively by three principal agents: wind, water, and other organisms.

Dispersal by wind and by water. The capability of wind to transport organisms is well known, but the extent and the scale of such transportation is only now becoming clear. Microscopic organisms, such as viruses and protozoa, are readily carried by the wind, but even some of the smaller vertebrate animals such as frogs can be carried along on strong winds. Spiders, mites, and insects have been collected on airplane flights across the Pacific Ocean at distances up to 3,000 kilometres from land. Because of the method of capture, specimens are dead, so there is no means of distinguishing whether they were or were not living when they were collected. Nevertheless, in view of these remarkable results, it is no longer surprising that even the most isolated islands have an extensive insect fauna. There are, for example, close to 4,000 insect species native to the geographically rather remote Hawaiian Islands. It has been suggested that these species evolved from 250 ancestral types that could have been windborne from continents.

In a similar way, winds carry insects from lowland regions to mountaintops, where they may be deposited in the snow. Those incapable of existing in this climatically rigorous environment (probably almost all) are eaten by the native arthropod fauna. In fact, the native arthropods may be dependent upon this wind-drifted food, just as some stream-dwelling insects are dependent for food upon the organic matter drifting down from upper levels of the stream.

A study of the spotted alfalfa aphid (*Therioaphis maculata*) in California indicates that an amazingly large number of organisms are passively transported by wind. Millions of these winged aphids alone float passively with the current in the spring months.

Structural adaptations facilitating dispersal by wind are strongly developed in some groups of spiders (class Arachnida). Gossamer, the silk secreted by the spider, is blown by the wind, and the spider, attached to it, is carried along, assisted sometimes by rising air currents.

Adaptations for dispersal

Water currents carry organisms in the same manner as wind. In the aquatic environment, however, the terrestrial organisms must stay afloat, avoid becoming saturated with water, and maintain their internal salt balance. Some outstanding examples of long-distance passive dispersal by ocean currents are known. Marine invertebrates are carried from West Africa to South and Central America on the main equatorial current of the Atlantic. Some water in this current comes from the Niger and Congo rivers, so it is possible for estuarine and freshwater animals to be transported to another continent by this means as well.

The distribution of several species of terrestrial animals is evidence that long-distance dispersal has been achieved by major ocean currents. One dramatic example is the distribution of a flightless staphylinid beetle (*Micralymma marinum*) believed to have been carried passively from the Western Hemisphere to the Eastern by means of the Gulf Stream.

The odds are considerable against terrestrial organisms surviving the ordeal of a prolonged journey in the surface waters of the sea. Some vertebrates may actively participate in the travel by swimming, thereby increasing their chances of survival. Others may cling to floating debris or natural rafts of land or ice. Pieces of river banks or fringing swamp vegetation and soil break off and are carried out to sea, particularly when flood conditions prevail, and with the rafts go the animals that were stranded on them. Pack ice, breaking loose, can similarly function as a raft for polar animals, such as polar bears, foxes, etc. A remarkable floating island was observed in the Atlantic off the coast of North America in 1892. It was estimated to be 30 metres square, with trees 9 metres high, and to have drifted at least 1,600 kilometres. Whether it reached land and its plant and animal passengers were able to colonize the new land is not known, but this example at least demonstrates the feasibility of distribution of organisms on rafts.

The distinction between airborne and waterborne means of dispersal is easy to make, but the two are not alternatives, because many organisms that start a journey on air complete it by water. Insects cross the Baltic Sea from Estonia in the Soviet Union to Finland, originally participating in mass flights, but reaching the other side in the water, many of them alive.

Dispersal by other organisms. The third means of dis-

persal—by other organisms—may be regular, as in the case of a host transporting a parasite, or irregular, as in the chance attachment of one organism to another. For parasites, finding a host is a major problem, and an efficient means of dispersal from one host to another, or to the environment in which another can be found, is essential. Consequently, parasites have developed remarkable specializations to the life history characteristics of the host that enable the parasite to disperse at the most favourable time. Dispersal is often assisted by the use of an additional species as a transporting agent, or carrier, from one host to another. A case in point is the myxoma virus, which parasitizes rabbits and uses mosquitoes as carriers. When aided by wind, an individual mosquito can carry the virus as far as 65 kilometres before infecting a rabbit; by this process of transporting and infecting, the mosquito population can disperse the virus rapidly over a large area.

Many organisms are distributed by attachment to mobile organisms. Small invertebrates such as snails and flatworms, for example, are accidentally transported on the legs of migrating birds. Undigested food items, such as the seeds of some plants, can be widely dispersed by their consumers in excrement deposited far from the source. There are numerous cases of mutualistic associations between animals and plants in which the plants are dispersed. One example is provided by flightless weevils of the genus *Gymnopholus,* which carry fungi, lichens, algae, or liverworts in depressions on their backs. From this association the plants gain a place to grow and a vehicle of dispersal. Plant dispersal is also carried out by mites that live for part of the time in the plant association growing in the weevils' backs. The advantage gained by the weevils is that of camouflage; to a predator they appear to be part of the vegetation-covered tree bark on which they occur (see also below *Biotic interactions*).

Introductions by humans

Probably the most effective agents of dispersal are humans, by virtue of their numbers, widespread distribution, and long-distance travel and shipping and of the objects they use. It is not surprising to find, for example, that ports of entry are the centres of distribution of some arthropod groups newly introduced to a continent. The arthropods, other small animals, and plant seeds are transported in boats with cargo. It used to be the custom for vessels to take on soil as ballast in England and unload it in Newfoundland, which resulted in the introduction to Newfoundland of a number of soil organisms not previously found there. These examples might be termed adventitious transport through the agency of humans. By contrast, there are also numerous instances of deliberate introduction of "exotic" species of plants and animals to new areas. In many cases, the newly introduced species perishes or simply does not spread from the point of introduction, as has happened with the skylark (*Alauda arvensis*), introduced from Great Britain to southern Vancouver Island, Canada. In other instances, the introduced species has exploited the new environment and spread rapidly. It took only 50 years for the Old World rabbit (*Oryctolagus cuniculus*), exported from Great Britain, to spread the length and breadth of Australia from a single farm, even in the face of efforts to prevent the spread.

The dispersal of plants depends largely on their exploitation of animals as carriers. Their structural specializations to being carried and dispersed exhibit great variety. These include a variety of hooks, barbs, bristles, and sticky secretions to aid attachment to animals; fleshy palatable fruit for consumption by animals (but enzyme-resistant seed walls); very small seeds light in weight but large in area to facilitate wind dispersal; and resistance to saltwater to facilitate dispersal by ocean currents.

The means of transport determine the type of environment to which the plant part (seed, fruit, spore, etc.) will be carried. In a study of flowering plants on various islands and island groups, it has been found that different types of island situations receive dispersing plant species by different means. Atolls, for example, receive most of their plant colonists by oceanic drift. Plants have colonized high islands most frequently by bird transport. Adherence of bristly seeds to the feathers of migratory birds is responsible for the extensive plant populations of dry volcanic islands. Air transport, whether aided by birds or not, is most effective for carrying plants to islands near the mainland.

Reduction in dispersal powers

Both plants and animals exhibit a loss or reduction of dispersal powers on islands. Examples include an increase in fruit size without concomitant increase in appendages such as hooks that serve in dissemination, a diminution or actual loss of those appendages, and an alteration of the mechanism of the release of fruits. These evolutionary shifts prevent the loss of reproductive products to unsuitable neighbouring environments in which they may not survive. There also tends to be a corresponding reduction in the number of seeds or fruits produced. And, in the case of increases in fruit size, there is the additional advantage of a greater volume of stored food, of particular help to seedlings growing under shady forest conditions. There are probably other less conspicuous changes in the dispersal abilities of plants, affecting such attributes as resistance to seawater, resistance to the digestive enzymes of animals, and the length of viability of seeds.

COLONIZATION OF NEW AREAS

From what has been said about the dispersal powers of organisms, it is a wonder that all taxonomic groups are not found everywhere. They are not, for two principal reasons. Different groups have different dispersal abilities, and environments have a limited capacity for supporting organisms, in terms of both number of individuals and number of species.

Effects of dispersal abilities. The probability that an organism will successfully colonize a new area is lessened by the distance it must travel and increased by the suitability of the new area. The most rapidly successful colonizing species are those with well-developed dispersal powers and a wide tolerance for environmental conditions. Plants referred to as weeds have these properties, and in addition they have the ability to exploit recently disturbed terrain. As a community of plants and animals undergoes a natural successional change toward greater complexity and stability, these pioneer organisms are replaced by organisms inferior in dispersal abilities but superior in competitive abilities. Such pioneer species, both plants and animals, tend to be found in ecologically marginal and geographically peripheral regions, from which position they are well suited to disperse to new areas. Many of them have other advanced biological properties that relate to their role of ecological opportunism, including the ability to reproduce without fertilization.

The future of initial colonists is partly determined by the ability of other competing species to reach the new area. The barrier to be crossed in reaching the new area—be it mountain, sea, river, or climatic zone—is not to be considered an insuperable one; it simply presents an obstacle that reduces the probability of transgression by some organism. There may be what has been termed a filter route across the barrier; that is, a route suitable for only some members of a fauna or flora, those with the most well-developed dispersal powers (Figure 7). An island 30 kilometres from the mainland can be colonized by birds flying across the water barrier, a route suitable only for the more dispersive members of the mainland bird life. Alternatively, the barrier may be almost impenetrable. For some organisms, the Sahara must be such a formidable barrier. The route across it is a "sweepstakes route"—to use the term coined by the American paleontologist George Gaylord Simpson—for the probability of any organism crossing the barrier is very low indeed, and yet some do cross it. Chance plays the major role in determining which ones get across, and when.

Chance in dispersal

The most strongly isolated environments may be so difficult to reach that the initial colonists have the opportunity to undergo evolutionary changes, unaffected by immigration of members of the same species or members of different species. In both plants and animals, these changes involve the relinquishing of their opportunistic characteristics.

Effects of the new environment. On the other hand, an organism arriving in a new area may have little chance of establishing itself, either because the habitat is unsuitable

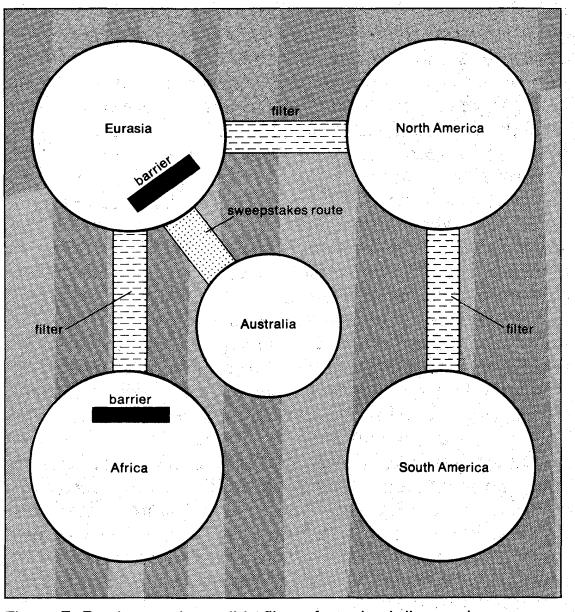


Figure 7: Barriers and possible filters for animal dispersal.

From *Life: An Introduction to Biology* by George Gaylord Simpson and William S. Beck, © 1957, 1965 by Harcourt Brace Jovanovich, Inc., and reproduced with their permission

or because numerous other species already occupy the area. There appears to be a maximum number of species that can be supported in a given area (or volume) of environment. Size of the environment itself is one determinant of this maximum; the larger the area, the larger the number of species (and individuals) present. Climatic and related physicochemical factors are other determinants; these are responsible for the gradual change in species from the equator to the poles. This variation in species number with latitude is well illustrated for the mammals. It is equally well illustrated by mollusks, snakes, birds, ants, or corals. In contrast, variation in species number with longitude is less pronounced and less predictable. The number of species at any point along the latitudinal gradient is more or less the maximum possible, given present-day physicochemical conditions and the total number of species in the world today. In other words, most areas are saturated with species; if a new species arrives in an area, it can establish itself usually only by displacing a resident species.

Species saturation

The process by which an area becomes saturated with species can be expressed in a simple model (Figure 8). Immigration and successful establishment of new species depends on the number of species already present. The rate of establishment of new species declines and the rate of extinction of older species rises until a point is reached when these two processes are equal and no further change in number of species takes place; the area is then saturated. There are interesting variations of this model that allow

Adapted from R.H. MacArthur and E.O. Wilson, *The Theory of Island Biogeography*, vol. 1 of "Monographs in Population Biology," copyright © 1967 by Princeton University Press

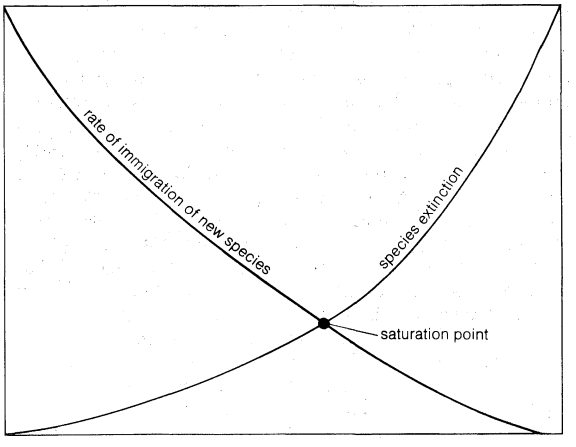


Figure 8: Model of species equilibrium on a single island (see text).

assessment of the importance of size of area and degree of isolation upon the two processes and upon the equilibrium number of species. Large size and weak isolation, for instance, are conditions conducive to a large equilibrium number (many species), and small size and large isolation yield a small equilibrium number (fewer species).

If the environment changes, however, the equilibrium number is subject to change. The shifting of average temperature over a long period of time in water, for example, may enable more species to exist at a given position in that body of water. In the terrestrial environment, the geographic distribution of climate has changed frequently in the past, which has caused a change in the distributions of organisms.

CHANGES IN DISTRIBUTION WITH TIME

Interpretation of past biological events will always remain a challenge because the necessary information will never be complete. The determination of past distributions of organisms, and changes in them, is biological detective work with generally few clues at hand. It is first of all dependent upon fossils that can be identified and located as to place and time of occurrence. Attention is focused less on individual species than on large groups (such as a family or order) or an ecological assemblage (such as a community or population). It is, secondly, based upon the principle that biological and ecological processes operating today are the same as those that operated in the past.

Centres of evolutionary origin. A survey of many groups of organisms shows a common pattern. It is characterized by evolution and proliferation (adaptive radiation) into a diversity of types adapted to performing a variety of ecological roles. The adaptive radiation of marsupial mammals in Australia from a primitive insectivore is a good example. The evolutionary processes are accompanied by expansion of distribution ranges and a geographic radiation. There follows recession or contraction of distribution ranges, of the group as a whole or of individual species, as new groups arise and partly supplant their predecessors. Superimposed upon this pattern are large-scale shifts in distribution, without necessarily any expansion or contraction, as environmental conditions change; for instance, there was a steady southward shift of northern vegetation zones during much of the Tertiary period (66.4 million to 1.6 million years ago), evidently caused by a cooling of northern climates. The distribution of organisms in the past can be viewed as an ever-changing kaleidoscope, the result of movements, countermovements, spreadings, competitions, extinctions, and replacements of a diversity of plants and animals over the whole of the diverse surface of the world.

An examination of the distribution of fossil material and the distribution of present-day organisms suggests that there were regular aspects to the kaleidoscope. It has been recognized, for example, that there were major centres of origin and evolution of terrestrial vertebrates, although the identification of these centres has not met with unanimous approval because the fossil evidence by itself is not convincing. Some biologists have suggested that the north temperate zone was the centre of origin because its variable climate continually presented challenges to vertebrates, to which the vertebrates responded evolutionarily by diversifying. Others have argued instead for the tropics as a centre of origin and evolution, in view of the greater number and variety of species in the tropics than in either temperate zone today. Furthermore, north temperate forms are not usually peculiar to the north, but are representatives of groups also present, often more abundantly, in the tropics; among birds, the hawks, owls, and finches may be cited as among these forms. Finally, many groups present in the tropics are absent in the north.

The tropics as the evolutionary centre

There are other reasons for considering the tropics to be an evolutionary centre, and here historical clues play a large part. It seems reasonable to expect that (1) the region in which the largest number of species and genera existed in the past is where the majority of them originated, (2) the region of greatest degree of differentiation might be the place of origin of a group, and, finally, (3) the region of largest area would have been occupied longest.

The fossil evidence interpreted in the light of these three statements points strongly to the Old World tropics as the main origin of terrestrial vertebrates.

The subsequent spread of organisms. *Environmental influences.* Major shifts in the distribution of whole faunal and floral assemblages can usually be correlated with inferred environmental change. At different times in the history of the Earth there have been different patterns of climate and, more particularly, the average temperature has been higher or lower than at present. Climatic changes have had profound effects upon animal and plant distribution, both directly and indirectly. In colder periods, for example, the Earth was covered with more ice and less water, and the sea level was sufficiently depressed that landmasses now separated by a water barrier were connected.

The best-known of these climatic changes is the most recent Ice Age, comprising several glacial (cold) and interglacial (warm) periods. The sea level was as much as 120 metres lower than at present during glacial periods, and as much as 50 metres higher than at present during interglacial periods. These fluctuations affected organisms most profoundly in polar and temperate regions, least in the tropics.

At the start of the Tertiary (Paleocene epoch), there were great differences between the faunas and floras of North America and Eurasia. There then occurred, according to the geologic and fossil evidence, an interchange of faunas across a land bridge in what is now the Bering Sea. Initially, the interchange involved large parts of the continental faunas and many major groups, most noticeably among the mammals. Subsequently, exchanges were increasingly selective and tended to involve smaller fractions of faunas, progressively less distinctive types of animals, lower taxonomic categories only, and ecological types less novel in the invaded regions. This has been interpreted to mean that the land bridge was increasingly difficult to cross, perhaps the result of an increasingly effective climatic barrier; additionally, each continent may have become increasingly saturated with the major types of animals, making it progressively more difficult for a potential colonist to establish itself. This whole process, extending down through the Tertiary period and up to the recent Ice Age (Pleistocene epoch), was evidently interrupted three or more times by submergence and subsequent elevation of the land bridge.

The dispersal of man

One of the species of mammals taking advantage of the land bridge across the Bering Sea was man himself. By late Pleistocene times man had become a dominant species throughout the Old World. In several waves of dispersal, man entered North America, probably at the last continental glaciation. Dispersal and changing (mainly expanding) distribution have continued to be characteristic of the species.

In contrast to the events in the far north, North and South America were separated during most of the Tertiary period. They were connected by land possibly at the beginning of and then again late in the Pliocene epoch. The effect of this Pliocene connection upon the distribution of organisms can be illustrated with mammals, of which many fossils are available. Before the connection was made, North America had about 27 families of terrestrial mammals, and South America had 29 families. Only one or two families were common to both continents. After the connection, the number of families in common rose to a maximum of 22. Incursions, withdrawals, and extinctions led to the stabilization of 23 families in North America and in South America, as before, 29. North America benefited by the arrival of opossums, three families of ground sloth, two stocks of armadillos, and many others. South America benefited by the arrival of several species or genera of cats (including saber-toothed cats), horses, camels, elephants, tapirs, and rabbits. Most have survived to the present day, although tapirs, camels, elephants, horses, and sabre-toothed cats have since disappeared from North America, and South America has lost a portion of its rich fauna also.

The recurring expansion and contraction of distribution ranges have produced some patterns exhibited by contemporary organisms that at first sight seem inexplicable. Continuous distributions are not difficult to understand; but what past events led to the highly discontinuous or disrupted distributions evident today, such as the distribution of horseshoe crabs in North America and Southeast Asia? Migration from one area to another is possible but unlikely; it is more likely that the distribution was once more widespread and either continuous or moderately discontinuous and that a shrinking of distribution range has occurred. The distribution of fossils indicates that horseshoe crabs were widespread and abundant in the Paleozoic era and that a restriction of range occurred in the Mesozoic, possibly as a result of competition with newly evolved marine invertebrates.

Other species were profoundly affected by climatic disturbances, such as the recent Ice Age, surviving partly in relatively small areas of suitable habitat, or refuges, and partly in larger areas farther away, to which they dispersed when the climatic change rendered the previously occupied area unsuitable. The evidence for this consists of numerous instances of much differentiated forms isolated from the main range of the group. These are referred to as relicts. Examples are Arctic plants found isolated on mountaintops in north and middle Europe. Even before Darwin's time, their occurrence was explained as due to survival in refugia, areas untouched by glaciers but once surrounded by them.

Relicts

Certain groups of coastal plants still may be confined to their glacial refugia because they are not evolutionarily versatile enough to break out from them; their variability became restricted during the past glaciation when new variants would have been at a disadvantage owing to the harsh conditions. In contrast, there are some species, termed disharmonious relicts, which occur in unusual and isolated habitats where they are thought to have survived because of broad ecological tolerance in the past.

Relict patterns of distribution are found on islands as well. These are in irregular and unpredictable patterns interpreted as having been formed by partial extinction of an old fauna or flora formerly common to all the islands. They are most pronounced on oceanic islands. An example is the distribution of frogs on islands in the Indian Ocean. The pattern is replete with discontinuities and mixed relationships, which indicates that the fauna is partly a relict one and that existing geographic relationships are partly the result of extinction and survival rather than of direct dispersal. The relict pattern of distribution contrasts with an immigrant pattern, characteristic of continental islands, in which the number of species in a taxonomic group decreases in more or less orderly fashion with distance from the continent, although modified by island area.

Geologic influences. Much has been said about the distribution of terrestrial organisms, as if the organisms have moved and the land has stood still. But, in fact, a large body of geologic and fossil evidence suggests that the continents have not always been in their present positions. In view of this evidence, the distributions of organisms, in relation to the Earth's axis, have undergone even more change than has been referred to so far.

Continental drift

The theory of continental drift holds that the continents were once connected but that they split apart during or after the Triassic period (245 million to 208 million years ago) and gradually drifted apart. In conjunction with seafloor spreading and mountain building, continental drift conforms to a broader theory according to which large plates making up the Earth's crust "float" on the molten mantle, variously colliding and separating (see PLATE TECTONICS). The earliest proponents of a once-unified landmass cited as evidence the present shapes of continents. The outline of West Africa, for example, is complementary to that of eastern South America. In fact, if the two continents are fitted together like a jigsaw puzzle, the Paleozoic and early Mesozoic rocks of the continents in the vicinity of their (present) coasts show remarkable correspondence in structure as well as magnetic orientation. The paleontological evidence is that the flora and fauna of the two continents showed a great degree of resemblance in those early times, and this resemblance may have diminished progressively thereafter.

Human influence. The agricultural, industrial, and commercial activities of humans have had an enormous impact upon the distribution of organisms. In the last 10,000 years human beings have replaced the climate as the most important agents of change.

Some organisms are more widely distributed now as a result of human activities, many more are more restricted, and a sizable number have disappeared altogether. Since 1600, almost 100 species of birds and 40 species of mammals are known to have become extinct, and it is probable that humans exercised an important, if not decisive, influence in these extinctions. Moreover, many species are now on the verge of extinction, again largely because of human activities.

Human influence upon the distribution of organisms has been, and continues to be, needlessly destructive. Admittedly, some destruction is probably inevitable, since species like man, with novel ecological features, spread and cause a reduction in numbers of other species. But this process leads to a state of dynamic equilibrium, to something approaching stability. Unfortunately, modern man shows no sign of approaching an equilibrial relationship with his environment in the immediate future. (P.R.G./Ed.)

Biotic interactions

As previously noted, no species lives in isolation from other species. Plants grow together in spatial patterns determined in part by competition and are fed upon by characteristic herbivorous animals, which in turn serve as food for carnivorous animals. Both plants and animals are attacked by parasites. Animals interact with each other competitively and sometimes beneficially. Some of these interactions are temporary, casual, and of minor importance, whereas others are permanent, vital, and of major significance.

Intraspecific and interspecific interactions

Biotic interactions may occur between members of the same species (intraspecific interactions) or between two or more species (interspecific interactions). They may involve nutritional benefits, space in which to live, shelter or protection, transport, or reproductive capability. Many interactions are highly specialized and complicated, involving adaptive changes in structure, function, behaviour, and ecology. Some create negative effects upon the interactants, reducing survival or reproductive success; other interactions produce positive effects, increasing survival or reproductive success. In some cases, members of an interaction are apparently unaffected.

Biotic interactions are significant not only because they influence individual species but also because they constitute the principal stabilizing, connective linkages among the various species contained in a biological community. In a general sense, species in a biological community persist in relative harmony because of biotic interactions, despite individual gains and losses. As a consequence of this web of interactions, the community as a whole persists, with all species ultimately contributing to its continuation.

INTERACTIONS WITHIN A SPECIES

Negative interactions. Intraspecific competition provides the ultimate force limiting the abundance of a species in a community. Natural enemies may be present or absent, but others of the same species are always present, providing the competition that insures against overexploitation of available resources and the irreversible destruction of a community due to overpopulation.

Cannibalism

Cannibalism, fighting, and territorial defense are some of the overt methods by which competition manifests itself. Many animals, both herbivores and carnivores, exhibit cannibalism when overcrowded, and some are cannibalistic whenever they contact another of their own species. The codling moth caterpillar (*Cydia pomonella*) in the apple is of the latter sort; hence, there is never more than one worm per apple. Mice and rats, when crowded, undergo psychological disturbances leading to excessive fighting and litter destruction. Territoriality, the dividing up of resources, mostly space, is a common means of defense against overcrowding. It differs from simple aggression in that, instead of involving individuals against individuals, it involves groups of individuals, breeding pairs, families, or herds. Migration away from overcrowded conditions is another common intraspecific response. It may be individualistic, with each member of a population going its own way, or it may involve mass flight, such as occurs with locusts, butterflies, and birds.

Positive interactions. On the positive side, numerous interactions among individuals of the same species lead to clustering, aggregation, swarming, herding, and, ultimately, development of social groups. Tent caterpillars (*Malacosoma*) live together in large colonies, foraging together in masses and resting at night in communally spun silk tents. Isolated caterpillars are unable to survive; only the aggregation thrives. Parental care provides the basis for many aggregations of like individuals. Parental protection of the egg, as occurs in birds, some insects, and some fish, is the simplest form of care. Feeding and protection of the young is a more advanced level of care. Insect societies, in which one reproductive adult, the queen, is served by her own progeny, represent a still more highly developed custodial configuration. In such cases the offspring remain with the parent, providing protection for the colony, care of younger members, nest building and cleaning services, and food gathering and storage. Division of labour to attend to these services is common.

Some species live in habitats where—because of small size and weak flight ability, limited and scattered resources, and relatively low population densities—the chances of the mature sexes meeting for reproductive purposes are low. One method for overcoming this hazard is the phenomenon of autoparasitization, in which one sex lives as a virtual parasite on the other, an unusual form of intraspecific interaction. Certain insects exhibit this mode of life: the males of scale-insect parasites develop as hyperparasites on their own females. Some deep-sea angler fish also employ this habit, with the very small adult male living in permanent attachment to the body of the larger female. In such situations the association of the sexes is facilitated, and reproduction in hazardous environments is ensured.

THE RANGE OF INTERSPECIES ASSOCIATIONS

Positive, negative, and neutral effects

One system of analyzing two-species interactions assigns positive (+), negative (−), or neutral (0) value to the interactants on the basis of how they affect each other. If the survival or reproduction of a species is enhanced, it is a gain and given a plus value. If survival or reproduction is reduced, it is a loss and given a minus value. If survival or reproduction is unaltered and the species unaffected, it rates a zero value. From such a rating scheme, a two-species (A, B), three-effect (+, 0, −) table can be constructed in which the bases for association (symbiosis) of all biotic interactions can be defined (see Figure 9).

species A (large, strong) \ species B (small, weak)	+	0	−
+	+, + interdependence (obligative mutualism) protocooperation (facultative mutualism)	+, 0 commensalism	+, − herbivory predation
0	0, + commensalism	0, 0 neutralism	0, − amensalism (competition)
−	−, + parasitism parasitoidism	−, 0 amensalism (antibiosis)	−, − mutual antagonism

Figure 9: Bases for association of all biotic interactions.

Mutualism (+, +) is an association of two different species that results in mutual benefit or gain. Obligative mutualism requires that both members interact together, or neither survives. Facultative mutualism, or protocooperation (+, +), refers to a less-rigid association in which one or both members can survive in the absence of the other. In commensalism (0, +) one member benefits while the other remains unaffected. A case in which the small or weak species is unaffected while the larger or stronger

benefits, especially in nutritional relations, is sometimes called allotrophy (+, 0).

In parasitism (−, +) the smaller organism benefits at the expense of the larger. Many parasites are microorganisms that cause disease, but not usually the death, of the host. A special case of parasitism, which involves insects that do kill their hosts, is called parasitoidism. Predation (+, −) refers to an association in which a large or strong animal consumes a small or weak animal. Herbivory (+, −), the consumption of plants by animals, is generally comparable to predation. Amensalism (0, −) occurs when one organism loses while the other organism is unaffected. Unilateral competition, to which the term competitive exclusion refers, and antibiosis, in which one species remains unaffected while a competitor is deprived, are examples of amensalism. In antagonism (−, −) both species lose or are harmed. Neutralism (0, 0) is an association in which neither species is affected by the presence of the other, the two species sharing a localized resource or habitat.

The following sections will provide selected examples of the various forms of biotic interactions, based on the definitions above. The emphasis rests on the pattern of the interactions rather than the symbiotic associations themselves.

CONSUMPTION: ORGANISMS EATING ORGANISMS

The basis of community food webs

Herbivory: animals eating plants. Herbivory constitutes one of the most important classes of biotic interactions. It serves as the main connection between all plant and animal life, the basic food association (trophic link) through which the composition, dynamics, and variation of community structure takes place. All higher organisms, primary and secondary carnivores and scavengers—essentially all the remaining links in the community food chains and food webs—are more or less dependent on this basic link (see Figure 10).

Herbivory is encountered at all levels of animal life, from the smallest to the largest animal, and in all habitats. No plant species is exempt, whether in the Arctic tundra, the tropical forest, or the desert and whether on the highest mountain, along the seashore, or in mid-ocean. It is a reciprocal interaction, or coaction, wherein the number or quantity of plants is altered by the number and consumption rates of the herbivores and in which, conversely, the number of herbivores is limited by the quantity of plants. This coaction, therefore, may approach a state of balance in which the amount of plant material is just sufficient to sustain the number of herbivores associated with it, and the herbivores collectively remove only the surplus plant material produced. The manner in which specific herbivores respond to the results of this interaction is dependent on their food habits. Species that require one kind of plant material are very sensitive to changes in abundance of their preferred food, whereas less-specialized species simply shift their attentions to other plants when the quantity of a given one fluctuates.

The degree of herbivory is rather wide. A plant may be totally consumed, as it is when a protozoan or a fish consumes unicellular planktonic algae. Vital parts of a plant may be eaten, leading to the eventual death of the individual, as occurs when bark beetles girdle a pine tree or when birds eat the tender tips of germinating plants. When only nonvital parts of a plant are eaten—including leaves, twigs, and roots—the plant may be stunted or not affected at all. In the latter case, which is quite common, the herbivore is truly living on "surplus" plant life.

Herbivory is one factor responsible for limiting the amount of plant life in a community and, therefore, influencing the composition of communities. It serves as an important means of recycling plant tissues back into the nutrient minerals, carbon dioxide, and water from which they were derived and enters into the determination of the numbers and kinds of animals associated with the community.

Herbivores to control weeds

Plant pests can drastically reduce the growth or productivity of crop plants. Successful attempts to use specific herbivorous insects to control certain weed species (*e.g.*, cactus in Australia or the poisonous Klamath weed in California) and the remarkable recovery of plant populations when an adverse factor is removed (*e.g.*, a pest) or a stimulating factor is added (*e.g.*, nutrients) suggest that herbivore–plant interactions are of critical importance to community composition. Plant–animal interactions are also illustrated in the overgrazing that occurs when deer, protected against predators, increase to such numbers that a plant community is threatened by excessive browsing. Grazing by stock animals, sheep, and cattle is an important factor in the composition of range and pasture lands; if excessive, grazing can actually destroy such communities. A rapid increase in plant growth brought about by excessive nutrient deposition (eutrophication) in bodies of water through runoff of soil fertilizers and discharge of municipal and industrial wastes results in "blooms" of aquatic algae and the disruption of the normal trophic balance between algae and algal feeders.

The concentration of a few plant species into large areas, as is the general practice in agriculture, provides a potential for excessive herbivory. Such large tracts of single crops (monocultures) concentrate and augment the numbers of herbivores (when insects, they are usually called phytophages) to the point that massive population outbreaks commonly occur. By contrast, natural communities, which are diverse in species, are much less prone to support pest outbreaks: an expression of the ecological principle of stability in diversity in community life.

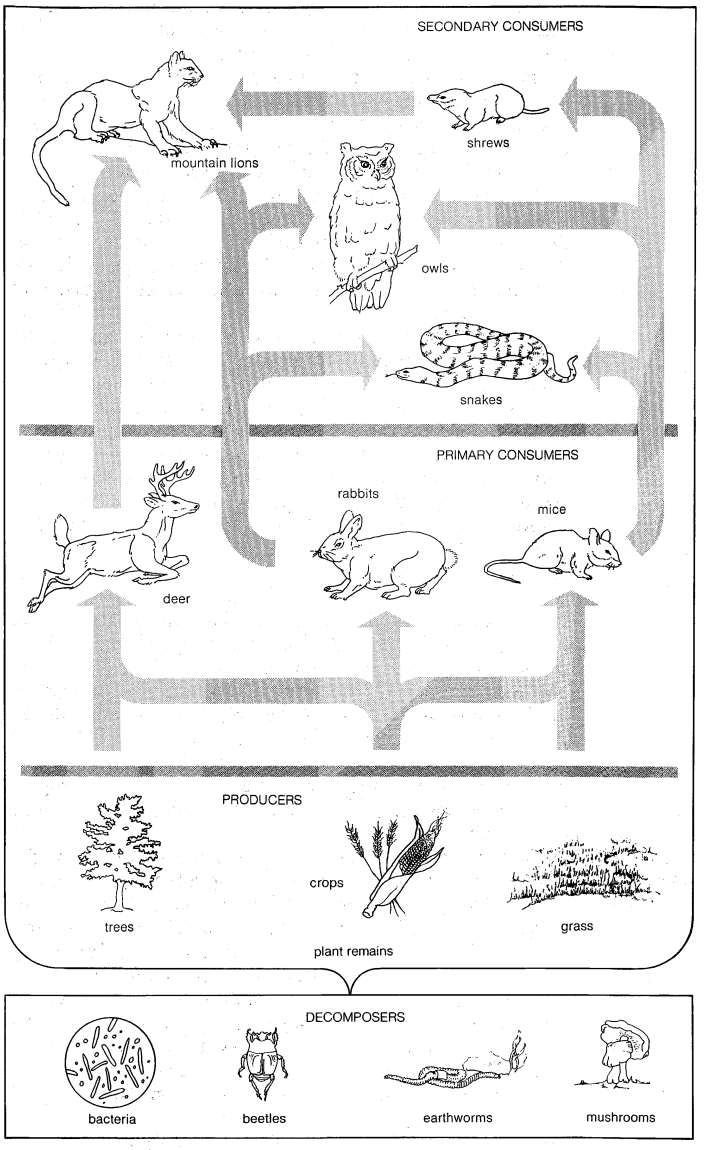


Figure 10: A representative food web comprising several interacting food chains.

From *Biological Science: An Inquiry into Life*, 2nd ed. (1968); Harcourt Brace Jovanovich, Inc., New York; by permission of the Biological Sciences Curriculum Study

Predation: animals eating animals. Just as most plant species are fed on by herbivores, so are most animal species subject to predation by other animals. By feeding upon and reducing the numbers of both herbivores and other carnivores, predators constitute the higher links in the nutritional pathways characteristic of communities. Predators substantially affect the numbers of herbivores and thus influence their impact on plants. When the numbers of predators fluctuate, their effects upon prey and plants vary accordingly (see below *Biological populations: Changes in population characteristics: Fluctuations in stable populations*).

Because of the obvious importance of and interest in the mechanics of predation, predator–prey interactions have been the subject of much scientific research. Laboratory and theoretical studies have been made of isolated predator–prey "systems" and of predator–prey "models." The degree to which such systems and models simulate natural predator–prey population fluctuations has been used as a basis for judging the "soundness and realism" of the scientific methods and for "explaining" the process of predation. Simple mathematical expressions have been devised that represent predator and prey numbers and indicate that predators vary in numbers in a smooth, oscillatory manner purely as a result of prey abundance changes, and vice versa. This latter result, that prey oscillate because of predation, is the important new "evidence" that prey–predator interactions are, indeed, reciprocally coupled.

The importance of predation in determining prey abundance and community stability can be illustrated by several examples. When man (a predator) overexploits his animal resources (in effect, prey), they decline in numbers sufficiently to affect subsequent utilization (consumption). Such is the result when fish, whales, sea otters, buffalo, and other animals are excessively hunted. Whales and sea otters have declined drastically; buffalo became exterminated as a wild animal, and grizzly bears, mountain lions, and elks similarly are threatened with extinction.

Predators to control pests

An illustration approaching the point from the opposite side is the employment by humans of predators to control pests, a technique called biological pest control. The effective suppression of the cottony-cushion scale (*Icerya purchasi*) pest of citrus in California in the late 1800s by use of the Australian ladybird beetle (*Rodolia cardinalis*) is a classic example. Since that time more than 100 successful cases of biological control, leading to permanent suppression of pests, have been accomplished.

The unintentional unleashing of new pests through the use of pesticides, which cause the destruction of natural enemy predators, illustrates the delicate balance between predator and prey that exists in most communities.

Occasionally, outbreaks of predators occur, with resultant drastic impact on prey populations. An example is in the Great Barrier Reef, Australia, and near Guam, in the western Pacific, where the crown-of-thorns starfish (*Acanthaster planci*), a predatory species, has increased in numbers that have seriously reduced the abundance of many corals. Sometimes the predator attacks a valuable resource of humans. The whelk, a marine snail, attacks oysters. An American whelk, the oyster drill (*Urosalpinx cinerea*), has invaded oyster beds off southern England, causing severe reductions in harvest since the late 1950s. And humans have always been uneasy about both bird and mammal predators—such as hawks, eagles, coyotes, wolves, and foxes—that attack their domesticated animals.

PARASITIC INTERACTIONS

Parasitism. In practically every community, plant and animal populations are burdened with parasites, which include viruses, fungi, protozoans, nematodes, many marine and freshwater coelenterates, flukes and tapeworms, leeches, some insects, certain birds, and higher plants such as dodder, mistletoe, and broomrape. A parasite usually does not kill its host (if it does, the parasite itself may die before reaching a new host), but it generally reduces the growth rate, survival ability, and reproductive capacity of the host. From an ecological standpoint, therefore, parasitism does constitute a considerable burden on a host population.

Parasites that live on the body surface of the host are called ectoparasites. They do not commonly cause disease in their hosts but rather suck blood or create superficial damage to the skin; examples include leeches, fleas, lice, and ticks. Other parasites live inside the host's body—either in cells or in spaces lined by cells (*e.g.*, intestine, blood vessels, mouth, etc.). Parasites that live within host cells—such as many bacteria and viruses—are called intracellular endoparasites; those that inhabit spaces within the host's body are called intercellular endoparasites.

Many disease-causing organisms, or pathogens, are endoparasites carried from host to host by some other organism; malaria, for example, is caused by a protozoan endoparasite transmitted by mosquitoes. In such situations, the biotic interaction involves three necessarily coexisting species: the pathogen, the carrier (or vector), and the host. A number of plant diseases transmitted by insects are dependent on carrier density, pathogen abundance and virulence (disease-producing power), and host density. Host susceptibility is also a significant factor in influencing the outcome of host–parasite interactions. In Europe the Dutch elm disease fungus (*Ceratocystis ulmi*) is a minor pest of elm trees (*Ulmus*). When this pathogen established itself in North America, however, it severely infected native American elms (*U. americana*), which were more highly susceptible than their European counterparts. A similar event occurred when the chestnut blight (*Endothia parasitica*), a fungus native to chestnuts in China and Japan, invaded North America and devastated stands of native chestnuts (*Castanea dentata*).

Not all invading parasites are microorganisms, however. The sea lamprey (*Petromyzon marinus*), an ectoparasite on fish, attaches itself to its host by means of a suckerlike mouth with which it sucks blood and soft tissues. The sea lamprey is native to the North Atlantic, living in the ocean but breeding in coastal rivers and streams. It was able to move into Lake Ontario but no farther, being blocked by Niagara Falls. After construction of the Welland Canal, however, which bypassed the falls, the lamprey invaded the remaining Great Lakes and began decimating the populations of commercially valuable fishes.

Intermediate hosts

An intermediate host—interposed between the target host and parasite—adds ecological complications to host–parasite interactions: the parasite, host, and intermediate host must all be present in the same community in suitable numbers at the same time. The liver fluke (*Fasciola hepatica*) that attacks sheep and cattle, for example, requires an aquatic snail as an intermediate host; there is no direct transmission of the fluke from sheep to sheep. Sheep, snail, fluke, pasture, and aquatic habitat must all coexist in the same community. Many other common parasites similarly require an intermediate host; *e.g.*, malaria, sleeping sickness, and wheat stem rust. One reason that such complicated three-component interactions are so common is that with at least one of the organisms the reproductive capacity is very high; this increases the likelihood of continuation of the pathogen. Furthermore, in the case of parasites carried by insects, the motility and numerical abundance of the insect carrier facilitates parasite dissemination.

Parasitoidism. Parasitoids—insects that parasitize other insects—differ from true parasites in that, like predators, they always kill their hosts. Most are free-living in the adult stage. In a typical life history, the female parasitoid lays an egg in or on a host; the ensuing larva feeds upon host fluids and tissues, eventually killing the host; and the fully developed larva pupates and later emerges as an adult. Most species of parasitoids have a distinct preference as to the stage of host attacked, whether egg, larva, pupa, or adult.

Parasitoids are among the most important agents in limiting the numbers of other insects. For this reason they, along with true insect predators, are frequently sought and used by humans as control agents for insect pests. Almost all insect species are attacked by one parasitoid or another. Some parasitoids are gregarious, with several living in one host simultaneously; others are solitary, with only one developing in a host. Certain wasp parasitoids exhibit polyembryony, wherein one egg deposited in a host may

give rise to dozens or thousands of parasitoid embryos, and the host is literally overwhelmed by the developing parasitic larvae. In some parasitoid flies the female lays her eggs on leaves, and the host becomes infected by eating the egg-bearing leaves. Other parasitoid flies lay their eggs on the backs of hosts, and the larvae hatch out and bore into the host body.

The parasitoid's life cycle is always well attuned to that of its host. If the host is a leaf-feeder and found only in the foliage of trees, that is where the female parasitoid searches for them. If the host has but one generation a year, so usually does the parasitoid. If the host passes the winter hibernating in the soil, the parasitoid generally does likewise, emerging from the soil in the spring soon after the host does.

Hyperparasitism

A parasitoid may itself sometimes be parasitized. In such cases its parasites are called hyperparasites. The hyperparasitic habit, commonly found in insects, may extend to several levels; for example, an aphid (order Homoptera) can be attacked by a solitary, endoparasitic aphid wasp (family Pemphredoninae), which, in turn, may be parasitized while still within the host's body by an endoparasitic gall wasp (family Cynipoidea). Moreover, the gall wasp can itself be attacked by an ectoparasitic pteromalid wasp. In theory even additional levels of parasitism are possible.

The free-living habit of the adult parasitoid enables it to spread rapidly from host to host throughout a community. Because the host almost always dies as a result of a parasitoid attack, few host-immunity responses have evolved. Only when parasitoids attack unusual hosts—those not in their normal host range—can a host survive and sometimes elicit an immune response to further attacks. The true parasite and the parasitoid can sometimes interact in their association with a host insect. Some parasitoid wasps coincidentally distribute pathogens from host to host when they deposit their eggs.

Brood parasitism. Brood parasitism, the laying of eggs in another animal's nest to be reared therein by the host, occurs in certain birds, including the cuckoo and the cowbird, and in certain insects, including cuckoo wasps. The female cuckoo bird (family Cuculidae) lays an egg in the host's nest, where the hatched cuckoo chick either kills or pushes out the host's brood, so that it is reared alone until it can fly. The cuckoo wasp (family Chrysididae) accomplishes much the same result, except that its host is one of the social wasps or bees. It enters a colony and lays an egg in one of the compartments, or cells. After hatching, the parasitic larva destroys the normal cell occupant and is taken care of by the colony workers.

The cuckoo habit is a balanced one, not being so efficient as to suppress unduly the host population but effective enough to assure perpetuation of the habit. In most cases, the reproductive habits of the host species enable progeny to be produced at times when the brood parasite is not active.

AMENSALISM AND ANTAGONISM

Competition. Competition occurs when two species utilize a common, limited resource (such as food, space, or moisture), and, in so doing, one species interferes with, injures, or deprives the other. When the competitors are of the same species, intraspecific competition occurs (dealt with earlier). Populations of two species cannot persist together for very long in the same community when both compete for and are limited by a common resource. This

Gause's hypothesis

principle, known as Gause's hypothesis, further suggests that the species having the greater competitive advantage—manifested in greater searching ability, more rapid exploitation of the resource, greater numerical growth rate per unit of resource attained, greater aggressive or fighting ability—in time gains more and more of the limited resource. Competitive interactions between species are resolved in either of two ways: (1) the exclusion or displacement of one species by the other or (2) the coexistence of the two competing species as a result of minor differences in habit or because their numbers are limited by factors other than the resource under consideration. When a species invades the habitat or community of another species and both occupy the same niche, either the incumbent will persist and the invader will disappear, illustrating competitive exclusion, or the invader will become established and the incumbent will disappear, resulting in competitive displacement.

Actual competition is difficult to see in nature, since a community usually contains "superior" competitors only, the lesser adapted species having already been ejected from the area. When a species invades a new habitat, however, either through its own dispersive powers (rare) or through human activities (common), competition can be observed in operation. Hence, when settlers established themselves in western North America and brought with them grains and forage grasses for their livestock, they also brought, unwittingly, many weed seeds from Europe. Many of those weeds and foreign forage plants became established, forcing out or eliminating many native plants. The Klamath weed (*Hypericum perforatum*) is a remarkable example of the way in which weedy competitors can succeed in a new habitat. Gaining a foothold in California a few years before 1900, Klamath weed established itself as the dominant perennial on low-elevation rangelands. Several decades later, millions of acres of land were infested. Only when the weed itself was suppressed by biological control agents in the late 1940s were the displaced native species able to return to their old habitats. In England the shrub purple-flowered rhododendron (*Rhododendron ponticum*), an invader from the Continent, has become a pest in wooded areas, displacing holly and preventing regeneration of oaks and other desirable trees.

Insect invaders

Many insect invaders have displaced incumbent species through competition. The Argentine ant (*Iridomyrmex humilis*) and the imported fire ant (*Solenopsis saevissima richteri*), both invaders of North America from South America, have displaced many native ant species. When the Oriental fruit fly (*Dacus dorsalis*), the maggots of which feed on tropical fruits, invaded Hawaii in the mid-1940s, it apparently displaced an earlier invader, the Mediterranean fruit fly (*Ceratitis capitata*), from all the low-elevation habitats that the latter had occupied for so many years. The Mediterranean fruit fly was forced up into higher elevations on volcanic slopes, where coffee remained its major host, leaving the lower slopes and shores to the newer invader.

In aquatic habitats, aggressive fish such as carp and perch commonly displace trout and other game fish. When the long-legged crayfish (*Potamobius leptodactylus*), a native of eastern Europe, was introduced into some central European lakes, it totally displaced the native broad-legged crayfish (*P. astacus*).

Laboratory experiments also show the nature of competition. When two species of the single-celled protozoan paramecia compete for food and space in small glass tubes, one displaces the other, even though both thrive when grown alone. Similar results are obtained when the confused flour beetle and the red flour beetle (*Tribolium confusum* and *T. castaneum*) are cultured together in jars of medium. Either species, living alone, limits its own population numbers (intraspecific competition) through such self-regulatory actions as cannibalism (adults and larvae feed on eggs and pupae), interference with female egg laying, and defilement of the common food supply. When populations of the two species are grown together, one species is eventually eliminated; the other persists. The interesting difference in these experiments is that, possibly because of slight variations in genetic composition of the organisms used, the same species does not always survive in every experiment. Rather, there is an element of probability as to the outcome. The probability of winning the competition can be varied, sometimes substantially, by altering the physical conditions for growth, such as temperature, moisture content, or humidity.

When two species of closely related water fleas, *Daphnia pulicaria* and *D. magna*, are cultured together with algae as food, both species increase in numbers for about three weeks; then the *D. magna* population ceases to grow and instead gradually declines to extinction at the end of the sixth week. The competition for the common food supply, not important while the populations were small, soon becomes severe enough to affect strongly the less adapted

of the two competitors. On a different food supply, yeast cells, *D. magna* actually increases at a greater rate at first, but after about 24 days it reaches a peak and declines to extinction, as in the previous case.

Coexistence of potentially competitive species

It often happens that two species with similar niches occur in adjacent geographic areas and partially overlap. Coexistence in the area of overlap is then possible as a consequence of the continuous migration of the "excluded" species from its area of exclusivity. Displacement takes place, but the dispossessed are constantly being replaced by new immigrants into the "overlap" area. Eventually, if the area of overlap is broad enough, niche separation may occur through the process of evolution. Niche separation, or displacement, is the shifting of one or the other competitor species from the common, limited resource to another resource not competitively limited. This can be a different food, different spatial habitat, or different nesting site. An example of niche displacement is that of the finches on the Galápagos Islands. Some of the 12 or so finch (order Passeriformes) species that have evolved from a mainland insect-eating finch have become vegetarian, whereas others have remained insectivores; some occupy ground-level habitats, others inhabit trees, and one lives on cactus; some eat seeds, others fruit. These are all ways by which the original stock extended into unutilized niches.

Occasionally, niche displacement is reflected in a structural or behavioral change in the species involved, a phenomenon called character displacement. The change in character, or feature, occurs in the region of overlap but not in the regions of exclusive occupancy. A good example is provided by the same Galápagos finches. Two species living on separate islands have approximately the same bill length; when they occur together on other islands, they exhibit altered bill dimensions, which reflect changes in food resources brought about by competitive interactions. A similar displacement of characters is seen in two related ant species—the yellow ant (*Lasius flavus*) and the good ant (*L. nearcticus*)—that coexist in the eastern United States and differ in at least eight features. These differences are not apparent in *L. flavus* in areas where it occurs alone in the western United States.

Antibiosis. Antibiosis occurs when one species interferes with or injures another through the secretion of a chemical substance. This commonly occurs among the bacteria and fungi and to a lesser extent among higher plants and some animals. The observance of such an effect in certain penicillium molds is well known. A bactericidal substance secreted into the growth medium prevented the growth of competing microorganisms and led to the discovery of the antibiotic penicillin. A more complicated example concerns needle, or wire, grass (*Aristida oligantha*), which invades old-field communities. It secretes phenolic acids, which inhibit the development of nitrogen-fixing bacteria and the blue-green algae of the soil. This, in turn, suppresses the production of available nitrogen in the soil, thus slowing the invasion of other competing, nitrate-requiring plants into the grass community.

Plant resistance

Antibiosis is the basis of the resistance exhibited by certain varieties of crop plants to insect attack. Corn varieties resistant to the European corn borer (*Pyrausta nubilalis*), for example, contain benzoxazolinones, chemicals that adversely affect larval growth and survival.

The discovery that many plants and some animals contain or secrete chemicals injurious to competitors or natural enemies has led to development of the study of allelopathy; the chemicals are called allelochemics. This phenomenon—the suppression of some higher plants by chemicals released by another higher plant—has been extended to include chemical defenses of plants against herbivores, phytophagous insects against predators, and the resistance of hosts to parasitoids.

Examples of plant-to-plant antibiosis based on allelochemics include the chaparral plants, whose toxic phenolic secretions are washed by rains into the soil, where they inhibit the germination and growth of herb seeds close enough to provide competition. The milkweed-feeding caterpillar of the monarch butterfly (*Danaus plexippus*) acquires toxic chemicals from its host plant, thereby rendering the adult butterfly distasteful to bird predators.

Antibiosis can take forms other than chemical defense. They include immune responses, whereby hosts defend themselves against invading parasites. Immune responses of a host to an invading organism can involve engulfment (phagocytosis), chemical inactivation (precipitation), dissolution (lysis), or encystment.

Mutual antagonism. Mutual antagonisms occur when a biotic interaction between two species results in harm or death to both. Most mutual antagonisms occur as a result of competition for a limited resource. In some cases two pathogenic species together invade a host and thereby bring about its death, but in the process they destroy themselves; either pathogen alone would normally not destroy the host. An insect parasitoid and a predator occasionally engage in competitive interactions that result in death to both competitors, a phenomenon called synnecrosis.

When more than one insect parasitoid parasitizes a given host, competition for the host results. For example, when the female wasp parasitoids *Praon exsoletum* and *Trioxys complanatus* both attack the same aphid host, usually only one parasitoid survives, but occasionally the competition is so disruptive that the host dies as well, and with it all competitors. A variation of this habit is superparasitism, or the depositing by one species of more eggs into a host than can survive. The usual, adaptive outcome of such a situation is the death of all but one, an example of intraspecific competition. When hosts are in relatively short supply, however, superparasitism to an excessive degree commonly takes place. The result often is the death of both the parasitic larvae and the host itself.

It has been said that the true parasite rarely exploits its host to the point of death, for the death of the host means the death of the parasite. The "prudent" parasite has evolved to avoid such drastic interaction, and the host likewise has evolved a resistance (immunity) to a parasitic species of long-standing association. When the pathogen, however, invades a new host species or subspecies or a previously unparasitized, susceptible host population, the resulting association may be one of extreme virulence, with rapid onset of death of the host.

Another example of mutual antagonism that is clearly harmful to both interacting species is the case in which cattle consume halogeton plants, whose leaves contain a poison that frequently kills the cattle.

COMMENSALISM

In many communities there are species that benefit through interaction with other species, while the latter remain indifferent or unaffected. Such unilateral benefits include attainment of nutrients, space or support, shelter or protection, and transport. The commensal population is, by virtue of the association, increased in numbers through improved survival. If the commensal is obligatorily associated with a host, it cannot live in a community that does not contain the host in question.

Nutritional commensalism. Nutritional commensals, or allotrophs, are well illustrated by the many organisms found in the alimentary tracts of higher animals. Such endosymobiotes use host waste products for food and acquire suitable water supplies and a suitable microclimate in which to live. No harm is done to the hosts. Endosymbiotic bacteria and yeasts are found in most large mammals, particularly cattle, horses, sheep, deer, and buffalo.

Endosymbiotes

Numerous bacterial and yeast colonies grow on the bark, twigs, and leaves of trees and large shrubs, particularly in tropical habitats. These nonparasitic organisms live on the nutritious exudates and the decomposing bits and flakes of dead bark that coat the surfaces of such plants. In the soil many fungi and bacteria live on nutrients derived from the root exudates of higher plants.

Aquatic organisms with poor dispersal powers or poor food-gathering capabilities often attach themselves to species on whom they can depend to provide those qualities. The small crab *Lissocarcinus* lives commensally on the surface of sea cucumbers (class Holothuroidea), being protected by colour camouflage from natural enemies and gaining its food simply by diverting bits of the plankton that is pulled into the sea cucumber's mouth by feeding currents. Clown fish (*Amphiprion*), which habitually swim

among the tentacles of sea anemones (order Actiniaria), are unaffected by the stinging cells (nematocysts) of the host, but they receive protection from small predators and also partake of surplus bits of any food captured by the anemone. The well-known remora (family Echeneidae) and the pilot fish (*Naucrates ductor*) also obtain their food from the activities of such hosts as a shark, marlin, or swordfish. The remora attaches itself to its host by a sucker organ, whereas the pilot fish swims in close association with the host. Both commensals feed on the leftovers of their host's meals.

Numerous birds are food commensals of other animals. Cattle egrets (*Bubulcus ibis*) in Africa follow herds of elephant, buffalo, or antelope, feeding on insects and grubs turned up by the grazing activities of such animals. The cowbird (*Molothrus*) behaves similarly with cattle in the Americas. In northern Europe, ptarmigan (*Lagopus*) travel with caribou in order to get the insects uncovered from the semifrozen sod. Gulls, lapwing plovers, and herons follow the farmer's plow to attain soil organisms turned up to the surface. Gulls and albatrosses follow ships at sea, feeding on garbage tossed overboard or on small fish stirred up by the passing vessel. Vultures follow carnivores such as lions, leopards, jackals, and hyenas and scavenge upon what is left of a killed carcass.

Insects such as biting lice, fleas, and louse flies, commonly categorized as ectoparasites, are more correctly designated ectosymbiotic commensals; they feed on feathers, sloughed-off flakes of skin, or waxy epidermal exudates. Only in the few cases of blood sucking are they regarded as true parasites.

Some insects and spiders closely resemble ants and associate with army ants in their foraging columns. These mimics feed on the booty flushed up by the foraging ant column. Certain ant mimics live within the nests of the host and as such are termed symphiles. Such symphiles include some of the rove beetles, claviger beetles, and certain mirid bugs.

Physical commensalism. The need for living space is a common basis for commensalism. In overcrowded communities, offshore marine habitats, or tropical rain forests, many nonmotile (sessile) microorganisms are unable to find adequate sites on which to become established; hence, many of them attach themselves to the surfaces of other plants or animals. Most sessile or slow-moving marine animals—sponges, corals, bivalves, snails, turtles, and whales—carry ectocommensals such as algae, hydroids, and barnacles about with them. Woody plants physically support numerous orchids, ferns, bromeliads, grasses, and mosses of the tropics in tropical habitats. Trees in the temperate zone provide support for mosses and lichens.

Insects and mites live in birds' nests or rodent burrows, feeding on dead organic litter. Nest inquilines—insects living in the nests of bees, wasps, termites, or ants—subsist on excess food stores, nest materials, or dead hosts. Some inquilines apparently find the internal physical habitat of these social insect nests favourable for their own well-being.

Batesian mimicry

Batesian mimicry, the morphological and colour-pattern resemblance of certain animals to other animals, is a form of protective commensalism. Edible butterflies, through natural selection by predatory birds or toads, come to resemble, or mimic, distasteful butterflies, which are avoided by these same predators. Drone flies (*Eristalis tenax*) gain protection from predatory toads by mimicking bees that sting and hence are avoided. In such situations the mimicking species benefits by avoidance of predation, whereas the model is little affected, being neither harmed nor benefited by the mimicry.

Examples of protective commensalism include the bird species that build their nests near those of aggressive species, such as predatory birds or certain venomous or predacious insects. In Europe, house sparrows (*Passer domesticus*) and starlings (family Sturnidae) often locate their nests at the sides of an eagle's nest. In North America, house sparrows and grackles (family Icteridae) often nest in close contact to nests of ospreys. In both cases the "protector" species are interested in other prey: the eagles usually attack large rodents or larger birds; the osprey preys on fishes. The cordon bleu weaver (*Uraeginthus*), of Africa, nests directly above colonies of the predatory, fiercely stinging papernest wasps, and the Asiatic woodpecker *Micropternus* often builds nests inside the larger nest complexes of pugnacious Texas shed-builder, or acrobat, ants (*Crematogaster*). These weavers and woodpeckers rarely feed on their insect protectors, and the latter appear never to attack the birds in their nests.

MUTUALISM

Facultative mutualism: protocooperation. Many plant and animal species interact facultatively in ways that are general and indirect but beneficial to both species. These relations, called protocooperation, can be considered the first evolutionary step toward mutualism.

The interactions that occur between soil bacteria and fungi and between them and higher plants growing in the soil are protocooperative. No species is dependent on such an association, but collectively all microflora and higher plants, with associated soil fauna, participate in determining soil composition, structure, and fertility. In the zone of plant roots, soil bacteria and fungi interact with each other, some producing nutrients (metabolites) required by others and all obtaining nutrients ultimately from root exudates and decaying organic matter. Plants benefit from the actions of this microflora by acquiring needed mineral nutrients and carbon dioxide.

Plants also interact protocooperatively with grazing herbivores. Although this relationship constitutes herbivory, grazing herbivores such as deer or cattle maintain a characteristic plant association in their habitual grazing habitats. Removal of the grazing herds soon exposes the range or pasture to invasion by aggressive pioneer plants, including woody shrubs, brambles, and trees. If the grazers are returned, the carrying capacity of the range, much reduced at first, is gradually restored as the grasses and browse plants return. The reason for this interaction is that grasses and browse plants grow from their bases rather than from their tips, so grazers rarely injure the growing crowns of such plants. Shrubs and seedling trees, on the other hand, grow from the tips of their stems and hence are usually destroyed by grazing.

Maintenance of pasturage by grazing animals

The natural control of herbivore populations by natural enemies provides an indirect, protocooperative benefit to plants. These natural enemies (predators, parasites, and pathogens), whose direct influence on herbivorous and other animals has already been described, are some of the major factors that keep herbivore numbers in check. At a community level, this interrelatedness of trophic levels, plant–herbivore–carnivore, constitutes a large-scale protocooperative interaction.

Some of the relationships exhibited between ants and aphids are examples of protocooperation. Benefits are provided to both ant and aphid, but the relationship is often quite loose and facultative (a few such ant–aphid associations are obligatory and are treated in the next section). Generally, the ant member forages for food on trees and shrubs infested with such honeydew-secreting species as aphids, mealybugs, and some scales, collecting the sugary material and transporting it to its nest as food for developing young. In some cases the ant actually stimulates the aphid to secrete honeydew directly into its mouth. Some ant species even protect the honeydew producers from natural enemies, the consequences of which are noticeable: ant-attended trees usually bear much heavier infestations of aphids.

Plants whose flowers are pollinated by insects and birds benefit protocooperatively, particularly when the pollinator is a general one and the plant is attended by many different pollinator species. Many plants, particularly those with colourful, showy flowers bearing nectar glands, are the beneficiaries of cross-pollination accomplished for them by insects. The insect, of course, benefits from the supply of food in the form of pollen and nectar. Honeybee (*Apis mellifera*) colonies are used commercially to ensure pollination of many agricultural crops.

An interesting and apparently widespread form of protocooperation, called cleaning symbiosis, appears most noticeably in birds and fish. The Egyptian plover (*Pluvianus*

aegyptius) picks insect pests from the backs of buffalo, antelope, giraffes, and rhinoceroses and even leeches from the open mouths of crocodiles. The cattle egret in America performs the same function. Certain fishes function habitually as cleaners of other fishes, nibbling away at ectoparasites, wounded tissues, and dead flesh. Even predatory fish search out such cleaning symbiotes and remain passive while they are worked over. Such fish cleaners are often concentrated in fixed sites, called cleaning stations, where other fish come to be cleaned.

Müllerian mimicry, the close similarity in appearance of two or more unrelated species in which each species is more or less distasteful to predators, is a form of protocooperation. Such mimicry presumably benefits all participating species.

Obligative mutualism: interdependency. In many communities the most stable, persistent, and interdependent associations between species are those based on obligative mutualism. Since the association is mandatory for the survival of each participant, coexistence in a given habitat is always required. Thus, ways have evolved to assure the perpetuation of the association from one generation to the next. Some of these means are intricate, involving adaptations in structure, behaviour, or life history so as to guarantee the coexistence of the partners. On the other hand, some mutualistic relationships are maintained simply through the high probability of mutual encounters as a result of the population density or great reproductive powers of one or both mutualists, or symbiotes.

How nature assures continuation of interdependencies

A number of bacteria–protozoan associations occur in aquatic environments. In certain cases, endosymbiotic bacteria, which exist in the cytoplasm of protozoan flagellates, are able to digest cellulose in quantities to provide for themselves and their hosts. Associations of fungi and single-celled algae include the well-known lichens, in which the fungal member penetrates algal cells with feeding tubes or haustoria, thus effecting an intimate physical interconnection. The mutual benefits derived from this association are nutrient exchange, maintenance of water and mineral balance, and resistance to drying or to extreme temperatures. Whether the association is truly mutualistic is difficult to resolve on formal grounds; some authorities contend that the fungi parasitize the algae. Ecologically, however, there is no question that the lichen is far better able to cope with its environment and to invade many more habitats than can either partner living alone.

Single-celled algae also enter into symbioses other than lichen formation, particularly in marine habitats where photosynthesis is restricted. Algae are symbiotic with, among other groups, protozoans, sponges, coelenterates, rotifers, flatworms, mollusks, echinoderms, and tunicates. The algal cell provides oxygen and manufactured food to its partner and, in turn, receives physical support, water, minerals, and a proper environment. Some algae are acquired by ingestion (the more primitive mechanism), whereas others are inherited; *i.e.,* transmitted during cell division of the host (the more advanced mechanism).

The association of algae with a motile coral polyp is accompanied by remarkable behavioral alterations in the host, which reinforce the benefits derived. Free-living larvae (planulae) of certain corals that contain yellow-green algae move toward light, whereas planulae lacking such algae are unresponsive to light. Corals endowed with algae grow more rapidly, take on different shapes, and are much denser than corals without algal members. In Caribbean mangrove swamps, certain algae-containing anemones distribute themselves in zones of intermediate light intensity; bleached anemones become indifferent to light or shade.

The important association of nitrifying bacteria with the roots of certain mostly leguminous plants is well known. Nitrogen fixation—the extraction of nitrogen from the environment—occurs only after the bacteria have invaded the plant roots and stimulated the host to form nodules that encapsulate the bacteria (a response of the host to infection). This association is closely related to parasitism. Nevertheless, the mutual benefits to the interactants are substantial: an environment and nutrients for the bacteria and improved fertility of the soil (by addition of nitrogen compounds) for the plant host.

The unusual case of nodule bacteria

Many insect species (perhaps most) contain microbial endosymbiotes, including bacteria, rickettsias, fungi, yeasts, and protozoans. These organisms provide required nutrients for the host—often vitamins, occasionally digestive enzymes, and sometimes simple foodstuffs such as glucose sugar. The host provides the symbiote with a protected microhabitat containing food, water, minerals, and the proper chemical environment. The endosymbiotes are rarely found in the free-living state, and their hosts are unable to survive in their absence. They may be either extracellular—residing in the mouth, gut, rectum, blood spaces, or excretory tubules—or intracellular—living in the cytoplasm of various cells of the host. Because of the intimacy of such symbioses and the dependent nature of the associations, intricate mechanisms have evolved for the transmission of the endosymbiote from one generation of the host to the next. In some conenoses (*Triatoma*) and seed bugs (family Lygaeidae), intestinal symbiotes are deposited as fecal material with the egg, and newly hatched larvae regain the symbiotes by consuming this material. In some stinkbugs (family Pentatomidae) the symbiotes are smeared on the oviposited egg shell, reinfecting the new hatchling as it emerges from the shell. In the olive fly (*Dacus oleae*) the egg is smeared with bacteria as it passes down the ovipositor, the bacteria penetrating the egg and becoming enclosed in the developing embryo. Finally, there are cases where the symbiotes penetrate the female ovary to invade the egg before it is even deposited; this occurs in some roaches, weevils, and ants.

One of the most notable mutualistic associations is that between certain wood-eating insects and cellulose-digesting protozoans. The wood roach (*Cryptocercus punctulatus*) acquires the intestinal cellulose-digesting symbiotes at an early age, retaining them for life. The termite, also a wood-eater, loses its symbiotes at each molt (shedding of skin) during its immature stages. The termites, however, reinfect themselves: newly molted nymphs ingest anal secretions from nonmolting nymphs and thereby obtain the intestinal organisms.

Numerous insects have become adapted to the use of fungi as specific food resources. These include the wood-boring insects that introduce and maintain fungal growths in their nest galleries and termites and ants that cultivate fungal "gardens" in their subterranean colonies. Various wood-boring insects possess this habit: ambrosia beetles (family Platypodidae), bark beetles (family Scolytidae), and sawflies (family Tenthredinoidea). The adult ambrosia and bark beetles introduce the fungus into the brood galleries they make when they invade a tree. The adult timber borers and sawflies lay their eggs on or beneath the surface of the bark of a tree, introducing the fungus along with the eggs. The fungi then grow in the galleries formed in the wood and provide food for the developing insect larvae.

Insects that cultivate fungi

Myrmicine ants and certain tropical termites also are fungal "gardeners." The fungus-raising termites (subfamily Macroterminae) apparently utilize the fungi to control the localized climate of the nest rather than to obtain food. In termite nests where fungal gardens thrive, both the relative humidity and temperature are maintained at relatively constant, high, favourable levels as a result of the metabolic activities of the fungi. Fungus-culturing ants, however, clearly acquire nutrients from their gardens. Leaf-cutting ants (*Atta*) clip off and carry to their nests pieces of fresh leaves on which the fungi grow. The gardens are established usually in enlarged underground cavities, well ventilated, and tended by a class of workers who remove undesirable fungi and bacteria. All members of the colony feed on strands of the cultured fungi.

Certain insects are obligatory mutualists of the plants they pollinate. The California desert yucca (*Yucca*) can be pollinated only by the yucca moth; the moth, in turn, is wholly dependent on the yucca flower ovary as a place to deposit its egg and develop its young. The common Smyrna fig (*Fiscus carica*) can fruit only after pollination by the fig wasp (*Blastophaga psenes*). This is a much more complicated process, since three different kinds of figs are involved in proper sequence: one for maintaining an overwintering population of wasps, one for producing adult wasps during the growing season, and one (the edible fig)

that requires pollination but is otherwise unavailable for wasp reproduction.

Insects also enter into protective relationships with plants. The acacia ant (*Pseudomyrmex ferruginea*), which lives on and derives its food from the bull-horn acacia (*Acacia cornigera*), protects the acacia from intruding vines, competing plants, and herbivorous insects. The ant depends on the acacia, and the acacia cannot survive without the ant. A concomitant development of this mutualism is that this particular acacia has lost the chemical defenses against insect defoliators that other acacias possess.

NEUTRALISTIC INTERACTIONS

The final category of interspecies association commonly encountered in communities is neutralism, the persistent appearance of two or more species together with neither benefit nor harm accruing to any. It is a commonplace that seems self-evident and without significance; thus, it is often ignored in discussions of symbiosis. Indeed, in some cases—perhaps most—this interaction is trivial or accidental, yet in many communities such interspecific associations are characteristic. Most often, neutralistic associations occur when a common resource for several species is highly localized within the community, as when numerous species aggregate about water holes or streams. Insectivorous birds and predatory insects and rodents are persistent followers along foraging columns of army or driver ants, not to prey on the ants but rather upon the organisms stirred up by the mass advance. Neutralistic associations are characteristic in the distribution of trees in forests. The common situation is mixed stands of different species rather than pure stands of single species.

Simple aggregations

POPULATION EFFECTS OF INTERACTION

At the population level the interaction between different species can have two different effects, particularly when the interaction influences the survival or reproductive success of either species (see also below *Biological populations*). One effect is on the numbers of individuals in the interacting populations. The other is on the qualities, or properties, of the individuals in these populations. The first effect—the quantitative one—is the basis for the balance of nature, the state of rareness or commonness of different species. The second effect—the qualitative one—is an aspect of evolution. Biotic interactions, when influencing the survival of a species, tend to produce changes in such aspects as structure, function, behaviour, and colour of the affected species. Since inherited changes in one species can serve as the basis for change in the other, biotic interactions between two species often lead to mutually induced changes in both, or coevolution.

Ecological aspects. The effect of one population on the numbers of another can be either positive or negative. A positive effect occurs when the survival or reproductive success of individuals of one species, or both, is enhanced and leads to an increase in the numbers of one or the other interacting population. Hence, an insect that consumes pollen in seeking a meal will generally bring about the pollination of the flowers of the host plant concerned; the result is an increase in the reproductive success of the host plant. Algae in such hosts as coral polyps stimulate the growth rate of coral colonies, and the enhanced growth of the coral benefits the algae.

A negative effect occurs when the survival or reproductive success of one or both interacting populations is reduced, usually leading to a reduction in numbers of the affected populations. Negative interactions are those resulting from herbivory, predation, or parasitism, in which the food, prey, or host species suffers and may therefore diminish in abundance, or from competition or mutual antagonism, in which both interacting species may suffer, be reduced in numbers, or even be eliminated from the habitat altogether.

In cases in which the biotic interaction is obligatory for one or both participating species, both species must appear together in the same community for the interaction to persist and for one or both participants to survive. A host species that is dependent upon an endosymbiote to provide necessary dietary components will usually not survive in the absence of the symbiote, and vice versa. A predator in consuming prey thereby reduces the numbers of the prey population to a point at which the prey becomes scarce and more difficult for the predator to find. The predator population then comes into balance with the diminished prey population. Any later increase in the prey results in an increase in the predator, and so on.

Interdependence of interactants' populations

Evolutionary aspects. Since all species gradually change in response to selective factors, biotic interactions can serve as the means by which one of the interacting species brings about evolutionary change in the other. Just as predators can alter the numbers of a prey, and vice versa, evolutionary change in one interacting population can result in the evolutionary response of the other. For example, if a prey, through natural selection, becomes swifter or more agile in escape, a predator may develop, also through natural selection, capabilities for better pursuit. Plants, under the selective pressure of herbivory, often evolve a capability to produce toxic poisons that discourage herbivores. Subsequently, however, the herbivore, especially if it is dependent solely on the particular plant group involved, tends to develop tolerance for or resistance to the poisons. These examples of coevolution illustrate the dynamics of the process, the constant influence of each participant upon the other.

Since two interacting species tend to influence each other's evolution, it is not surprising that evolutionary convergence occurs in dissimilar organisms involved in similar biotic interactions. Convergence is the acquisition, through natural selection, of similar structure, function, or behaviour by unlike and totally unrelated species. One interesting example of convergence is a consequence of insect herbivory. A variety of unrelated insect species attack members of the mustard family (Cruciferae or Brassicaceae), the plant family that also includes the cabbage, brussels sprout, and radish. All crucifers contain chemical irritants or poisons that affect many animals and many microorganisms. Such toxins likely were evolved as chemical defenses in response to herbivore pressure. The insect species that now attack crucifers, however, have not only become capable of detoxifying the poisons but in some cases are even able to utilize them, either as attractants or as feeding stimulants.

Through coevolution, interspecific interactions among mutualists have led to mutual adaptations of unusual complexity. The mutualistic associations between the yucca plant and the yucca moth, the Smyrna fig and the fig wasp, and the bull-horn acacia and the acacia ant have already been described. An even more striking example concerns the mimicry developed in the fly orchid, or bee orchid (*Ophrys insectifera*). The flowers of this unusual orchid—pollinated chiefly by a species of bumblebee—simulate very closely the shape and colour of a female bumblebee, so much so that male bumblebees frequently try to mate with it and in that attempt pollinate the flower.

Mutual adaptations

Polymorphism, as mentioned before, is the occurrence in the same species of two or more different forms, distinguishable by colour, pattern, function, body dimensions, shape of the appendages, or behaviour. Often, polymorphism in one species occurs as a consequence of biotic interactions with another species. Character displacement in overlapping competitor species may present a situation in which the taxonomist often wonders whether he is not dealing with three or four different species rather than just two. Similarly, the discovery of noninterbreeding species, identical in every morphological detail but coming from different regions, also poses problems for the taxonomist. Such so-called sibling species often can be recognized only through differences in the nature of their biotic interactions. An example is the two parasitoid species *Trioxys complanatus* and *T. pallidus*. They are virtually identical in structure, yet each attacks and is specific to altogether different aphid hosts. (P.S.M./Ed.)

Biological populations

Animal and plant populations (the numbers of individuals of given species living in a particular area) tend to be stable—that is, their fluctuations are small compared

with what would be theoretically possible if no limits were placed upon their increase. Only a relatively small number of species are either expanding in numbers or decreasing to extinction. Stability is possible only if numbers tend to rise when low and to fall when high, through the action of density-dependent factors. An attempt is made here to summarize what is known about the critical factors concerned.

Human populations

Human populations are currently unstable. More specifically, they are increasing at an unprecedented rate, owing in part to improved food supplies in many regions of the world and to the eradication of diseases. In the course of a few generations, these advances have significantly reduced infant and child mortality as well as increased the life expectancy of adults. In addition, human beings have characteristically insulated themselves from factors such as natural predation and severe weather that serve to regulate other species' numbers. But a population must eventually cease to increase. It can do so in one of two ways, either through higher mortality (from starvation, disease, or war) or lower natality. While a further increase in food supplies may be desirable on humanitarian grounds, it can at best provide only temporary alleviation.

THE CHARACTERISTICS OF BIOLOGICAL POPULATIONS

Populations consist of individuals of different ages, of two sexes, and of varying heredity.

Age distribution. In some animals, the young are similar to the adults except in size; they eat similar types of food and are subject to similar predators, so their numbers and those of the adults may interact with each other. But in other kinds, the young are extremely different in appearance, live in different habitats, eat different types of food, and are subject to different predators, so that adults and young have no direct influence on each other's survival.

In many small invertebrates, notably many insects, the eggs are laid in one short period of the year, hatch into larvae more or less synchronously, pupate synchronously, emerge as adults together, and die after breeding. Hence at any one time, almost all the population consists of animals of the same age, and the breeding stock of one year is derived wholly from eggs laid the previous year. The situation is slightly more complex in species that have a similar life history but raise two successive generations each year, or, alternatively, that take several years to raise one generation. Most sockeye salmon (*Oncorhynchus nerka*), for example, live for four years. Each individual breeds once and dies; its young move from fresh water to the sea to complete their growth and then return to breed. Thus the populations coming up the rivers of British Columbia in four successive years are nearly separate. The point has commercial importance because in some areas one of the annual populations in a four-year period is much larger than the other three, probably because the average chance of survival is greater for a young fish hatched in a peak year, when they are much more numerous—relative to the numbers of predators—than in other years.

Age distribution in salmon populations

The age distribution (*i.e.,* the proportion of individuals in each age group) is much more complex in most other vertebrate animals and in some of the larger invertebrates. In these groups, most adults breed every year until they die, so that the breeding population at any one time consists of individuals of a mixture of ages. In most stable populations the age distribution remains relatively constant. But in certain fish and marine invertebrates, many more young are produced in some years than in others, and the predominance of a particular year class may persist for some years.

Various small plants, like various small animals, are annuals; they winter as seeds, breed once, and then die. Other species are biennials, and many more, including trees and bulbs, are perennials. Some large trees live much longer than any known animal—as long as several hundred or, in rare instances, 1,000 years.

Sex distribution. In most animals the male-female sex ratio is approximately equal. In various animals with an almost equal sex ratio at birth, one sex may become more common than the other later in life. In many monogamous birds, for instance, adult males are more common than adult females. This is probably because the females take the greater part, and in some species the whole part, in incubating and raising the young and thus are subject to heavier predation than the males. On the other hand, in polygynous birds (*i.e.,* those in which each male has several mates), breeding males are much scarcer than breeding females, perhaps in part because their exceptionally conspicuous plumage and behaviour patterns render them more subject to predation than the females. In one such species, the boat-tailed grackle (*Cassidix mexicanus*), the conspicuous males disappear faster than the less-conspicuous females in winter, perhaps through predation. Among humans, about 51 percent of all births are male, but in modern Western civilization, females survive slightly better than males, so that during adolescence the sex ratio becomes equal, and among adults there are slightly more women than men.

In certain arthropods, notably aphids and water fleas, some generations consist solely of females that lay eggs that develop without fertilization by a male (parthenogenesis), though at other times of the year males are present and fertilize the eggs, which are of a different type. In another group of insects, the social bees and wasps (order Hymenoptera), most of the population consists of workers, which are sterile females. In various rotifers, only females are known. The spoonworm (*Bonellia*) is at first undifferentiated, but environmental factors determine sex; while other marine worms change their sex with age. Earthworms and snails, along with a few other animals, are hermaphrodites, with male and female organs on each individual. Hermaphroditism is, of course, common in flowering plants, both sets of organs usually being on the same flower. Asexual reproduction is common in plants, but in nearly all of them sexual reproduction is also the rule, and only a few species have dispensed with it entirely.

Hermaphroditism

Genetic differences. In addition to the differences linked with sex, individuals of the same population differ somewhat in other genetic characteristics. These constitute the subject matter of population genetics, but it should be noted that the extent of genetic variation may vary with the state of the population; in particular, it may be unusually great in an expanding population because selection is temporarily lightened. Some researchers have argued that the regular population cycles discussed later in this article result from the production, in different years, of individuals of different genetic makeup; but there is no positive evidence for this, and it seems unlikely to be an important cause of cycles (though some genetic change is to be expected in populations that alternately increase and decrease in numbers on a big scale).

NUMBERS AND DENSITY

The numbers of large animals can often be counted directly, and aerial photography has been developed as a method for counting both big mammals and large birds. The numbers of various other animals have been assessed by capturing some of the population, marking and releasing them, and in a further series of captures finding the proportion of marked individuals in the new sample. Given large samples and random means of catching the animals, this "capture-recapture" method can provide accurate estimates of numbers. The results can be misleading, however, if the samples caught are biased as to age or sex, particularly if this varies seasonally, and also if previous capture affects the chances of later capture—*e.g.,* if an animal learns to avoid being caught. In the case of some animals, including small but abundant species, total numbers are normally estimated from those counted in sample areas.

Some animals are much more numerous than others. In general, large animals are sparser than small, because the larger the animal, the more food needed to sustain it; but there are so many exceptions that this generalization is of little use. Predators are much scarcer than their normal prey and are usually larger in size. Among the smallest existing populations are those of various land vertebrates on remote islands, though this is partly the result of destruction by humans. Where human destruction is

checked, even a severely reduced population may persist, as shown by the conservation and recent small increase of the whooping crane (*Grus americana*) on the American mainland. The most numerous large animal in the world is man. It would be extremely hard to determine which is the most numerous small animal in the world.

The population density of a species typically varies in different parts of its range. Population biologists have tended to study animals where they are common—in the middle parts of their range and in optimal habitats—but their relation to their environment may be different where they are sparse. In the few cases in which a population has been studied in two areas, it has been found that its density, natality, mortality, main predators, and even main foods may differ from one area to another. Most animal populations have inevitably been studied under conditions affected by humans. This need not affect the validity of the general conclusions reached, but the proportionate importance of various factors influencing birth rates or death rates may be different in human-modified and in natural habitats.

'he quality etween atality nd ıortality

Natality. The reproductive rate—in mammals the birth rate, in plants the rate of seed production—is the rate at which new individuals are produced. It should be distinguished from the recruitment rate, which is the rate at which new adults (not new eggs or new young) enter the breeding population. Most animal populations are not, on the average, either increasing or decreasing markedly, and in such populations the recruitment rate equals the adult mortality, while the natality or reproductive rate equals the total mortality of eggs, young, and adults.

Some biologists have argued that the equality between natality and mortality is an evolutionary mechanism to prevent overpopulation. This suggests that natality is smaller than it might otherwise be, particularly in long-lived animals. A different view holds that the implication of evolution by natural selection is that those hereditary types (genotypes) that leave most offspring must come to predominate over the rest. In other words, the reproductive rate of each species has been evolved through natural selection to correspond with that which gives rise to most surviving young per adult. This does not mean that each species must produce an immense number of eggs or young. In some this may, indeed, be the most efficient means of reproduction, but others are unable to form sufficient food reserves for many eggs, or cannot find food for a large family without serious risk to themselves, or need large food reserves for the young at hatching, and hence produce few (but often large) eggs or seeds. In this view, natality and mortality are roughly equal because mortality rises when the population is high.

Natality differs significantly in different kinds of animals. An oyster can probably lay a hundred million eggs during its life, and even so complex an animal as the Atlantic cod (*Gadus callarias*) produces some four million. In contrast, an elephant matures late, has one offspring at a birth with an interval of several years between births, and may produce only five young in its life. Similarly, the number of seeds produced by a plant differs greatly in different species.

Birds. The average clutches of different species of birds vary from one to fifteen; the largest in a species that feeds its young is ten, in the blue tit (*Parus caeruleus*). Clutch size is characteristic for each species and presumably depends on hereditary factors, although there are also phenotypic (nonhereditary) variations, the same individual laying clutches of different size under different conditions.

Egg ıumbers

In birds that feed their young, clutch size corresponds to the greatest number of young for which the parents can find enough food without serious risk to themselves. For instance, the European common swift (*Apus apus*) usually lays three eggs. If it is experimentally given four, then, as shown in Table 1, so many nestlings starve that each pair raises, on the average, fewer young than pairs starting with three. Hence, any tendency to lay a clutch of four will be eliminated by natural selection. Similarly, the most frequent clutch size in first broods of the European starling (*Sturnus vulgaris*) and the great tit (*Parus major*) corresponds with that brood size from which the most young survive; in these species, however, the young in larger broods do not starve in the nest but leave it underweight and die soon afterward.

Table 1: Number of Young Raised by the European Common Swift (*Apus apus*)

brood size	broods observed	number hatched	number fledged	proportion raised (percent)	number raised per brood
2	72	144	139	97	1.9
3	20	60	51	85	2.6
4	16	64	27	42	1.7

Source: Observations by C.M. Perrins at Oxford in the years 1958–61. Broods of four were not natural but were made up by adding newly hatched young from other broods, which does not disturb the parents. Some broods of two came from clutches of two, but others from clutches of three, either through failure of an egg to hatch or because a nestling was removed to make up a brood of four.

Many phenotypic variations in clutch size are adaptive. The Eurasian nutcracker (*Nucifraga caryocatactes*), which stores the food for its young the previous autumn, lays four after a good and three after a poor season for nuts. Various hawks and owls that prey on lemmings or voles lay unusually large clutches in the occasional years when their prey is extremely abundant. Regular seasonal variations in the clutch size of songbirds in north temperate regions broadly fit the seasonal variations in the availability of their food. Individuals breeding for the first time lay slightly smaller clutches than older individuals and are also known to be less efficient at finding food. If such phenotypic adaptations were perfect, each species would always lay the clutch best suited to its circumstances, but this would require precise prediction of future feeding conditions for the young.

In hawks, storks, and other large birds, incubation starts with the first egg, so that the young hatch at one- or two-day intervals. This is an adaptation, for if food is sparse, the older nestlings are so much stronger than the younger that they get all of it, while if the food is plentiful, the larger nestlings are sated and the younger are fed; hence the brood size comes quickly to correspond with the feeding conditions at the time, and food is not wasted on young that would starve later. Another adaptation is the growth rate of the nestlings, for the same quantity of food per day can be used to raise a larger brood more slowly or a smaller one more quickly. This is probably why hole-nesting songbirds have larger clutches than those with open nests; the latter suffer such heavy predation that there must be strong selection for rapid growth, but fewer young can then be raised together. A slow growth rate may also be advantageous for species that find difficulty in bringing enough food for even one nestling, such as various large raptors and seabirds.

Forms of adaptation

Another general tendency is for clutches to be larger at higher than at lower latitudes. At the higher latitudes, summer days are longer, so the parents can collect more food per day. Many species breed with the flush of insects after the winter dearth in which many adults have starved, so that food becomes plentiful in relation to adult number and large broods can be fed. Similar conditions hold in tropical savannas, where birds breed with the rains after a long dry season. In tropical evergreen forests, however, food may be available in nearly the same quantities throughout the year; hence, the number of adult birds may stay close to the limit set by food, there being only enough extra food for small broods.

Although ducks (family Anatidae), unlike most birds, do not feed their young, there is a broad inverse correlation between the number and size of the eggs in different species. This suggests that the female's food reserves may be used for more and smaller, or fewer but larger, eggs. Presumably, natural selection has favoured the evolution of a larger egg (and therefore a smaller clutch) in those species in which it is advantageous for the young to be larger, or to carry greater food reserves, at hatching. In a few other birds, including the common swift and the red grouse (*Lagopus scoticus*), the clutch is slightly larger

when more food is available for the laying female, but this is a minor factor.

Natality is also affected by the number of broods raised. Most species raise one a year; many songbirds raise two, some even three, and the Senegal fire finch (*Lagonostricta senegalla*) raises four. At the other extreme, probably because their food is sparse, a few large raptors and seabirds take more than one year to raise a nestling; they breed successfully only once every two years.

The young of some finches and doves first breed when two to four months old; most birds when just under a year old; various water birds, raptors, and others not until several years old; and large albatrosses not until about the age of 10. Such "deferred maturity" (like a small clutch) has been seen as an adaptation to reduce natality in long-lived species. But it can as readily be explained in terms of natural selection, provided that those individuals that defer maturity ultimately leave more descendants than those that start breeding when younger. This will hold if (1) younger individuals are unlikely to raise young and (2) the attempt to breed involves a greater chance of death for the adults. It has been found that in pelicans and herons, which have deferred maturity, immature individuals catch food less efficiently than adults and thus might be unable to feed a brood. Again, in the royal penguin (*Eudyptes schlegeli*), subadults weigh less than adults, presumably because they find it harder to obtain food, and only the heaviest birds breed; presumably this is because big food reserves are needed by the female to form eggs and by the male to take the first long incubation stint. Many species with deferred maturity lay only one egg, which suggests that they find it hard to obtain food for their young. In various seabirds, however, deferred maturity may simply result from an insufficient number of safe nesting sites. Finally, deferred maturity is also found in the males but not the females of various polygynous and promiscuous birds, in which there is such intense competition for mates that younger males may well fail to secure them and in which the elaborate plumage displays probably involve extra mortality from predation. One may conclude that, in birds, the age of first breeding may well have been evolved in relation to the maximum number of young that the adults can raise during their lifetime. Still, the alternative view—that breeding age is adapted to mortality—has not been disproved.

Deferred maturity

Mammals. Mammals produce from 1 to about 12 young at a birth, the latter being the figure for the African multimammate rat (*Rattus coucha*). In some species, individuals breeding for the first time have smaller litters than older parents; in others, such as hares and rabbits, there are seasonal variations in litter size; and in some species litter size varies with latitude. These parallels with birds suggest that similar factors operate in both birds and mammals. It has been shown that, in laboratories, infant mortality in the guinea pig (*Cavia porcellus*) rises with increasing litter size, but the extent to which it may do so in the wild is not known. Similarly, in man, the proportion of stillbirths is lowest for single babies, higher for twins, and much higher for triplets; it is likely that, in primitive conditions with poor food supplies, infant mortality also rises with the number of young at a birth.

Many rodents have several litters a year; many other mammals breed once annually and some only once in two or more years. The female meadow vole (*Microtus arvalis*) is fertile when 25 days old, many mammals when one year old, and other mammals not until several years old. As in birds, larger mammals tend to mature later than smaller ones, but there are many exceptions: at least two species of large whales first breed when two years old, whereas elephants and humans mature in their teens.

Phenotypic effects on maturity

Maturity is also influenced by phenotypic factors. Thus a reduced diet delays maturity in various domestic and wild species; this is because it retards the growth rate, and the species concerned breed at a particular size rather than a particular age. Humans are capable of producing young when quite undernourished, but in the near-starvation conditions of some concentration camps in World War II, many women ceased to menstruate and could not have conceived. Conversely, the marked lowering of the average age at which menstruation begins among girls of late-20th-century Western societies has been attributed to improved nutrition.

Fishes. In fishes, as in ducks, there is a broad inverse correlation between the numbers and sizes of the eggs; this implies that egg production is limited by the food reserves of the female and that larger eggs may in some circumstances be sufficiently advantageous to offset the disadvantage of smaller numbers. In general, marine fishes with planktonic larvae have many small eggs, and the larvae hatch with abundant food around them and so do not need large food reserves; on the other hand, various deep-sea fishes and those that breed in polar coastal waters have fewer and larger eggs, because food is less abundant for the young and it is advantageous for them to be larger at hatching. Some species that have only a few large eggs protect them by making a case for them or even by staying beside them, and a few fishes are viviparous (producing their young alive from eggs retained in the body). This has given rise to the idea that, because the eggs are protected, fewer are "needed"; but, as suggested earlier, natural selection must favour those individuals leaving the most offspring. Presumably, the species in question leave the most offspring when they utilize their food reserves to produce a few well-nourished and well-protected eggs rather than many small and unprotected eggs.

Most temperate-zone fishes, once they mature, breed in successive years until they die. A few species living in lakes that dry up each summer breed once and die, their eggs withstanding the drought. More remarkably, various salmon and eels grow for several years and then undertake a long migration from the sea to fresh water (or conversely in the case of eels), where they breed once and die. The long distance between the growing and breeding grounds of salmon and eels doubtless makes it disadvantageous for the female to breed in successive years. If a female can put virtually all her food reserves into eggs, she can produce more at one time than if she has to survive afterward; even so, most species of fishes can probably lay more eggs in a lifetime by breeding in successive years.

The relation between breeding and food supplies

The fishes, like various mammals, first breed at a certain size rather than at a certain age, and they grow faster and mature earlier when food is more plentiful. Each individual also lays more eggs when it is larger. The young of many marine fishes have grown faster and matured earlier after commercial fishing made their food more plentiful by removing many of the larger fishes.

Freshwater crustaceans. The inverse correlation between egg number and egg size is also found in crustaceans. Thus, the "clutch" (retained in a brood pouch) of the freshwater copepod *Eudiaptomus graciloides* consists of 9 to 18 small eggs in spring and of 4 larger eggs in summer, the total volume being in both cases the same. Food is plentiful for the larvae hatching in spring but not for those hatching in summer, for which food reserves in the egg are therefore advantageous. Similar variations have been found in other species. In the water fleas, as in fish, breeding starts at a particular size rather than a particular age, and the number of eggs laid tends to increase as the body size increases. Water fleas also have asexual generations, and the descendants of one female (the members of a single clone) have the same genetic constitution. Individuals of different clones living in the same environment at the same time may differ consistently from each other in egg number, which suggests that there are hereditary differences in the size of the clutch.

Insects. In insects there is also a general tendency for species in which the larvae hatch with abundant food to lay many small eggs and for other species to lay fewer but larger eggs. This reaches its extreme development in the few viviparous species, of which the tsetse flies (*Glossina*) produce only one young at a birth. In some insects, such as the bordered white moth (*Bupalus piniarius*), the number of eggs laid is proportional to the size of the female's body and hence to the amount of nourishment she received as a larva.

In insects that lay their eggs on a limited source of food for the larvae, such as a flow of sap from a tree, a ripe fruit, a cowpat, or a corpse, there is some adjustment of

Adjustment to food source

the number of eggs laid to the duration of the food supply and to the number of eggs or larvae already present. An insect is likely to leave more survivors if it seeks another source of food than if it lays its eggs in a crowded or deteriorating source. Aphids, for instance, lay more eggs on fresh young leaves rich in nutrients than on old leaves poorer in nutrients, and a predominance of old leaves may inhibit breeding. Comparable adaptations are found in parasitoid insects. Those that parasitize species of similar size to themselves lay only one egg in each host and can also detect whether another individual of their species has already laid an egg there. In contrast, species that are much smaller than their host lay many eggs in it (see above *Biotic interactions*). Finally, in the specialized social insects, such as ants and bees, many sterile workers provide food for, and protect, the larvae hatching from eggs laid by the queens. This significantly increases the number of eggs a queen can lay, since egg laying is thus her main or sole duty; the modification reaches an extreme among those termites in which a queen lays 6,000 to 7,000 eggs every 24 hours for a period that may extend over years.

Flatworms. Small animals are often thought of as having more young than large animals. But a freshwater flatworm in the genus *Polycelis,* which is less than 1 centimetre (0.4 inch) long, produces only 2.7 young per cocoon; and the average natality is only 1.0 per adult because many adults do not breed. This low reproduction rate has been shown to be the result of a limited food supply: in one experiment, natality rose by seven times when much extra food was provided in a pond.

Flowering plants. In certain fishes and insects, as already mentioned, the number of eggs laid by different individuals of the same species varies somewhat with their body size. This variation is far greater in flowering plants. In the corn poppy (*Papaver rhoeas*), for example, the number of seeds produced by different individuals of the same genetic constitution, but grown under different conditions, varies between 1 seed capsule with 4 seeds and 400 seed capsules with 2,000 seeds each. In some trees, the number of seeds produced by the same plant varies enormously in different years, depending partly on climatic conditions and partly on a tendency for a good crop to be followed by a small one.

When species of plants are compared, it is found that the size of the seeds tends to be smaller when there are more of them. Small seeds are advantageous, not only because more of them can be produced but also because they are dispersed more readily in the air. They are characteristic of those species that colonize broken ground. Because their seeds disperse so readily, species characteristic of broken ground, such as many of the daisy family (Compositae), are among the first to reach oceanic islands where, in the absence of competitors, some of them have evolved into trees that have larger and much less dispersible seeds. The seeds of forest trees usually have to germinate in deep shade, so it is advantageous for them to have large food reserves to give them a good start. Some plants have two means of reproduction: a seed requires fewer nutrients from the parent plant than a vegetative shoot but has a much smaller chance of survival than a shoot that remains attached to the parent plant until it is viable.

Advantages of small seeds

Mortality. The death rate varies at each stage of the life cycle. In long-lived species, the "potential age" to which survival is possible is much higher than the average age at death. Hence it is useful to calculate the mortality and the expectation of further life separately for each age group.

If an animal population is kept under favourable conditions in the laboratory, with plentiful food and without enemies, the annual adult mortality is much lower, and the expectation of further life much higher, than in a wild population of the same species. In the laboratory, mortality may be fairly high in the small young, but it is low in large young and in young adults, and, while it rises with increasing age, it is high only in the elderly. At the opposite extreme, among wild birds in natural conditions, the losses of young and juveniles are extremely high. Adult mortality is also relatively high but does not rise with increasing age, nor does the expectation of further life decrease, until the individual reaches senility, as in the case of the lapwing. Relatively few individuals reach that age, however. A similar contrast is found in human beings: mortality in Western civilization is similar in pattern to that of a protected laboratory animal, whereas that of Neanderthals and of people living in Roman times seems to have been similar to the pattern of various wild mammals, though not so extreme in its characteristics as that of wild birds.

Birds. Two methods are available for determining the mortality of birds. The whole of a particular breeding population may be marked, and the number returning in each later year recorded; or the age can be determined of all those birds, banded by many different observers, that are later found dead and reported by others. The former method gives a lower figure for the annual mortality than the latter, probably because breeding birds constitute the most successful part of a population.

Death rate variations among birds

In all the birds so far studied, mortality is much higher in juveniles than adults, and it is particularly high soon after the young become independent of their parents. The highest known adult mortality is 70 percent per annum (which means an expectation of further life, at any given time, of only about 11 months), found in the Senegal fire finch. The lowest figure is 3 percent (expectation of further life, 33 years) in a very small population of the royal albatross (*Diomedea epomophora*) in New Zealand. As already mentioned, in a stable population, mortality equals natality; correspondingly, the Senegal fire finch has four broods each year with a mean of 3.4 young per brood, while the royal albatross raises, at most, only 1 young every two years. The mortality of various birds is summarized in Table 2.

Table 2: Annual Adult Mortality in Various Birds

species	average annual mortality (percent)	expectation of further life (years)
Fire finch *Lagnosticta senegalla*	70	nearly 1
Typical songbirds, ducks, gallinaceous birds, and doves shot for sport (northern Europe, North America)	50 (45–55)	1½
Herons, gulls, terns, waders, plovers, and one species of pigeon, in northern Europe and North America	30–40	2–3
Swifts *Apus melba* and *A. apus*	18–20	5
Penguin *Megadyptes antipodes* (New Zealand), kittiwake *Rissa tridactyla* (Britain), manakin *Manacus manacus* (Trinidad)	10	9½
Gannet *Sula bassana* (Britain) and Australasian shearwaters *Puffinus tenuirostris* and *P. griseus*	6	16
Manx shearwater *P. puffinus* (Britain)	4	25
Royal albatross *Diomedea epomophora* (New Zealand)	3	33

Mammals. The age of some mammals can be told from the annual rings on their teeth or horns, and that of a few others has been determined by marking. In all the wild mammals so far studied, adult mortality rises steeply with age after the first few years. In the Himalayan tahr (*Hemitragus jemlahicus*), a goat introduced into New Zealand, about half the young die in their 1st year; in the 2nd year mortality is only 1.3 percent, after which the figure rises steadily to 38 percent in the 12th and 13th years; omitting the 1st year, overall average mortality is 15 percent. In the male waterbuck (*Kobus defassa*), annual mortality in the 2nd to 4th years is only 2 percent, but in the 8th year it is 14 percent; thereafter it increases steeply, and the overall figure is similar to that in the tahr. Because mortality rises with increasing age, the average annual mortality is a less meaningful figure in wild mammals than in wild birds. In young adult mammals, mortality ranges between 50 and 67 percent in various larger rodents and smaller carnivores; it is between 15 and 40 percent in various deer; but it is apparently only 5 percent in the mountain sheep (*Ovis dalli*) and slightly less in the hippopotamus (*Hippopotamus amphibius*).

Fishes. The ages of fishes that live in high latitudes can be determined from the annual rings on the scales, bones, and the otoliths in their ears. Fish fry have a huge mortality. In the few unfished populations so far studied, adult mortality rises with age as it does in mammals. In

High mortality among young fishes and insects

the Lake Superior whitefish (*Coregonus clupeaformis*), for example, the figure rose in one lake from 8 percent in the 12th year to 17 percent in the 20th year and to 45 percent in the 27th year; in another lake it rose from 41 percent in the 7th year to 59 percent in the 13th year. Such figures show that annual mortality may differ greatly in different populations of the same species. In the Queen Charlotte Islands' herring (*Clupea pallasii*), annual mortality increased from 28 percent in the 6th year to 72 percent in the 11th year. With heavy commercial fishing, however, annual mortality becomes much higher and may be almost independent of age.

Insects. In most insects, as already noted, the adults die after breeding; interest therefore centres on the losses in earlier stages. The spruce budworm (*Choristoneura fumiferana*), the larvae of which eat the balsam fir, may be taken as an example. Each female lays about 200 eggs, of which an average of 19 percent are lost (nearly half through insect parasitoids and one-third through predators), leaving 162. Of the newly hatched larvae, 82 percent are lost (mainly through failure to establish on suitable parts of the food plants), leaving an average of 29.2. Of the surviving larvae, 86 percent are lost (nearly half through insect parasitoids and more than one-tenth through predators), leaving 2.7 to emerge as adults. Of the adults, about one-fifth are lost before they breed, leaving one male and one female to repeat the cycle.

CHANGES IN POPULATION CHARACTERISTICS

Growth of populations. If a few fruit flies (*Drosophila*) are put in a milk bottle with banana mash suitable for their larvae, or a few flour beetles are put in a tin of flour, or a few water fleas are put in a jar of pond water, they increase through breeding, first slowly, then rapidly, then slowly again. If their numbers are plotted against time, as in Figure 11, they form an S-shaped curve with the ends flattened. The initial slow, and later rapid, increase may actually represent a constant rate of change. A population that doubles its numbers in each generation seems to increase slowly from 2 to 4 to 8, and much more rapidly from 256 to 512 to 1,024. This type of population growth is called exponential or geometric.

From W.C. Allee *et al.*, *Principles of Animal Ecology*, 1st ed. (1949); W.B. Saunders Company

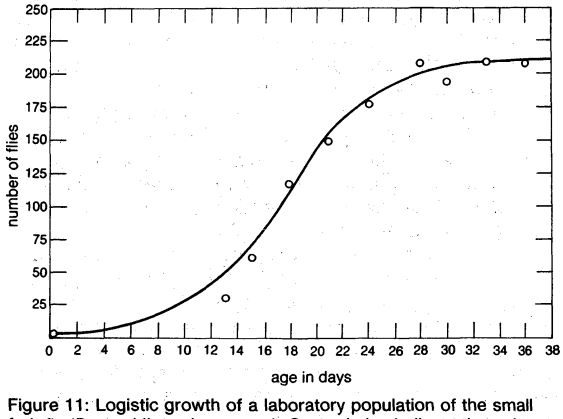


Figure 11: Logistic growth of a laboratory population of the small fruit fly (*Drosophila melanogaster*). Open circles indicate the points resulting from one experiment.

Variable rates of increase

When an animal is introduced by humans to a new region suitable for it, it increases at first slowly and then rapidly, probably in an exponential way. After an animal has become abundant where introduced, its numbers in that area may stabilize while at the same time it is spreading into new areas. In this way the house sparrow (*Passer domesticus*) spread from the Atlantic to the Pacific coasts of the United States in less than 40 years. Similar spectacular increases have been recorded among plants: the water fern (*Salvinia auriculata*) was introduced to Ceylon (Sri Lanka) in 1939 and only 16 years later had put more than 25,000 acres of rice paddy out of production.

Some laboratory populations eventually level off in numbers, but the situation is artificial in that their food supply is constantly renewed. The evidence suggests that, in nature, a successfully introduced population usually declines after a time and does not again reach its first high level, probably because its food supply declines. A few reindeer (*Rangifer tarandus*) introduced to St. George Island in the Pribilovs in 1911 increased to 200 by 1922, fell to 50 in 1926, and fluctuated somewhat below this figure in succeeding years. Another 29 reindeer introduced in 1944 on St. Matthew Island increased to 6,000 by 1963, decreased to less than 50 in the following year, and two years later were practically extinct. They had destroyed their food supply. In the late 19th century the cattle egret crossed from the Old World to the New, where it is increasing and spreading rapidly, but in a habitat (grass pasture with cattle) made by humans. Humans have also caused the rapid increase of various species in their natural habitats by changing their food situation. The herring gull (*Larus argentatus*) has been increasing rapidly on the coasts of Europe and North America; in the latter it has been doubling in numbers every 12 to 15 years since 1900, mainly from feeding on fish waste and on rubbish dumps.

Examples of rapid increase by natural means in natural and undisturbed habitats are far less common because there are usually no ecological niches to be filled. A rapid though temporary natural increase can be observed when numbers have been greatly reduced by a catastrophe, such as a severe winter; the species involved usually attain their former numbers within a few years.

While the rate at which a population can increase is influenced by natality, this is not the main factor determining eventual numbers. Species with a high natality are not likely to be more abundant than those with a low natality. As pointed out by Darwin, the fulmar petrel (*Fulmarus glacialis*) is extremely abundant (and, since he wrote, has greatly increased in western Europe) but has an exceptionally low natality since it lays only one egg a year and requires several years to mature.

Fluctuations in stable populations. All animal populations eventually stabilize at a level around which they fluctuate between relatively narrow limits compared with what is theoretically possible. The extent of the fluctuations, however, does vary significantly; for example, in studies over a period of 40 years of the gray heron (*Ardea cinerea*) in the Thames basin in England, the largest breeding population seen was about double the smallest. At the other extreme, in studies of the bordered white moth (*Bupalus piniarius*) in a German pine forest, the largest population in a period of 60 years was 100,000 times the smallest. But even the largest observed variation is negligible compared with what is theoretically possible. The duckweed (*Lemna minor*) can double in numbers each 2½ days, and at this rate an initial square inch of duckweed would cover one acre in 55 days: yet the waters of the world are not all covered with duckweed. A pair of European robins (*Erithacus rubecula*) producing 10 young a year could theoretically increase to more than 120,000,000 in 10 years. Since even the largest observed fluctuations are negligible compared with what is theoretically possible, animal populations are said to be "stable"; it should be understood, however, that this term allows for fluctuations of the size mentioned, provided that, in the long term, numbers tend to fall when high and to rise when low. The only unstable populations in nature are those recently introduced to a new area and those decreasing to extinction.

Density-dependent factors

Most biologists consider that populations as stable as those found in nature must be regulated by "density-dependent" factors—*i.e.*, factors that tend to reduce numbers when they are high and to allow them to rise when they are low. The contrary view, still advocated by some, is that animal numbers fluctuate irregularly by chance and should not be considered controlled or regulated. The conflict has been partly the result of a difference in interest. Some biologists are interested primarily in annual fluctuations (particularly if a pest is numerous only in occasional years), which may be caused by climatic or other factors acting independently of population density. Others are mainly concerned with the level about which such fluctuations occur, and it is in determining this level that density-dependent factors are important. Density-dependent factors may affect natality, mortality, or movements

of the population; though some populations may be regulated by a single factor, in others two or more factors—for example, food supply and predation—may interact, or the effect of a density-dependent factor may be greatly influenced by that of a density-independent one.

Seasonal fluctuations in numbers. Nearly all animal populations show seasonal changes, often large. In small organisms, some changes of this type are linked with the time needed to complete one generation, or in marine organisms they may be related to tidal cycles; but by far the biggest fluctuations are connected with seasonal change in the environment—from winter to summer at high latitudes, and from the dry to the rainy season in tropical grassland and savanna. This is because nearly all the animals in such places breed at one time of the year, usually in late spring or summer at high latitudes and in the rainy season (or just after it) at low latitudes. They increase significantly in numbers toward the end of the breeding season and thereafter decline until breeding commences the following year. Hence in studying statistics of animal populations based on annual censuses, it should be remembered that between the annual counts there is each year a big and rapid increase in numbers, followed by a corresponding but more gradual decline, and that these seasonal fluctuations in numbers may be much bigger than those between one year and the next.

In a few natural habitats, notably evergreen forests, tropical lakes and oceans, and the deep-sea bottom, seasonal changes are small or almost absent. Even in a rain forest, however, there are small regular differences from month to month in rainfall, vegetation, and the availability of insects; these seem to affect the breeding seasons of many birds. On the other hand, many tropical seabirds that feed on marine organisms on or near the sea surface may breed in any month of the year, which suggests great uniformity in the numbers of their prey and hence also in the food of their prey.

mportance
of the
limatic
actor

Irregular annual fluctuations. All populations fluctuate in numbers to some extent, most fluctuations being irregular. (The few regular cycles are described later.) Large annual fluctuations are usually connected with climatic factors that affect natality or the survival of the young; they are most marked in animals or plants that live for only one year or at least have a high adult mortality, since in such species the young of one year form all, or at least most, of the adult population of the following year. Large annual fluctuations in the production of young (or of seeds) may, however, cause little change in the numbers of a species in which the adults have reached the limit set by space or food supply, especially if the adults have a low mortality. Also, not all annual fluctuations result from variations in the birth and survival of young; in a species living near the northern limit of its range the adult mortality linked with winter cold may vary significantly from year to year.

Irregular fluctuations in numbers have been of special interest to entomologists, particularly in respect to species that remain comparatively scarce for many years and then have a huge outbreak. Examples are the spruce budworm moth, the larvae of which occasionally devastate the forests of balsam fir in northeastern North America, and several other species of moths, the caterpillars of which occasionally strip German pine forests of their needles. It is thought that in most years the numbers of such species are held down by insect parasitoids or predators, but that if climatic factors lead to an exceptionally high production of larvae, especially in two or more consecutive years, the natural enemies do not have time to increase proportionately; thus, the caterpillars then increase rapidly until they are so numerous that they defoliate the trees and eventually die in large numbers of starvation.

Such occasional huge outbreaks are rare. A more usual picture is presented by the annual fluctuations of the red grouse (*Lagopus scoticus*). The adult grouse eat mainly heather. The average clutch is slightly higher and the proportion of surviving young much higher in summers when the heather has a high nutritive value. The actual mechanism of the link between the state of the heather and juvenile survival is not known. In the years when many young survive, the size of the autumn territories is smaller, and the breeding population of the following year is larger. These fluctuations in numbers are proportionately greater in areas of low, rather than high, average density of grouse.

Density-dependent natality. Density-dependent effects may be of two types: direct, in which increasing population exerts a negative effect on factors causing population growth; or indirect, in which increasing population exerts a positive effect, tending to promote further population increase. The latter would not, however, regulate numbers. The natality of many animals is known to be influenced by food supplies, and, since food may be scarcer when the animals are more numerous, it is at least theoretically possible that natality varies inversely with population density. A probable example is the southern elephant seal (*Mirounga leonina*). On the island of South Georgia, where adults are killed in large numbers, cows first have pups at the age of three; bulls first arrive on the breeding ground at the age of four and first hold harems at the age of seven. In contrast, on Macquarie Island, where the animals are protected, cows first have pups at a mean age of six; bulls first arrive at the age of six to eight and breed much later (one marked bull at thirteen or fourteen). The age of first breeding depends on body size, and young seals grow much faster on South Georgia than on Macquarie, presumably because the removal of many adults on South Georgia leaves more food for the young. In the North American mule deer (*Odocoileus hemionus*), the average age of first breeding is also known to vary somewhat with population density. The same is doubtless true of other mammals.

Natality rates of the southern elephant seal

In fish subjected to heavy fishing, the natality of the remainder rises, because with fewer numbers more food is available and the fish grow faster; both the age of first breeding and the number of eggs vary with body size. In the sycamore aphid (*Drepanosiphum platanoides*), and doubtless other species, crowding inhibits breeding; this may reflect reduced food supplies, since breeding is also inhibited by food shortage. The number of eggs produced per female is also density-dependent in the pine beauty moth (*Panolis flammea*), because when the larvae are crowded each forms a smaller pupa and, hence, in the following year a smaller adult, which lays fewer eggs than a larger adult. Again, the number of eggs laid by an insect parasitoid is equal to the number of unparasitized hosts it finds, and the latter number depends in part on the population density of the parasitoid species.

Density-dependent mortality. A simple example of density-dependent mortality was cited earlier for nestling European swifts in which the mortality from starvation is much higher in broods of four than of three. Broods of young are not, however, independent populations. Three examples from such populations follow.

In British Columbia the adults of sockeye salmon come up the rivers to spawn in fresh waters between July and November. The proportion of eggs that fail to hatch rises markedly with an increasing density of adults. Latecomers may expose eggs laid by earlier females; many individuals are forced to lay in suboptimal places susceptible to either drying out or freezing; and crowded eggs may die from fungal infection or shortage of oxygen. These factors are not sufficient to regulate the size of the population, however, since the total number of young that hatch continues to rise, although at a slower rate.

Salmon fry

The small fry suffer heavy predation, chiefly from fishes of other species, on the way from where they hatch to the lake nurseries where they grow. Probably this predation is proportionately less when population density is high. This is evidently because, in years when prey is sparse, the numbers of the predators are held down by the numbers of their prey, and hence in a year with far more fry than usual there are proportionately fewer predators (as their numbers have not yet increased). In such years, small fry have a greater chance of survival than usual, which is probably why, as mentioned earlier, one of the four successive sockeye populations is much larger than the others.

After the fry reach the lake nurseries there is further heavy mortality from predatory fishes, especially in the

first 100 days. This is directly related to population density and therefore affects their numbers. While the immediate cause of their mortality is predation, it is closely linked with food shortage, because when the fry are at higher densities their food is sparser; they grow more slowly than at other times, and because they are smaller they are more vulnerable to predation.

The growth rate of the fry is inversely related to their population density, and so is the average weight of the smolts when they set off for the sea. Moreover, the proportion of smolts returning later to breed is strongly correlated with their weight at departure and hence with their density as fry in the lakes. The population density of the fry is therefore inversely correlated with their subsequent survival at several stages of their life history; its overall effect on the production of adult salmon is summarized in Figure 12. The density-dependent mortality of the salmon fry regulates their numbers, bringing them down when they are high and allowing them to increase when they are low. The causal factors are predation acting in combination with food shortage.

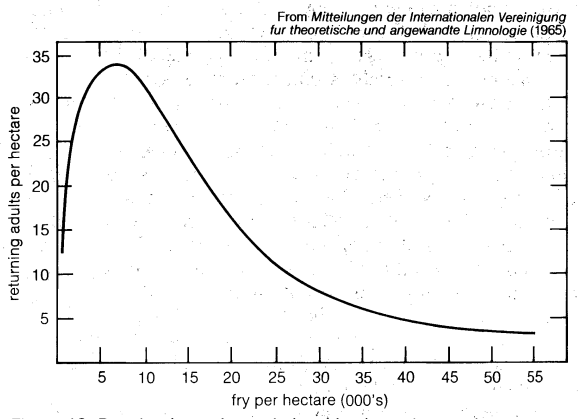


Figure 12: Density-dependent relationships in sockeye salmon (*Oncorhynchus nerka*): relation between density of fry in lakes and that of returning adults.

Food shortage also limits the population density of freshwater flatworms (genera *Polycelis* and *Dugesia*), though in a complex way because these relatively simple animals have the power, when food is short, of feeding from their own tissues. A *Polycelis tenuis* 8 to 9 millimetres (about 0.33 inch) long can survive without food for eight months, during which it shrinks to 1.5 millimetres (0.06 inch). Only the largest individuals of the species breed. Three species studied in North Wales feed on oligochaete worms, one of them also on pond snails. When their food supplies start to increase in spring, the surviving large adults breed. But their numbers are sufficiently close to the food limit to mean that the appearance of many small young produces a shortage, as a result of which many large adults become small and many young fail to grow. A little breeding continues until autumn, at which time the food supply is low and hardly any large individuals remain; the heavy mortality of small individuals from starvation allows the remaining adults to grow larger. As mentioned previously, the experimental removal of two-thirds of the adults from one pond, and the provision of much extra food in another, caused a large increase in natality. In addition, in both ponds the mortality fell and the proportion of big adults increased. In an experiment in which a population was kept without food, the large individuals shrank to medium size, the medium-sized became small, and the small died.

Adult flatworms

Poppies have small seeds with small food reserves. When experimental plots are sown with seeds at different densities, the mortality of the seedlings becomes higher at the higher densities, until it is so high that the addition of further seeds does not produce any further increase in the number of surviving plants and may even cause a decrease. The mortality among densely planted seedlings in this and other species has a number of causes, including higher humidity and hence greater fungal infection, more blanching because the plants partly screen each other from the light, and a great attraction of predators such as wireworms.

In similar experiments involving plants with large seeds, such as the corn cockle (*Agrostemma githago*), each seedling has more food reserves and mortality does not rise with increasing density, so that the greater the density of seeds, the greater the density of plants produced. At the higher densities, however, the rate of growth of each plant is significantly reduced, and above a certain density there is no further increase in the total weight of plant matter per unit of area, or in seed production per unit of area. Corn or maize (*Zea*) is another large-seeded plant in which the rate of growth decreases markedly with increasing density of seedlings; above a certain density, the number of seeds on the plants falls, not merely per individual plant but per unit area, although the total weight of green shoots continues to rise. These findings are of importance in determining the most economic density at which to sow crop seeds.

Predator–prey oscillations. In the examples so far considered, mortality was directly dependent on density, in the sense that the higher the density, the greater the proportion of individuals dying. But in some cases the effect of population density on mortality is delayed, notably when a prey species and its predator (or an insect parasite) each limit the numbers of the other. If, for example, both species are scarce, the prey can increase rapidly through reproduction; but the predator can then also increase in numbers, and the process continues until the predator is so numerous that it begins to reduce the numbers of its prey; this in turn brings down the numbers of the predator, and the cycle starts again (see above *Biotic interactions*). In such a cycle, the relationship between prey numbers and prey mortality is not simple and direct. Instead, the prey mortality is low throughout the phase of increase and high throughout the phase of decrease. A particular density of prey is followed by an increase in numbers at one phase of the cycle and by a decrease at the other. This is because the increase of the predator lags behind the increase of the prey.

Mutual population limitations in predator–prey interactions

Early mathematical analyses suggested that, in a predator–prey interaction of this type, the numbers of both species should oscillate regularly and that the oscillations might increase in amplitude with time. Efforts to test this idea in the laboratory ran into difficulties in achieving ecological balance: the predator tended to increase, exterminate its prey, and then die out. Regular oscillations were finally produced in a closed laboratory population (Figure 13), but only when the situation was complicated by the addition of launching posts to assist the prey to reach undiscovered food sources within the culture and of partial barriers to reduce the movements of the predator between these sources.

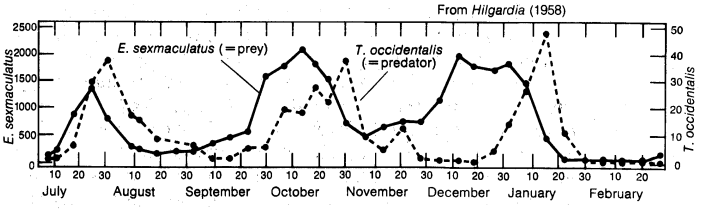


Figure 13: Laboratory study of oscillations in population density of a predator and its prey. The predatory mite (*Typhlodromus occidentalis*) preys upon the orange-feeding, six-spotted mite (*Eotetranychus sexmaculatus*).

Few natural populations of predators and prey oscillate regularly in numbers. In some cases in which a predator has been successfully introduced to control a pest, the predator has at first increased greatly; as a result the prey has decreased greatly, and after it the predator, following which both species have remained extremely scarce. This happened when the Australian ladybird beetle was introduced and successfully controlled the cottony-cushion scale insect in the citrus groves of California. The same result followed when the cactus moth (*Cactoblastis cactorum*) was introduced and successfully controlled the prickly pear (*Opuntia*) in Australia. As of 1971 there were more than 60 cases in which an insect pest had been significantly reduced through the introduction of a

Effects of predators on prey populations

predator or insect parasite and a number of others in which a harmful plant had been similarly controlled by a phytophagous (plant-feeding) insect. These large-scale field experiments show that a single predator is capable of holding the numbers of its prey to a level far below that which prevailed before its arrival. It is generally supposed that the prey continues to survive and to multiply in little pockets, but before it has multiplied far, the pockets are discovered and eliminated by the predator, by which time, however, further pockets have appeared elsewhere. This capacity for movement of both prey and predator can only with difficulty be reproduced in the laboratory.

A phytophagous mite that damaged strawberry plants was maintained by a predator at a level low enough to cause minimal harm. The reverse can occur, however, in populations that lose their predators. An example is the mule deer (*Odocoileus hemionus*) on the Kaibab Plateau in Arizona, following the removal of carnivorous mammals and human hunters. A population of about 4,000 deer in 1906 increased to about 100,000 in 1924 but then fell dramatically to about 20,000 in 1931 and to less than 10,000 in 1939, probably because overgrazing prevented adequate regeneration of the food plants (a delayed density-dependent interaction). Similar effects occurred in other parts of the United States following the removal of the natural enemies of the deer. In parts of Alaska where wolves (*Canis*) are present, mule deer have a light winter mortality from starvation and the range remains in fair condition; but, in parts where wolves are absent, the deer have a heavy winter mortality from starvation and there is marked deterioration in the condition of the range. Thus, the interaction between the wolves and the deer significantly modifies the interaction between the deer and their plant food.

A similar situation is found on Isle Royale in Lake Superior, to which a few moose (*Alces americana*) crossed early in the 20th century. They increased to about 3,000 by 1930, when their food supplies were badly damaged, decreased through starvation to 200 in 1935, rose again to 800 in 1948, and decreased to 500 in 1950. Shortly before 1950, 20 gray wolves (*Canis lupus*) arrived; after that, up to the last published count in 1961, moose numbers stabilized near 600. The wolves tended to take calves or debilitated animals.

Regular, large, and continuing cycles occur in the numbers of some wild mammals. If a cycle is regular, one may be reasonably certain that its cause is simple. The best attested example is that of the lynx (*Felis lynx*), which the records of the Hudson's Bay Company, covering more than a century, show to have a remarkably uniform periodicity of 9 to 10 years. The years of peak numbers are similar in its prey, the varying hare (*Lepus americanus*), and clearly its numbers are determined by those of the hare. This is probably not a simple predator–prey interaction, for there is no evidence that the numbers of the hare are brought down primarily through predation by the lynx. The regularity in the successive declines of the hare suggests a simple interaction with some other factor, and the continuation of each decline for several successive years is characteristic of a delayed density-dependent factor; interaction with plant foods is perhaps involved, but this has not been investigated.

Lemming cycle

The best-known regular fluctuations in numbers are the 4-year cycles of lemmings. The only population so far studied for a long period is that of the brown lemming (*Lemmus trimucronatus*) at Barrow, Alaska. Lemmings are vegetarian, and at the peak in numbers, about every fourth year, there is heavy overgrazing followed by starvation. Grazing transfers most of the nutrient salts from the vegetation to the lemmings, but the nutrients eventually return to the soil and thus, after delay, become available again to the plants, which regenerate only slowly in the cold soil. Once the plants have regenerated, the lemmings increase again, and the cycle is repeated. Removal of the plant cover by the lemmings also exposes them to predatory birds and mammals, but the latter are not the primary cause of their decline.

Several alternative theories of the lemming cycle have been proposed. One is that lemmings regulate their own numbers by emigration. Emigration is particularly characteristic of the Norway lemming (*L. lemmus*), which moves out of areas where food has become sparse to others where food is present. But when all suitable ground has been colonized and overgrazed, emigration is of no further benefit; the lemmings then descend in large numbers from the tundra and may even cross rivers or fjords. Another theory is that physiological stress caused by overcrowding (not starvation) induces metabolic deficiencies in the lemmings, especially in the endocrine system, that reduce fecundity or kill them. A third theory advanced to explain the continuation of the declines after numbers have fallen is that genetic changes occur in the population at different stages of the cycle.

The mammals and birds that prey on lemmings, like the predators of the varying hare, fluctuate in numbers with their prey. The goshawk (*Accipiter gentilis*) and the American red fox (*Vulpes fulva*) are of special interest in this connection because in some areas they prey on lemmings and have a 4-year cycle and in other areas they prey on varying hares and have a 10-year cycle. The numbers of grouse and ptarmigan living in the same areas also fluctuate in parallel, for unknown reasons, but presumably because their numbers are linked with those of the rodents through common foods or common predators.

The examples discussed so far concern simple, or apparently simple, interactions; but most animals, especially those living in natural habitats outside the Arctic, eat diverse foods and have diverse enemies. Because a diversity of factors is involved, their fluctuations are likely to be irregular. The numbers of a species may be regulated in different ways in different parts of its range. For example, much of the mortality of young barnacles (*Balanus*) on the rocks in Scotland is the result of intraspecific crowding, but in San Juan Island in the state of Washington, it is the result of predation by snails (*Thais*); these very different factors govern both the distribution and abundance of the barnacles in the areas concerned.

Movements. Movements often have an important effect on the size of a population, especially in mitigating temporary food shortage or overcrowding, and are of four main types: migration, emigration, dispersion, and dispersal.

Migration. As commonly understood, migration is a regular, seasonal, long-distance, large-scale movement of populations in set directions between two areas, one of which (rarely both) is used for breeding. The term is also used for short movements, in any compass direction, to a special habitat for breeding, such as are found in many types of land, freshwater, and marine animals.

Migratory species

Many birds breed at high latitudes in summer and move for the winter to low latitudes or to coasts or to the sea. Many others have extensive movements between a breeding and a nonbreeding area within the tropics. A few, notably the common redpoll (*Carduelis flammea*) in northern Europe, the Old World quail (*Coturnix coturnix*) around the Mediterranean, and the quelea (*Quelea quelea*) in tropical Africa, almost certainly breed in two areas in succession, but this is not yet fully proved. In mammals, also, certain bats and the caribou (*Rangifer tarandus*) migrate to breed at high latitudes in summer; various whales feed and form fat stores at high latitudes in summer and breed at low latitudes in winter. Various antelope migrate in the grasslands of tropical Africa. Some fishes and at least one squid migrate regularly like birds. In various eels and salmon, each individual moves once in its life from a breeding to a feeding area, returning after several years to the breeding area and then dying. Various northern butterflies breed in the north in summer; their offspring migrate south and breed; and their offspring in turn fly north the following spring, so that, while each individual migrates only once, there is a twice-yearly migration of the population. The monarch butterfly (*Danaus plexippus*), on the other hand, migrates from a northern breeding area to hibernate in the south, the same individuals returning to the northern breeding area the following summer. Other butterflies migrate within the tropics.

The migratory habit enables a species to breed in one area and to survive the winter in another, or to grow large in one area and to breed in another, or even to breed in

two areas at different seasons. Nonmigratory animals may have to survive cold or drought in a resting stage as eggs or pupae, or, if adult, in hibernation or aestivation.

Certain northern birds are partial migrants, some of the population leaving and the rest staying for the winter in the breeding area. In such species, the advantages of leaving or staying must be almost equal; this view is supported by the fact that the proportion migrating increases toward the north of the range, where the winters are more severe. Like other adaptations, migratory movements are anticipatory. If food will become scarce, it is advantageous for an animal to move away before this occurs, and similarly it is advantageous to leave a wintering area so as to arrive on time in the breeding area. Many migrant species have evolved special abilities that enable them to sense time, accumulate fat reserves, and navigate over long distances—strong circumstantial evidence that migration is advantageous to the species concerned (see also BEHAVIOUR, ANIMAL: *Basic behavioral activities of individuals: Migratory behaviour*).

Emigration. Emigration differs from migration in that it occurs at irregular intervals, in a variety of directions, without a set return movement, and in direct response to food shortage or overcrowding. The European red squirrel (*Sciurus vulgaris*) and the Siberian nutcracker emigrate from the taiga in years when they are numerous and coniferous seed crops fail; the common waxwings (*Bombycilla garrulus*) and some other northern birds emigrate when particular berry crops fail; the lemmings, as mentioned earlier, move from the tundra when they have overeaten their vegetable foods; the hawks and owls that prey on lemmings also move when the latter decline; water birds breeding beside shallow lakes in the steppes leave in the occasional summers when the lakes dry out; locusts move when overcrowded, producing a special migratory form; and various aphids, after multiplying significantly on a food plant, produce winged forms that drift away. Some emigrations are combined with migration, in that they may occur at set seasons and in set directions with later return movements, but the numbers participating are much greater in occasional years than in others.

Emigration and natural selection

Emigration is advantageous for any individual that has a greater chance of survival if it moves than if it remains where it is, and hence the habit is explicable in terms of natural selection. It may be specially advantageous in cases of local crowding or local food shortage, when small and inconspicuous movements suffice to take the emigrants to suitable new areas. Even in occasional catastrophic years, when many emigrants die, the chance of survival may be higher for an individual if it moves than if it stays; recoveries of marked birds have shown that some of the emigrants, in fact, survive and return later to their breeding grounds. There is no positive evidence for the alternative idea that the emigrants constitute a doomed surplus that moves out to die in order to save a remnant of the species from starvation; no means are known by which a species can evolve behaviour that benefits unrelated members of the species but is harmful to the participating individuals.

Dispersion. The small and usually inconspicuous movements by which various animals space themselves out, especially for breeding, lead to what is called dispersion. Birds that are solitary breeders tend to be spaced out fairly evenly but with a tendency to be denser where food and other conditions are more favourable; in colonial species, both the spacing and the size of the breeding colonies vary somewhat with the availability of resources. Since adult birds usually breed from year to year in the same place, dispersion is brought about mainly by the younger individuals that are settling for the first time; they first of all seek the preferred habitat of their species, but, if it is densely occupied, they may choose to settle in suboptimal habitats. The latter may give them a better chance of raising young than would otherwise suitable but crowded areas, and therefore suboptimal habitats are occupied chiefly in years of high numbers. Similarly, the young adults of colonial species tend to settle in established breeding colonies but may move elsewhere if these are crowded in relation to resources. In winter, birds join feeding flocks if this helps them to obtain food and leave them if it does not. Similar principles apply to many other kinds of animals.

Dispersal. Dispersal is often used as a general term for animal movements but also in a special sense for the movements of young animals or plant seeds from their birthplace to where they settle down. It has little effect on numbers, although it is an important means by which an increasing species gradually extends its range. In various sedentary or sessile marine animals, however, and in many land plants, specialized larvae or seeds have adaptations for active or passive transport that enable them to settle at long distances from their birthplace (see above *The distribution of organisms: The nature of dispersal*).

(D.L.L./Ed.)

Biological communities, biomes, and ecosystems

A biological, or natural, community consists of two or more interacting and interdependent populations, such as the plant and animal inhabitants of a lake or forest. The major biological communities of the world, such as the Arctic tundra or the Amazon rain forest, are called biomes. Similar biomes considered collectively in terms of predominant vegetation constitute biome types—*e.g.*, tropical rain forest, which comprises biomes in Southeast Asia, Africa, South and Central America, and northeast Australia. A biological community, biome, or biome type considered together with the nonliving components of its environment (*i.e.*, soil and water conditions and climate) is called an ecosystem.

COMMUNITY STRUCTURE

Growth forms and life-forms. The different types of plants represented in a community, in terms of height, form, manner of growth, and kind of foliage, are termed growth forms. Every community includes a number of different growth forms, each represented by a number of plant species. The one or two growth forms that are most important are said to dominate the community. The major growth forms of the community determine its overall appearance or structure. The structure of a plant community on land, as determined by its growth forms, is called its physiognomy.

The plant types, as distinguished in another conceptual system, are termed life-forms. Instead of the many characteristics of plants by which growth forms are distinguished, life-forms are defined by the relationship of the embryonic growing tissues from which new stems and branches will come to the ground surface. The life-forms are (1) phanerophytes—trees and shrubs with growing tissues in buds well above the ground, (2) chamaephytes—low shrubs with buds within 25 centimetres of the ground, (3) hemicryptophytes—perennial herbs with growing tissues at ground level, (4) geophytes—perennial herbs with underground bulbs, buried horizontal stems, etc., (5) therophytes—annual herbs with the growing tissues surviving unfavourable seasons only in seeds, and (6) hydrophytes—aquatic plants. The percentages of the life-forms in a list of plant species for a community or landscape constitute a life-form spectrum. Because the different life-forms represent the ways in which plants are adapted to survive unfavourable seasons, the life-form spectrum expresses aspects of the community's adaptation to its environment. Thus, the phanerophytes predominate in the life-form spectrum of a tropical forest, the hemicryptophytes in temperate forests and grasslands, the therophytes in some deserts, and the hemicryptophytes and chamaephytes in Arctic tundra. Life-form and growth form concepts are most commonly applied to land plants, but similar categories can be defined for animals and aquatic communities.

Stratification of communities

Vertical and horizontal patterns. Growth forms and life-forms relate to stratification—a general characteristic of communities. Plant communities on land show vertical differentiation when the different life-forms and growth forms bear their leaves at different distances above the ground. This stratification is strongly related to light conditions because a given stratum is adapted to the light intensity at its own level and it reduces the light inten-

sity for lower strata. Light intensity in a forest decreases downward; the light reaching leaves of the different strata decreases from full sunlight for the uppermost trees, to approximately 10–50 percent of full sunlight for smaller trees, 5–10 percent for shrubs, and 1–5 percent for herbs. The animal taxocenes of a community also show stratification. In a forest three groups of bird species may be distinguished: those feeding in the tree canopy, those near ground level, and those in foliage of shrubs and low trees between these. Strata involving roots of different plant species, and different animal species, may be recognized from the soil surface downward. Stratification in relation to light and depth occurs also in aquatic communities. Marked vertical differentiation is observed in shore communities and involves water depth and light intensity below tide levels, exposure to air and other factors within the tidal belt.

Many communities also show patterns of horizontal differentiation. On the forest floor scattered patches of different herb species may be observed. In some cases these patches are caused by environmental differences within the forest—differences in light intensity or small undulations in the soil surface, for example, that affect distributions of plant species. In some cases the patches result from growth of plant colonies that spread by underground stems from a common parent; in others, they may result from interactions between species, as in a parasitic plant that forms patches where the roots of its host species occur. Individuals of plant species are not usually scattered at random through the horizontal space of the community; they show clumped or clustered distributions. Many animal populations, both on land and in aquatic communities, also have clumped distributions. A less common condition is spacing that is more regular than a random distribution. Shrubs in some deserts have regular distributions, and singing birds and some other animals appear to occupy definite territories with roughly equal areas, so that individuals or breeding pairs are evenly spaced.

Internal patterning

In the community some species tend to occur together in the same patches, others to occur separately. The tendency to occur together has different causes for different associations of species. Two species may be responding in similar ways to place-to-place differences in light intensity or other environmental factors within the forest; but another pair of species may occur together because one is dependent on the other. The tendency of individuals to occur separately may result because each species responds differently to environmental factors, or because effects produced by one species tend to inhibit or exclude others. Species thus relate differently to environments and to one another within the community, and complex patterns of patchiness and horizontal differentiation of the community can result from these different responses.

Time relations. Communities also show differentiation in time. In plankton different species of algae appear and disappear to give way to other species during the annual cycle of seasons. In a broad-leaved deciduous forest one group of herbs blooms early in spring before leaves are on the trees; other groups bloom in later spring, early or later summer, or fall. Different insect species appear and disappear as the seasons progress, and bird species respond differently to seasons, some migrating and others remaining throughout the year. Organisms also show rhythms of behaviour related to daily time. Some animals are active in the daytime, others at dusk, still others at night. Flowers of different plant species are open at times of day that coincide with the activity of the animals that pollinate them. Many plankton animals migrate upward toward the water surface at night and downward away from intense light in the daytime, but different species have different patterns and extents of vertical migration. The complex rhythms of the tides govern the activities of many shore-dwelling organisms. Plant and animal species of a community thus differ in their relationships to time (see BEHAVIOUR, ANIMAL: *Nature and patterns of animal behaviour: Periodic biological phenomena*).

Interactions. Species within the community also differ in sources of food and ways of interacting. Three major ways of obtaining food may be recognized in a community: photosynthesis, ingestion (eating), and absorption. These three means of nutrition represent major directions of evolution, corresponding, with a few exceptions, to major groups of organisms (plants, animals, and fungi). As treated earlier, animals exhibit a variety of feeding adaptations—*e.g.*, herbivory, carnivorous predation, carnivorous parasitism, and carnivorous scavenging. Many aquatic animals filter a mixture of dead particles and living microorganisms from the water and use some of this organic material as food.

Thus, most species are food specialists, differing from other species in the same community in the source of their food. Species differ in other ways of relating to one another. They may, for example, variously affect one another's environment. An herb growing beneath a tree in a forest must be adapted to the low light intensity of the tree's shade and to the chemical effects produced by the tree in the soil where the herb grows. Some plant species release toxic materials to the soil, preventing some other species from growing nearby. Some species rely on another species for shelter, support, transportation, or concealment. Many plants live as epiphytes (plants not rooted in soil but living among the branches of trees or on other supports) supported by another plant; the lichens on the bark of a tree and the orchids borne on high branches of tropical trees are epiphytes. Many marine animals live on the surface of another animal. A close, sustained living together of two species or kinds of organisms is referred to as symbiosis. Many species, rather than benefiting one another, are competitors. If two species live together and use the same resource, and use of the resource by one limits the growth or population level of the other, the species are in competition. Species may compete for food sources, light, soil nutrients, space, and other resources needed for the support of their populations (see also above *Biotic interactions: Amensalism and antagonism: Competition*).

Symbiosis

Niches and species diversity. Species differ in their positions in a community—in their relation to vertical and horizontal space, time, resource use, and manner of interacting with other species. The position and function of a species in the community in relation to other species is termed its niche. Competition relates to the niche concept. In an experimental culture if two species are in direct competition for the same resources, the population of one species will decline to extinction in the culture. From similar observations on competition and on species positions in communities, the principle of competitive exclusion can be stated: two species cannot occupy the same niche, in direct competition with one another in the same stable community. The species of the community consequently evolve toward niche difference—*i.e.*, different positions in the community by which competition between them will be reduced. A community is, thus, a system of interacting, niche-differentiated species.

Because of niche diversification, many species are able to live together in a community without direct competition. The community's richness in species is referred to as its species diversity, a quality most directly measured in terms of the number of different species found in a sample of standard size taken from a community. Another characteristic of communities is to be recognized, however, in the manner in which relative importances or abundances of species relate to one another. In some communities one species is strongly dominant, and the abundances decrease rapidly in the sequence from the first to the second and third species and beyond. In other communities no species is conspicuously more important than others, and the decline in abundance in the sequence of species is much less pronounced.

Species diversity of communities is affected by evolutionary time (the time during which a community changes and evolves), environmental stability, and favourableness of the habitat. During evolutionary time new species enter a community, each new species adapting itself to a niche different from those of other species in the community. Even though some species become extinct, species tend to accumulate in a community, and diversity tends to increase through evolutionary time. In more rigorous and unfavourable environments, however, fewer species are

able to adapt to the environment itself. Such environments imply that selective forces acting on the species are directed primarily toward adaptation to environmental hazard, rather than toward niche differentiation in relation to other species. Many communities in rigorous environments have low species diversity and high concentration of dominance. In more stable and favourable environments (such as that of the tropical rain forest) selection toward refined niche differences, acting through a long evolutionary time, has made high species diversities possible. Many of the aspects of community structure—*e.g.*, differences among species in vertical and horizontal position, time, food relations, and other interactions—are in part consequences of evolution of niche difference. Complexity of community structure tends to increase through evolutionary time as species of different positions in the community accumulate. Community structure and appearance reflect adaptation to the environment on two levels: (1) the limitations imposed by the environment determine what growth forms can survive in the community, and (2) the niche differentiation among plant species is in part expressed by growth-form differences in stature, form, leaf type, and seasonal relations.

Community structure and niche difference

COMMUNITY FUNCTION

Productivity of organic matter. The existence of the community and of all organisms in it requires energy. Photosynthesis is the source of biological energy for communities, and the creation, through the process of photosynthesis, of organic compounds from inorganic materials by the green plants of a community is termed primary productivity.

Primary productivity. Primary productivity may be measured either in terms of the amount of energy incorporated in organic compounds by photosynthesis or in terms of the dry mass of the organic material produced. In approximate terms 1 dry gram of plant tissue represents 0.4 gram of organic carbon and 4.25 kilocalories of sunlight energy incorporated in organic compounds. The community's gross primary productivity is the total energy captured by green plants (or the mass equivalent of this energy) per unit area and time. The community's net primary productivity is the organic mass (or its energy equivalent) produced by photosynthesis that remains after some of that organic matter has been used in respiration (chemical reactions supporting life processes) by the photosynthesizing plants. In forests 60 to 75 percent of gross productivity may be respired by the plants; the remaining 40 to 25 percent is net productivity. In aquatic communities less than half of gross productivity may be respired by the plants.

Amounts of photosynthesis vary widely in different communities; communities may be grouped in terms of four ranges of net primary productivity as expressed in units of net dry weight in grams per square metre of earth surface area per year ($g/m^2/yr$). The ranges correspond to different environmental and vegetational conditions as shown in Table 3. For a discussion of photosynthetic efficiency in terms of overall energy flow in the biosphere, see above *Energy flow and material cycling.*

The amount of gross primary productivity represents the total energy available to carry on all the biological activities of the community if no organic matter is brought into the community from the outside. The amount of net primary productivity represents the total energy available for use by the community's heterotrophs (nonphotosynthetic organisms that do not manufacture their own food like plants but that must obtain it elsewhere), the animals, bacteria, and fungi. The rate of utilization of net primary production by heterotrophs is important in determining the amount of the community's plant biomass. Biomass (or standing crop) is the amount of organic matter present in a community at a given time potentially available for harvest and use by heterotrophs. In a land community the biomass primarily comprises living and dead tissues of plants. A grassland, for example, might have a stable net primary productivity of 600 grams per square metre per year, and one-third of this plant tissue might be eaten by animals (or decomposed by bacteria or fungi) the first year, one-third of the remainder the second year, and so on. A forest, on the other hand, might produce 1,300 grams per square metre per year and then lose by various means only one-fortieth of it each year. In each case the portion of past production remaining in the community is its biomass. Eventually the grassland biomass will remain steady at about 1,800 grams per square metre and the forest biomass will reach about 52,000 grams per square metre. The forest and grassland differ widely not only in amounts of biomass but in biomass accumulation ratios, the ratios of plant biomass to net annual primary productivity (3 and 40 in this example). Biomass accumulation time and standing biomass are compared for several communities in Table 4.

Biomass

The plant food harvested by herbivores is used for their growth and activity. Herbivores in turn are harvested by a series of carnivores. Such a sequence of organisms along which food is passed—for example from an oak tree (the producer) to a caterpillar (the herbivore) to a warbler (the first carnivore) and finally to a hawk (second carnivore) is called a food chain.

Secondary productivity. When organisms of the community are grouped by position along food chains, the groupings represent trophic levels. Productivities of the trophic levels above the primary level are referred to as secondary productivities. Because many animal species feed upon a number of other species, food chains are linked together into community-wide food webs. A species' place in a food web is an important aspect of its niche. At each trophic level some of the organisms' energy is used in their own life processes, and some organisms die and decompose without being eaten by other animals. The productivity of a given trophic level, therefore, can never be more than a fraction of that of a preceding trophic level. The ratio of energy on a given trophic level to that of the preceding level is known as the efficiency of the given level. For the first and second consumer levels ef-

Food webs

Table 3: Communities Arranged by Net Primary Productivity

net primary productivity*	description and productivity value	some examples of communities with typical productivities in each range
3,000 or more	very high productivity in particularly favourable conditions of water and nutrient availability and temperature	some young tropical forests, salt marshes, coral reefs, rice paddies, sugarcane fields
1,000–3,000	high productivities in generally favourable environments	most forests and some highly productive grassland, some estuaries and nutrient-rich lakes
200–1,000	intermediate productivities in environments in which water or nutrients or temperature are in some respect limiting	many temperate grasslands, semiarid woodlands and shrublands, many lakes and coastal waters of oceans
200 or less	low productivities in environments that are severely limiting in some respect	dry and cold deserts, some nutrient-poor lakes, and many open waters of oceans

*Net dry weight in grams per square metre of earth surface area per year.

Table 4: Comparison of Communities by Biomass

community	accumulation time	standing biomass (kg/m^2)
Forest	10–50	10–70
Shrubland	3–10	2–10
Grassland	2–5	0.2–5
Field of annual herbs	1.0	0.1–1
Plankton	0.01–0.05	0.001–0.04

ficiencies tend to be about 10 percent. Efficiency of producers, comparing their net primary productivity with their energy source in sunlight in the visible range, is much less—*e.g.*, about 1 percent for many forests and even as low as 0.1 percent for some of the open-ocean plankton. Because of the loss of energy on each level, productivity must decrease up the sequence of trophic levels, forming what is known as a pyramid of productivity. In many, though not all, cases, biomasses and numbers of organisms will also decrease up the sequence of trophic levels and for illustrative purposes these can also be visualized as pyramids.

Decomposition of organic matter. Only part of net primary productivity is harvested by animals. The remainder is decomposed by bacteria and fungi, or is transported out of the community, or accumulates as net ecosystem production. In many communities the fraction of net primary production harvested by animals from tissues of living plants is small compared with the fraction utilized after death of the tissues by scavengers, bacteria, and fungi. Less than 10 percent of leaf tissue and less than 1 percent of wood tissue of living trees in a forest may be harvested by animals. The remainder falls to the ground to form the litter covering the soil surface and is utilized by the soil community. The soil community includes animal scavengers that eat dead plant tissue, scavengers on dead animal tissue, bacteria and fungi of decomposition, and animals feeding on these organisms. Although animals contribute to the breakdown of the litter, the bacteria and fungi have the most essential role in reducing dead organic matter to inorganic end products. The major modes of nutrition previously referred to thus appear as the three major trophic parts of the community: producers (photosynthesizing plants), consumers (ingesting animals), and reducers or decomposers (absorbing bacteria and fungi).

The biomass of the reducers may be small compared with that of the consumers and very small compared with that of the producers. The activities of this small mass of reducers are, however, of great importance in community function. The reducers break down almost all remains of dead organisms in the community; food chains normally end in reducers. Decomposition by the reducers in most cases prevents continuing accumulation of dead remains of organisms. When such accumulation does occur, as in the peat formed in bogs, community productivity may be limited by the fact that nutrients are locked up in dead tissues. The reducers thus make possible the steady-state condition of a stable community's biomass. In this steady state the pool of organic matter in the community is relatively constant, while matter is being added to it as net primary production and subtracted from it by animal harvest and decomposition, at equal rates. Decomposition of the dead remains of organisms makes available to the soil or water the nutrients that were contained in those tissues.

Decomposition

Nutrient circulation. The reducers thus make possible the circulation of nutrients in the community. The producers take up inorganic nutrients (including nitrogen, phosphorus, sulfur, calcium, potassium, magnesium, and other elements) from the soil or water. The producers use these for the synthesis of certain organic compounds and to maintain the levels necessary for the composition of protoplasm and the functioning of cells. Animals and reducers obtain these nutrient elements in their food. The nutrients are passed along food chains until released back into the environment by decomposition or excretion. In the nutrient circulation of a forest a nutrient atom or ion may be taken up from the soil into a tree root, transported upward through the tree's conducting tissues to a leaf, taken in by a caterpillar that eats the leaf, consumed then by a bird that eats the caterpillar, until, with the death and decomposition of the bird, it is released back into the soil for renewed uptake by a plant root. Many nutrients are returned from forest trees to the soil by shorter routes—by the fall of dead plant tissues to the litter and decomposition, or by washing down from plant surfaces to the soil in rainwater.

The community's biomass has a marked effect on nutrient circulation. For some elements in some communities, the greater part of the ecosystem's stock of that element is held in the tissues of the plants and a smaller part is free in the soil or water. The amount of phosphate in solution in the water, for example, may be a small fraction of the amount in plankton cells and particles. Tropical forests hold the stocks of some nutrients in relatively "tight" circulation: much of the nutrient stock is held in plant tissues, and nutrients released by leaching or litter decay by fungi are quickly reabsorbed into plants. When a forest is cut or burned, abrupt and extensive loss of nutrients from the ecosystem may occur by erosion or by downward movement of nutrients in soil water. The plankton of the open oceans is low in productivity because the settling of plankton cells and dead particles carries nutrients downward, leaving only low concentrations of critical nutrients in the lighted surface water where photosynthesis occurs. It is thus true both that nutrient resources of the environment affect community function and that community function affects observed nutrient levels in the community and its environment.

Movement of nutrients by water

Water movement transports nutrients between the different communities—plankton, shore, and bottom—of a water body and relates these communities to one another as parts of the water ecosystem. Land ecosystems and lakes and streams receive some nutrients from the outside—by rain, dust, groundwater movement, immigration of organisms, etc.—and lose some nutrients to the outside—by settling, water movement, emigration, erosion, etc. There is, thus, some movement of nutrients between the communities of a given landscape. On a broader scale nutrients are taken up from the ocean surface into the air in the spray from waves, transported over the continents in air currents, and carried downward into land communities in rain. Nutrients from land communities are carried into streams and transported in stream water into the oceans. The ecosystems of the world are linked together by the transfer of nutrients between them into the biosphere.

COMMUNITY SUCCESSION

When gross primary productivity is greater than total community respiration, and net primary productivity is greater than the rate of harvest and decomposition, organic matter accumulates. Coal and petroleum represent surpluses of productivity over respiration accumulated in past geologic time. Such surpluses are bases also of shorter-term growth of communities, as is illustrated in the contrast of a mature, stable forest with a young growing forest shown in Table 5. The accumulating surplus in the young forest may be termed a net ecosystem production. If not exported from the community, net ecosystem production implies growth in the community's biomass. Thus the young forest, with biomass of 10 kilograms per square metre, may in time mature into a forest of 50 kilograms per square metre. The process by which communities grow toward a stable, mature condition is called succession.

Table 5: Contrasts in Productivity and Respiration Between a Mature and Young Forest*

	gross productivity	plant respiration	decomposer respiration	animal respiration	net ecosystem respiration
Young forest	2,650	1,450	580	80	540
Mature forest	3,350	1,950	1,250	150	0

*Net dry weight in grams per square metre of earth surface per year.

Developmental communities. If the forest is destroyed by fire, a new forest is established by developmental communities, which replace one another in sequence (*e.g.*, a field of annual weeds, a meadow of perennial grasses, a community of shrubs, a young forest, a mature forest like the one destroyed). If the soil has not been lost, such a succession replacing a former community is a secondary succession. A succession that develops a new soil in a bare environment is a primary succession. If a forest on a mountain slope is destroyed by an avalanche, a succession on the exposed, bare rock surface may lead back to forest by way of lichens, mosses, grasses, shrubs, and trees as dominants of successive stages. Each stage paves the way for the next stage. Thus the grasses are able to begin growth in the sparse soil formed and collected by the mosses. Growth of the grasses suppresses the mosses and forms a meadow in which a deeper soil develops. This soil permits shrubs to enter the meadow, to grow taller than the grass, and to kill the grass by shading, etc. The soil and shelter of the shrub stage permits trees to enter, grow above, and replace the shrubs. A number of developmental trends—progressive changes in community characteristics—are observed in most successions. During a succession there is usually not only increasing accumulation of biomass but also progressive increase in community height, differentiation into strata of the plant community, and consequent structural complexity, productivity, effect of the community on environment, soil development, stocks of circulating nutrients, species diversity, longevity of dominant organisms, and relative stability of the community. There are exceptions. Productivity and species diversity in particular often decrease during the late stages of a succession and are lower in the stable end community.

Stable communities. The mature, relatively stable community in which a succession finally ends is termed the climax. The climax is characterized not by maximum productivity but by maximum biomass and biomass accumulation ratio and by low or zero net ecosystem production. More broadly, the climax is characterized by steady-state function—that is, a dynamic equilibrium, a condition of relative constancy in an open system. Energy, materials, or individuals may flow through the system, but if the input and output of these are equal, the system can remain stable in its characteristics. Thus in the community steady states may be recognized in these three categories: (1) population balances, (2) energy flow, and (3) materials turnover.

Climax

The individual species populations of the climax are stabilized (relatively, and despite some fluctuation). The rate of addition of individuals by birth (and in some cases immigration) is balanced by rate of loss of individuals by death (and emigration).

Energy intake in photosynthesis is balanced by energy loss in respiration, while the community's pool of energy in organic materials remains relatively constant. Intake of materials by photosynthesis and nutrient uptake is balanced by loss through respiration, decomposition, excretion, and leaching, while the community's biomass and nutrient stock remain relatively constant.

The community is thus (like an individual organism or a cell) an open system through which energy and materials flow. The intimate relation of the community to its environment in this flow provides justification for the concept of ecosystem. Like an individual organism, the community sustains itself by continuing the intake of available free energy in organic compounds. Like the organism, it uses this energy to maintain its function and complex structure. Like the organism, the community grows with increase of mass and structural complexity to a final, mature state. There may, however, be different reasons for growth and maturity. The manner of growth, functional organization, and point of reproductive maturity are determined in the organism by its inherited genetic instructions. There are no corresponding community-wide instructions determining succession and climax. These phenomena are resultants of the manners in which species populations interact and of overall balance of materials and energy flow. The character of the climax is determined by the resources and limitations of environment and by the characteristics of the species that interact and maintain themselves in that environment.

Open system concept

The stability of the climax is suggestive of the homeostasis (the constancy of function and internal conditions) in the organism, but the mechanisms are different. The community has no central regulatory system such as the nervous, circulatory, and endocrine systems provide in higher animals. Yet the community tends to stabilize itself and to return to normal after disturbances. The expression "balance of nature" refers to this condition.

COMMUNITIES IN SPACE

Landscape patterns. Many of the habitats of a landscape are related to one another along environmental gradients (adjacent regions of changing elevation, soil characteristics, surface moisture conditions, etc.). Some of these are interrupted by cliffs or other barriers, but the habitats may be conceived as forming a pattern of environmental gradients. At each habitat, or point in this pattern, a climax natural community may develop. Characteristic species and functions of the community are adapted to the habitat in which they occur. Along a continuous environmental gradient the characteristics of one community change, often smoothly, into those of other communities. The environmental gradient is thus paralleled by a gradient of communities developed in response to the environmental gradient. An environmental gradient comprising many environmental factors that change together along spatial gradients may be termed a complex gradient; its corresponding gradient of communities is a coenocline. The complex gradient and coenocline together form an ecosystem gradient, or ecocline (see Figure 14).

Ecocline

A pattern of climax communities corresponds to the environments of the landscape. Climax communities over much of the area may have been replaced by disturbance and successional communities; but the climax pattern represents the potential maximum development of natural communities in equilibrium with environment for that landscape. A landscape is a pattern of ecosystems related to one another by (1) intergradation as parts of ecoclines, (2) occurrence of populations of the same species in different communities, (3) movements of materials and organisms, and (4) developmental relations of communities to one another (if some are successional).

Climax interpretation. There are three major approaches to interpreting climax communities: (1) monoclimax theory, (2) polyclimax theory, and (3) climax pattern hypothesis.

Monoclimax theory emphasizes the convergence of successions in a given landscape from different beginnings toward similar climaxes. Thus, in a forested area successions on a rocky hillside and a valley bottom will both lead to forests, though these forests may consist of different tree species. Because of this relative convergence, one may consider that in principle all the successional communities of an area could converge on a single (broadly defined) climax community. This hypothetical single climax is termed the climatic climax or monoclimax. If one of the communities of the area is interpreted as the climatic climax, the other stable communities present are termed proclimaxes.

Polyclimax theory recognizes more than one possible climax community. As already observed, the communities of an area form a pattern in which a number of types of stable or climax communities may be distinguished. Many ecologists prefer to grant the occurrence of a number of climax communities in an area, hence they accept a polyclimax interpretation. One of these communities may be considered most typical, or most representative of the general climate of the area, and this community may be regarded as the climatic climax. Other stable communities may be termed topographic climaxes (differing from the climatic climax because of topographic position) or edaphic climaxes (differing because of soil characteristics).

The climax pattern hypothesis allows one to visualize the landscape as a pattern of intergrading communities corresponding to the pattern of environmental gradients. The type of community that forms that largest fraction of the climax pattern and is most widespread in the landscape (if

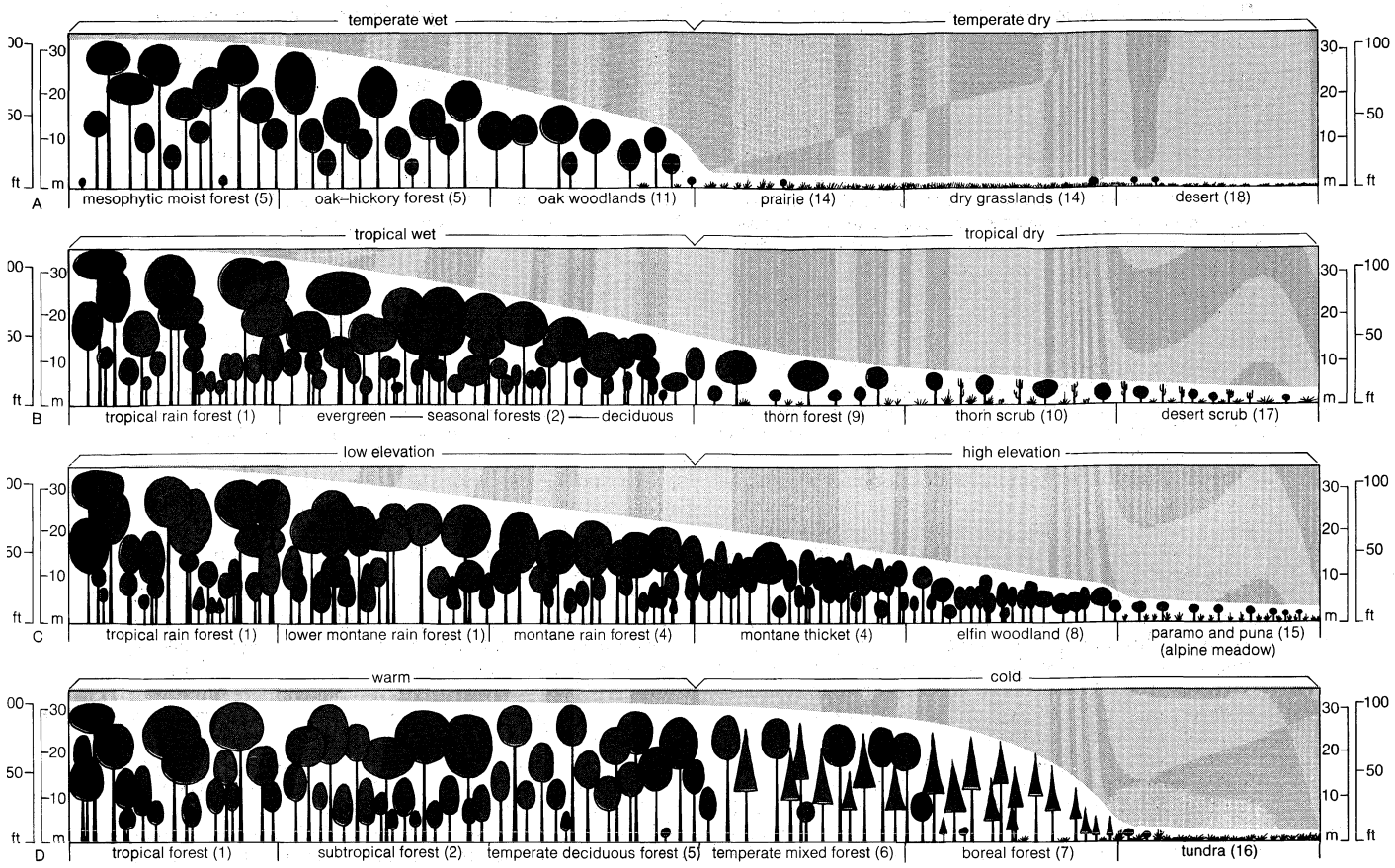


Figure 14: *Four ecoclines representing major climatic gradients.*
(A) From moist to dry climates in the temperate zone. (B) From moist to dry climates in the tropics. (C) From low to high elevations in the tropics. (D) From the tropics northward to cold climates of the far north. Numbers refer to formation types of Table 7.

From (A,D) R.H. Whittaker, *Communities and Ecosystems* (© Copyright, Robert H. Whittaker, 1970), The Macmillan Company; (B,C) Beard, *Ecology (U.S.)* (1955), by permission of the Duke University Press

undisturbed) is then regarded as the prevailing climax or climatic climax.

Community gradients. The appropriateness of the climax-pattern-hypothesis interpretation results from the manner in which species populations and communities relate to environmental gradients. Community samples, including counts of plant populations, can be taken at intervals along an environmental gradient to form a transect (a line along which measurements are taken). The rise and fall of species populations along the gradient can then be observed. Results of interest emerge; especially significant are the symmetrical, bell-shaped form of the majority of such population curves and the scattered positions of the modes or peaks of the species distributions scattered along the gradient. The fact that the curves generally overlap broadly, rather than forming abrupt breaks where one species excludes another, is also apparent. The coenocline, or gradient of communities, may appear as a continuum when species populations are observed along a transect, or when growth forms are observed along a climatic gradient, as in Figure 14.

Sampling communities along transects

Such study supports two principles that are basic to the interpretation of communities. The first is the principle of species individuality, which states that each species is distributed on the basis of its own genetic, physiological, and life-cycle characteristics and ways of relating to environment and other species; consequently no two species will be distributed alike. The second is the principle of community continuity. It states that along continuous environmental gradients, natural communities in general intergrade continuously, rather than appearing as distinct species combinations that give way abruptly, along distinct boundaries, to other species combinations.

Although they are basic, some exceptions to these principles must be noted. Certain pairs of symbiotic species have parallel distributions. Populations of certain animal species exclude one another along a boundary, rather than overlapping broadly. There are some discontinuities between communities that are not produced by soil or topographic changes. The forest edge between forest and grassland is in some areas such a discontinuity, the abruptness of which may be increased by effects of fires that burn the grassland up to the forest edge but do not burn the forest. Steep community transitions, known as ecotones, often show an "edge effect." The transition zone is a distinctive community of high species diversity, combining species of both communities that are bounded by the edge with other species that occur primarily in the edge itself.

Ecotones

Habitat differentiation. Underlying the manner in which species are distributed along a gradient is an evolutionary phenomenon related to the principle of competitive exclusion referred to earlier. Species may escape direct competition either by exploring different niches or by occupying different habitats. Species that are partial competitors (with overlapping niches) evolve toward different locations of their population centres along environmental gradients. Many species that adapt to an environmental gradient and to interaction with one another do so by evolving toward a scattering of population centres along the gradient. Through evolutionary time additional species fit themselves into a gradient of communities (coenocline), with population centres between those of other species. As they do so they tend to narrow the distributions of the other species that are nearest to them along the gradient and that overlap with them in niche. Coenoclines may thus evolve from the condition of relatively few species with broad distributions to that of a larger number of species with narrower distributions.

Gradient analysis. An arrangement of community samples (or species) in relation to one or more environmental gradients or axes, by either direct or indirect means, is an ordination and an essential means of accomplishing gradi-

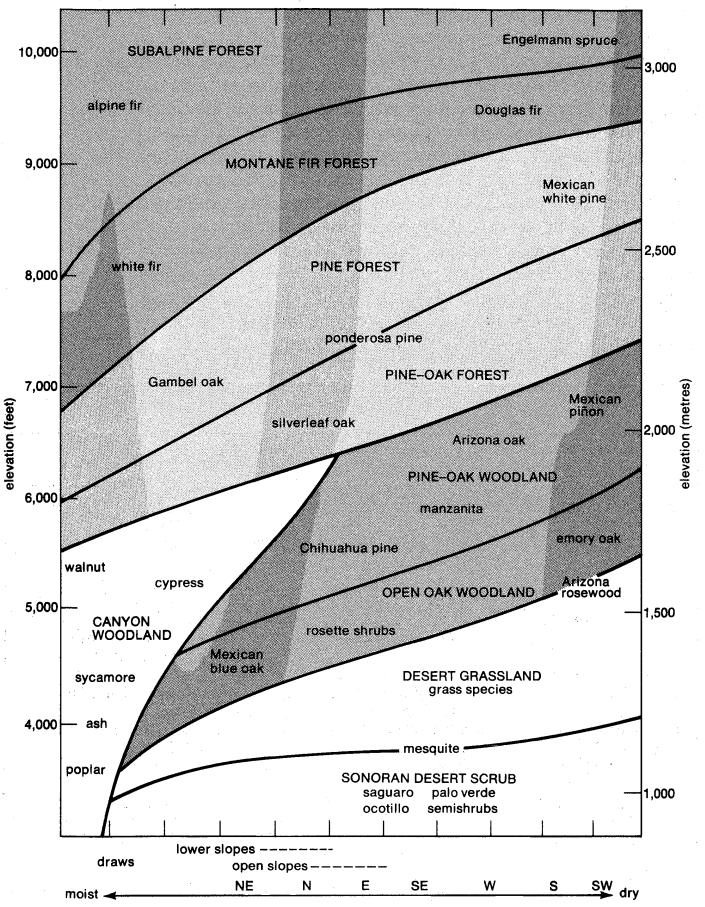


Figure 15: *A mountain vegetation pattern in relation to environment.*
Distributions of major kinds of plant communities are shown in relation to elevation and topographic moisture gradients in the Santa Catalina Mountains of Arizona, U.S. Locations of boundaries are approximate.

From R.H. Whittaker, *Communities and Ecosystems* (© Copyright, Robert H. Whittaker, 1970), The Macmillan Company

ent analysis. Gradient analysis is the study of communities in terms of how gradients of environment, species populations, and community characteristics relate to one another. Through gradient analysis the communities of a landscape may be analyzed and related as a pattern of intergrading communities. In one approach community samples are arranged on a diagram with major environmental gradients as axes as in Figure 15. The samples are classified into types of communities, and boundaries between these may be drawn onto the diagram. Types of communities are then shown in their relations to one another and to the environments of the landscape pattern. The types in Figure 15 have oblique boundaries because of the effects of topography, as well as elevation, on temperature and moisture conditions.

Alternatively, from a set of community samples pairs of end-point samples can be chosen to represent extremes of environmental gradients. Other samples can then be arranged relative to these end-point samples to form what is known as an indirect ordination. This ordination can also be displayed as a diagram that will show the changes in community composition in response to the environment.

Chief aims

Whether direct or indirect, gradient analysis has as its objectives the observation and interpretation of relationships among environments, species, and communities, and the representation of the range of community variation of a landscape as a comprehensible, unified pattern.

COMMUNITY CLASSIFICATION

Bases for classification. In most ecological studies communities are classified, whether or not they are also ordinated. Because of the individuality of species and the continuous intergradation of communities, communities do not form clearly defined natural units of classification. Classification units do not simply emerge from observation; they are created by an ecologist's choice of criteria of classification (*i.e.*, characteristics of communities by which they are to be grouped into classes or units). Application of different criteria of classification to communities of the same landscape will result in the use of different units of classification for those communities. There is consequently no universally accepted "correct" approach to classifying communities, comparable to the system of species and other categories into which organisms are classified.

Table 6: Some Approaches to Community Classification

criteria upon which judgments are based	terms used for the units of classification
Physiognomy	biome or formation
Dominant species of major stratum	dominance type
Dominant species of the different strata	sociation
Species composition using full range of plant species present	association, etc.
Composition of the undergrowth independent of the dominant species of canopy	forest site type
Quantitative comparisons of composition of community samples	nodum
Characteristics of habitats	habitat type
Kinds of communities recognized as belts forming a sequence along a major environmental gradient	life zone, littoral zone
Characteristics of landscapes as patterns rather than particular communities	landscape type

Means of classification

Some of the principal approaches to community classification are given in Table 6. Although these and other approaches are in use, the classifications most widely in use are those that have been based on species composition and on physiognomy, that is to say, the associations and biomes, or formations.

Associations. The system of classification most used for intensive study of particular landscapes has as its basic unit the association. The system considers all plant species

From P. Dansereau, *Biogeography: An Ecological Perspective*, Lieth, *Berichte der deutschen botanischen Gesellschaft*, L.R. Holdridge, *Science*, vol. 105, pp. 367–368 (April 4, 1947) in R.H. Whittaker, *Communities and Ecosystems* (copyright 1970), The Macmillan Company

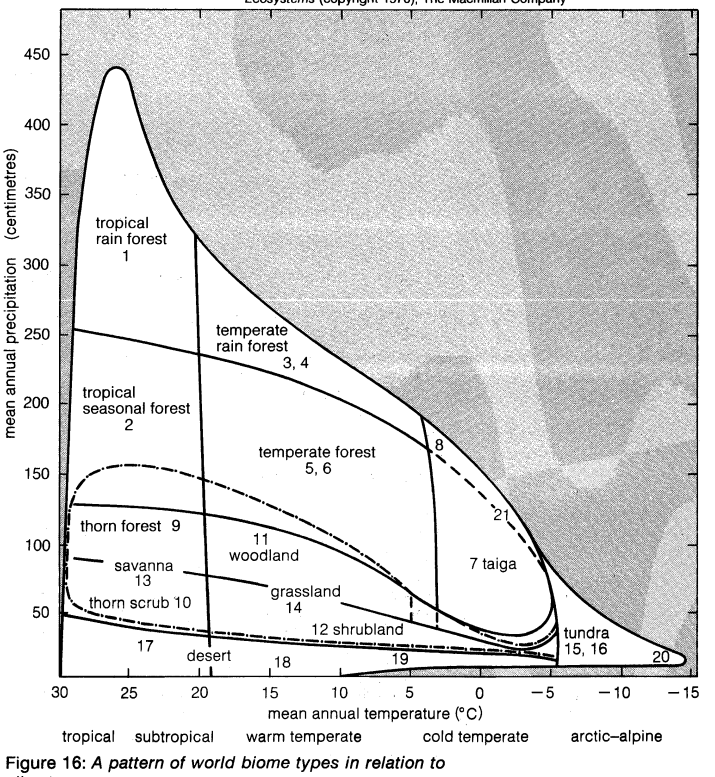


Figure 16: *A pattern of world biome types in relation to climates.*
The numbers refer to biome types listed in Table 7. The dot-and-dash line encloses a wide range of environments in which either grassland or one of the types dominated by woody plants may form the prevailing climax vegetation in different areas.

in community samples as a basis for classifying those samples into units. In grouping samples into associations the ecologist seeks, however, for character species—*i.e.*, those species that are centred in or largely restricted to the communities considered to belong to a given association. These character species are emphasized as means of defining associations, though dominance and other criteria may also be considered. Associations and other units are given latinized names (*e.g.*, the Caricetum curvulae is an association characterized by the sedge, *Carex curvula*). Associations may be grouped into units termed alliances on the next higher level; alliances may be grouped into orders, and orders into classes. These higher-level units may also be defined by character species, which are species that are too broadly distributed to become character species of associations but are centred in or are largely restricted to one of the higher-level units. Vegetation of an area may thus be classified into a formal arrangement in a graded series of units from classes down to associations and lower units. The units below the association (subassociation, variant, and facies) may be defined by differential species and quantitative relations of species. (Differential species are those that tend to be present in one and absent in the other of two units being compared, regardless of their distributions in relation to other units.) The lower-level units make possible effective use of plant communities to indicate habitat properties and the potential of land for agriculture or forestry. This system has been important not only in the development of vegetation mapping but also in the use of vegetation for its indicator value in the

Table 7: Biome Types of the World

biome type	characteristics and environment	distribution
1. Tropical rain forest*	large forests with many tree species, epiphytes (unrooted plants supported on trees), and lianas (vines) in area of abundant rainfall throughout the year	Southeast Asia, Africa, South and Central America, northeast Australia
2. Tropical seasonal forest and monsoon forest	forests adapted to climates with a marked dry season in which some or all leaves are shed	extensive in Southeast Asia; occurs in other tropical areas
3. Giant temperate rain forest	very large forests (trees up to 100 metres tall) in areas of abundant rainfall	northwest coast of the United States, eastern Australia
4. Montane rain forests and thickets	smaller evergreen forests of wet temperate climates of mountains	throughout the tropics and in the Southern Hemisphere
5. Temperate deciduous forests	trees with broad leaves that are shed in winter, mainly in continental, moderately humid climates	Northern Hemisphere mostly, especially in the eastern United States, eastern Asia, and western Europe
6. Temperate evergreen forests	either needle-leaved or broad-leaved evergreen trees or both, in moderately humid temperate climates	all continents (excluding Antarctica)
7. Taiga or boreal forests	needle-leaved forests, mostly evergreen, of cool temperate (subarctic and subalpine) climates	across northern North America and Eurasia and southward at high elevations in mountains
8. Elfin woodlands	dense evergreen thickets with heavy moss and lichen growth in cool subalpine belts of tropical mountains	mountainous tropics
9. Thorn forests and thorn woodlands	communities of acacias and other trees armed with spines, often in open growth, in moderately dry climates	widespread in the tropics
10. Thorn scrub	dense communities of spiny shrubs and small trees, often with succulents, in dry climates of the tropics	widespread in the tropics
11. Temperate woodlands	communities of small trees in open spacing, generally with well-developed undergrowth, in climates too dry or otherwise unfavourable for forests	on all continents
12. Temperate shrublands	diverse shrub communities in temperate climates including as types macchie, or maquis, garigue, chaparral, heath, etc.	on all continents
13. Savannas	tropical grasslands, often with scattered trees, in moderately dry climates or in consequence of soil conditions and fire	extensive in Africa, also in Southeast Asia and South America
14. Temperate grasslands	plains, prairies, steppe, pampas, and veld, moderately dry climates	North America, Eurasia South America, southern Africa, New Zealand
15. Alpine meadows	treeless vegetation above the subalpine forests or thickets in cold mountain climates; sedge or grass meadows, paramo and puna, tussock grasslands	worldwide at high elevations
16. Arctic tundra	treeless vegetation of cold climates north of the taiga, formed by varied combinations of lichens and mosses, grasses and sedges, and dwarf shrubs	northern North America and Eurasia
17. Tropical deserts	sparse vegetation in very dry subtropical climates	North Africa, Asia Minor, southwestern Africa, western South America
18. Warm-temperate deserts	scrub of small-leaved or spiny shrubs or both, in arid climates	south-central Asia, south-western North America, Australia, Argentina
19. Cool-temperate semidesert scrub	shrub communities of sage-brush or other shrubs with grasses in semiarid climates	western United States, interior Asia, Southern Hemisphere
20. Arctic-Alpine deserts	sparse vegetation of widely scattered plants, or lichens only or rock or snow, in extremely cold climates	at high elevations, in the far north in America and Eurasia, and in Greenland and Antarctica
21. Bog†	wet communities with sphagnum moss, mostly in cold climates	worldwide, most extensive in the Northern Hemisphere, especially Ireland and Scotland
22. Tropical fresh-water-swamp forest	evergreen forests of soils seasonally or permanently submerged by water	Amazon basin and other areas of tropical forest
23. Temperate freshwater-swamp forest	cypress swamps and other forests of seasonally or permanently submerged soils	southeastern United States and other humid temperate areas
24. Mangrove swamps	forests of small evergreen broadleaf trees, growing in shallow brackish water or wet soils	tropical coasts and estuaries
25. Marshes	wet soil communities of grasses or grasslike plants	worldwide
26. Marine pelagic‡	plankton (small plants and animals suspended in the water) of the open oceans, fish, marine mammals and birds	worldwide
27. Marine benthos	communities of the ocean bottom, at greater depths limited to microorganisms and animals living on, or rooted in, or burrowing in the bottom mud or ooze	worldwide
28. Marine rocky littoral	animals and algae living on rocky coasts, many of them attached to the rocks and subject to wave action and tidal exposure	marine coasts throughout the world
29. Marine sandy littoral	animals and algae living on or in the sand of beaches and sandy shores in shallow water	marine coasts throughout the world
30. Estuarine and marine mudflat	microscopic algae of the mud surface, animals living in the mud, and plankton of bays and estuaries	marine coasts throughout the world
31. Freshwater lentic	communities of inland lakes and ponds, including in the biome type the plankton, bottom, and shore organisms	on all continents
32. Freshwater lotic	communities of streams and rivers, with organisms of fast-flowing water and rocky bottoms, and of slow-flowing water and sand or mud bottoms, as major subdivisions	on all continents

*Numbers 1–21 refer to biomes shown in Figures 14 and 16. †Numbers 21–25 are biomes determined primarily by wet soils. ‡Numbers 26–32 are aquatic biomes.

intensively used land of the western European continent.

Worldwide classification systems

Biomes or formations. A system of classification is also used for treatment of communities on an extensive scale for whole continents and the world. Treatment of vegetation on a continental scale generally deals primarily with prevailing climax vegetation, uses physiognomy as the criterion of community units and community expression of climate, and employs the formation as its unit of classification. A formation groups together the communities of a given continent that are of similar growth-form structure in response to broadly similar environmental conditions (especially climate). In studies concerned with animals as well as plants, formations are generally termed biomes. Similar formations or biomes occur as convergent responses to similar climates on the different continents. A grouping of convergent formations or biomes of the different continents is a formation type or biome type. Many systems of biome types have been proposed by ecologists and geographers; Table 7 employs one of these many systems for its elaboration of the biome types of the world.

The relations of biome types to climate are represented in Figure 16, which represents the adaptation of world vegetation to world climate. Mean annual temperature and precipitation are considered to be the principal factors determining physiognomy. On the basis of these factors approximate boundaries between major physiognomic types are shown, and the relations of the biome types in Table 7 to these is indicated by their numbers. Community gradients involving some of these formations are illustrated in Figure 14. The pattern of Figure 16 cannot adequately represent (1) the effects of different seasonal relations of temperature and precipitation, notably the contrasts of maritime and continental climates, (2) the effects of fire in determining occurrence of communities dominated by grasses rather than woody plants in many areas, (3) effects of differences in soil, and (4) the continuous intergradations between formations. Despite these simplifications the figure offers, as does Figure 15, a representation of the broad pattern of natural communities on land in relation to climates. (R.H.W./Ed.)

AQUATIC ECOSYSTEMS

An aquatic ecosystem consists of plants and animals interacting with the chemical and physical features of a watery environment from the smallest puddle to the world ocean. And, while conditions in a short-lived puddle differ considerably from those in the ocean deeps, there are many points of similarity imposed upon the communities that inhabit such environments by the unique properties of water. Life forms have adapted to these environmental extremes and to the myriad intermediate aquatic environments available on Earth.

Types of aquatic ecosystems

A common distinction is usually made between seawater and fresh water or between oceans and inland waters. The oceans of the world constitute a deep, interconnecting system of saline water (salty with dissolved minerals) surrounding the continents. This world ocean is characterized by complex circulation and tidal movements. Inland waters, on the other hand, comprise a widely disparate series of bodies of water that tend, generally, to be fresh (*i.e.*, low in dissolved salts) and to drain into the oceans.

Many variations in the aquatic environment are products of organisms themselves; for example, the diatom (one-celled organisms with skeletons composed of silica) population of a lake in May is determined by chemical nutrients, but the chemical composition of the lake, particularly the aspect of it most important to diatoms, is very much the product of diatom activities in the same lake a month earlier. It is impossible to understand the life of waters without taking into account a reasonably large part of the web of circular causality that binds the system together.

Water as a medium for life. Because organisms are themselves aquatic systems, water provides a more kindly environment than does land (and the air above it), the other principal medium of the living world. Yet problems exist that must be overcome. Maintaining the proper internal concentration of water may cost organisms some energy in salty environments. An organism living in a concentrated desert lake or coastal lagoon may be faced with problems of water balance similar to those of its terrestrial neighbours. At the other end of the salinity scale, in freshwater environments, organisms must work to keep an excess of water from flooding in and disrupting their metabolism. In general, though, the problems of gaining and holding an adequate supply of water are very much less in aquatic ecosystems than they are in terrestrial systems.

Water has a high specific heat, which, simply stated, means that water gains or loses a large amount of heat before its temperature changes appreciably. This property of water moderates seasonal, daily, and local extremes of temperature. In even the coldest polar environments, aquatic ecosystems commonly provide refuges beneath the winter ice where the temperature is above the freezing point and far above that of the overlying air. In hot regions, evaporative heat loss rises with increasing temperature under most circumstances and keeps the water relatively cool. As a further consequence of the high specific heat of water, aquatic ecosystems do not show the pronounced microclimatic variability of many terrestrial systems. The contrast between the higher temperatures close to the ground in a sheltered hollow and the lower temperatures a few feet above an exposed ridge—a characteristic of high mountains in summer—is not found in mountain ponds, at least the surface water of which is nearly the same temperature throughout. High specific heat is also an important factor in maintaining the temperature of deep ocean water, even at the equator, always a few degrees above freezing.

The supporting ability of water

Water is much denser than air, so the structural difficulties of maintaining a large organism in it are much less. In water, whales and sharks, much larger than any terrestrial organism—including the dinosaurs—are supported with ease. It is largely this property of water that has permitted the evolution of a style of life—that of the plankton—that has no analogue in air. Plankton is a collective term for all the aquatic free-floating organisms that are buoyant or nearly so. They sink so slowly that normal turbulence keeps most of them in suspension. Birds, bats, insects, and other organisms use the air as a medium of transport, to get from place to place, but typical plankters, by contrast, are completely at home in the water—breeding, feeding, and dying there. Their movements, if any, simply keep them in suspension and do not afford transport.

Water exhibits some important differences from related substances—such as ammonia, hydrogen fluoride, and hydrogen chloride—that theoretically could have provided a fluid medium for the maintenance of life. The hydrogen atoms in the water molecule are on one side, the oxygen atom on the other. The hydrogens being positively charged and the oxygen negatively charged creates a strong polarity and is responsible for the unusual physical and chemical behaviour of water as a solvent for many solids. Some idea of the importance of this capacity in organic evolution can be gained by comparing the relative abundance of the elements in the Earth's crust and in organisms. The elements that accumulate in living things are those that form soluble ions (charged particles) under the conditions of oxidation and acidity normally found in aquatic ecosystems.

The asymmetric arrangement of the hydrogen atoms also creates strong forces of attraction between the molecules of water, so that they tend to clump together. With, on the average, six molecules clumped together, water has much higher values for properties such as viscosity (resistance to flow), surface tension, specific heat (heat required to raise the temperature of one gram of water one degree Celsius), melting point, and boiling point than it would have if its molecules remained separate.

One other important consequence of the asymmetry of

the water molecule is worth particular note. Most substances contract (*i.e.,* become denser) as they become colder. Water is unique in contracting and increasing in density down to about 4° C (39° F), after which it starts to expand again, undergoing a great expansion as it freezes into ice. If it were not for this anomalous thermal expansion of water, ice would form on the bottoms of polar seas and temperate lakes and rivers. Instead of winter water being insulated from further freezing by a surface layer of ice, summer ice would be insulated from melting by a surface layer of water. The ice would gradually increase in volume until all the aquatic ecosystems that are commonly ice-covered in winter became solid ice, and life in them would be impossible.

Physicochemical control of life in water. Water, however, does place limitations on its denizens that air in terrestrial ecosystems does not. Since water is denser and more viscous than air, it is more difficult for an organism to move through water, a matter of consequence to an organism attempting to travel quickly or to move large quantities of water over its gills. Mixing of ingredients occurs slowly in water by comparison with air. Although a productive forest on a clear day or a downtown avenue at rush hour may show departures from the normal levels of carbon dioxide or monoxide, generally the atmosphere is well mixed. In aquatic ecosystems, mixing is so slow that local depletions and excesses of biologically important substances are common.

Mixing of ingredients in water

The mass and metabolic activity of organisms in the water are governed by many physical and chemical factors. For the Earth as a whole, the most important of these is certainly light. If the light intensity throughout the deep ocean water were raised to the level in the upper 10 metres (33 feet), the effects would be to increase the amount of photosynthesis and the activities of plants and animals dependent on it more than any other single change in physical and chemical factors. The fundamental control imposed by the supply of solar energy is so all-pervading that it is usually taken for granted, and investigations are limited to the experimental manipulation of other factors, particularly dissolved chemical substances known to be essential to plant growth. For particular places at particular times, it is possible to enhance photosynthesis by experimental additions of a wide variety of substances. Phosphorus, nitrogen, sulfur, silicon, iron, and other elements may all produce important increases in productivity when used singly and dramatic increases when used in combination; commonly, however, the biggest increase for a single element is obtained with phosphorus. Potassium, a common constituent of agricultural fertilizers, is seldom or never in short supply in aquatic systems.

It is reasonable to expect an enhancement of photosynthesis if the supply of carbon dioxide is increased in a system with a high light intensity. Certain elements required in small amounts and vitamins, known to be essential to the nutrition of many aquatic plants, are occasionally scarce in aquatic ecosystems, but for the most part these substances seem to be available in adequate supply. This is undoubtedly because nutrients are more readily circulated through the water than they are on land.

Animals are affected by anything that influences the plants on which they depend for food. In many aquatic situations, they also must contend with a shortage of oxygen. Many plants and animals are excluded by the extremely acid or extremely alkaline waters of a few lakes in volcanic regions or by the extreme salinity of coastal lagoons or lakes in arid lands saturated with sodium chloride. Whereas the entire range from salt-saturated lakes to fresh snow meltwater provides living space for communities of organisms, the constituent species are seldom able to tolerate more than a small part of that range.

Major divisions of the aquatic environment. The world ocean constitutes the largest ecological subdivision on Earth. Inland waters, which occupy far less area, are generally grouped into two categories on the basis of movement: standing water, or lentic environments; and flowing water, or lotic environments.

The marine division. The world ocean is enormously bigger than any lake, is much older, and has much greater local variation in surface illumination and temperature. Its biota is richer. Its chemical composition is relatively uniform, with about 3.5 percent minerals, mostly salts, and a smaller percentage of organic matter in solution. Unlike the separate lakes that constitute the rest of the standing water of the hydrosphere, the ocean may be thought of as a single system, with no part completely isolated from any other.

The inland divisions. Inland waters include the standing waters of puddles, ponds, and lakes; the critical element is size. Ponds and lakes present many of the same limiting factors and stratification. Smaller bodies of standing water, such as temporary ponds and puddles, are severely limited by drying out and thus maintain transient and quite variable communities.

In common usage many lakes are referred to as seas—*e.g.,* the Caspian Sea, the Dead Sea, and the Aral Sea. The important distinction between a lake and a true sea, such as the Baltic Sea or the North Sea, is that a sea is in open communication with the world ocean; however, a lake, if connected to an ocean at all, is connected through a river that flows in only one direction, from the lake to the ocean. The common criterion of saltiness, which separates most lakes so clearly from the ocean, is not universal. Some lakes are about 10 times as concentrated as open ocean water, and there are parts of the ocean, such as the upper Baltic, that would hardly be considered oceanic on the basis of salinity alone.

Distinction between a lake and a true sea

The other inland waters are the flowing waters, which include springs, brooks, creeks, streams, and rivers. Although springs present a wide spectrum of temperatures and chemical properties, they remain relatively stable—in some cases very stable—with time. The march of the seasons does not affect them to nearly the same extent that it does other ecosystems, so the observations on one aspect of their function may legitimately be compared with observations on other aspects taken at different times. Although the inorganic chemical composition of springs is diverse, springs share the feature of having a small part of their total content of nutrient salts combined into organic matter, either living or dead. During its long sojourn underground, springwater has lost by oxidation most of the dissolved and particulate organic matter it once contained but tends to be richer in available nutrients than most river water and therefore capable of supporting active plant growth as it issues forth into the sunlight. Rivers and streams are much more uniform in chemical composition than are springs, being integrations of the discharge of many individual springs and shallow seepage and surface runoff water. They are highly responsive to the vagaries of weather. From week to week, season to season, and year to year, the discharge and chemical composition of a river may show great variation. During a flood, when the water is laden with silt and spilling over the bank onto the floodplain, a river is a different ecological system than it is during the clear, quiet times of low discharge.

The ocean and its communities

THE OCEANIC ENVIRONMENT

The ocean is by far the largest of all aquatic ecosystems. Once the ocean basins seemed stable and permanent features of the Earth, but it is now known that they are continuously changing in size and configuration at rates that are almost directly observable. Seafloor spreading from the mid-oceanic ridges carries the entire mass of suboceanic sediments, together with everything living on them, on a giant "conveyor belt" that spreads out from the middle ridge, apparently to plunge into the relatively small and localized oceanic deeps.

There is good reason to believe that the ocean basins have been growing steadily since their beginning. The ecosystem consequences of the concept of young and dynamic ocean basins have not been fully examined, but it is likely that they will influence knowledge about the geochemistry of the oceans, and possibly the age, origin, and distribution of the ocean biota as well.

The water of the ocean circulates on a global scale with a pattern that is vital to the organisms (particularly the

Oceanic circulation

deep-sea organisms) dwelling in it. The circulation pattern of the ocean is determined both directly, by observation, and indirectly, by inference from the distribution of water density. The basic data have been obtained from oceanographic vessels and consist of vertical profiles of temperature and salinity, the two main variables influencing the density of seawater. The driving force of oceanic circulation appears to be a combination of wind stress on the water surface, so that the sea is carried along by the atmospheric circulation, and differential heating and evaporation of the surface water at different places in the ocean.

One of the most important features of the entire circulation, with biological consequences for the deep waters of the entire world, is the cooling of surface water in the Arctic North Atlantic to the temperature of maximum density. This water is fully charged with oxygen, then sinks to the ocean floor and spreads out to all the oceans of the world. As it moves south and away from the source area there is some mixing with warmer surface water, so that a perennial thermocline (a middle layer of water with a temperature gradient) is set up at a depth of many hundreds of metres. In the upper part of the mixed layer over this permanent thermocline in temperate climates, a seasonal thermocline develops, with a thin layer of light, warm, surface water riding over the much deeper layer of seasonally mixed water.

Much of the particulate organic matter carried down in the cold, descending water masses of the North Atlantic decomposes as the water sinks. In addition, there is a steady rain of dead organisms and feces from the overlying surface water in all parts of the ocean into the cold, slow-flowing, deep water of North Atlantic origin. The decomposition of organic matter consumes oxygen, but the rate of renewal of the deep water and its oxygen content are so great that the oxygen depletion of the deep water, though easily measurable, is not sufficient to prevent oxygen-requiring animals and bacteria from living in most of the deep sea. In only a few localities, such as the Black Sea, the Cariaco Basin, off Venezuela, and many fjords of heavily glaciated coastlines, as along Scandinavia and Alaska, is the renewal of water restricted enough to produce severe oxygen depletion comparable to that in the hypolimnion (stagnant water below the thermocline) or monimolimnion (water rich in dissolved salts, below the hypolimnion) of lakes.

As in lakes, plankton and detritus sink to the deep water, releasing the nutrient salts needed to sustain plant growth. When this nutrient-rich water wells up to the lighted surface, as it does in a ring around Antarctica, off the coast of West Africa, off California, off Chile, and in various other places, it supports rapid growth of planktonic algae, which, in turn, supports large populations of animals. Many on the major fisheries of the world, such as the Peruvian anchovy fishery, are sustained by the productivity of these localized areas of nutrient upwelling.

Mineralization

The mineralization of plankton and detritus in the sea is never complete. A residue of undecomposed organic matter remains in extremely dilute solution and extends to the greatest depths. Since living creatures are abundant only in the illuminated surface layer of water, the mass of dissolved organic material in the sea greatly exceeds the total mass of particulate matter of all kinds—from phytoplankton (*e.g.*, algae) to whales, both living and dead.

The upper atmosphere mixes with ocean surface water rapidly—in about 10 years—but exchange of gases at deep levels is much slower—on the order of 1,000 years. In other words, an average molecule of deep oceanic oxygen or carbon dioxide left the atmosphere about 1,000 years ago.

The ocean is the great sump, or drain, of the hydrosphere. Its chief contributors, the dissolved and suspended loads of rivers, are relatively easy to measure. Other gains and losses are less easily measured, but the ocean appears to have been in a chemical steady state before the advent of industrialized society. The main buffer in the system consists of reactions between seawater and solid compounds (silicates). This, for example, is the mechanism that buffers the ocean in the face of a continual input of acidic gases from volcanoes, although short-term and local buffering is dominated by reactions with carbonates and bicarbonates.

The ocean receives much of the total of human industrial, municipal, and agricultural wastes. Some persistent synthetic chemicals, such as the chlorinated hydrocarbon insecticide DDT, are carried to the sea, where they accumulate in the fatty tissues of organisms and become concentrated as they are transferred up the food chain. The lead content of surface seawater appears to be increasing rapidly. Oil spills and leaks from oil wells, which formerly had disastrous effects only locally, now appear to be polluting the surface water on a worldwide scale. The ecological effects of these chemical changes are cause for serious concern. If they are reinforced by further expansion of the present regrettable practice of using the deep sea as a dumping ground for noxious substances, such as radioactive wastes and nerve gases, it is possible that the deep parts of the entire ocean will be profoundly modified. Because of the slowness of the deep circulation, serious and irreversible harm to the ocean ecosystem could be inflicted before the effects became noticeable.

(D.A.L./Ed.)

CHARACTER OF OCEANIC COMMUNITIES

The fossil record shows that a rich and diverse marine fauna has persisted on Earth since the beginning of the Cambrian period (about 570 million years ago). The present oceanic fauna—clearly and smoothly traceable from that period—is incomparably richer than the fauna of any lake or river.

Habitat zonation

Life in the open sea occupies the vast extent of the sea bottom (benthic zone) and overlying water (pelagic zone). Assuming that 200 metres (660 feet) below the surface is the upper limit of the habitat of the deep-sea fauna, the surface of the deep-sea region lies under about 92 percent of the total surface of the sea, equivalent to about two-thirds of the total surface of the globe. This three-dimensional region of faint light or complete darkness, with its strange creatures, possesses monotonous uniformity and stands apart from other biological realms on Earth.

The lighted, shallow, populated pelagic zone extends upward and oceanward from the upper boundary of the continental slope. The dark, deep, sparsely populated pelagic zone extends outward and upward from the continental slope and the deep-sea floor and abyssal region. The corresponding zones of bottom-dwelling, or benthic, organisms are the archibenthic (800–1,100 metres, or about 2,600–3,600 feet) and abyssobenthic (below 1,100 metres) zones. The latter zone is sparsely populated and contains some of the most unusual animals known.

Organisms of upper layers of the open sea (Figure 17) are exposed to the full daylight spectrum. Relatively small numbers of organisms inhabit the offshore oceanic waters. The more conspicuous components of the surface population are the Portuguese man-of-war (*Physalia*), the purple sailor (*Velella*), the purple storm snail (*Janthina*), and a nudibranch mollusk, *Gaucus*.

The transient animal life of the surface layers includes many fishes. The relatively permanent surface fauna, apart from such large forms as cetaceans (whales), turtles, and sea snakes, are of two main types: animals adapted entirely to a surface existence partly in air and partly in water and animals that inhabit the immediately subsurface layers. The strictly surface forms usually have floats of various types; some are definite organs of buoyancy, while others are formed of bubbles or trapped air. The only known open-ocean insects are water striders (family Gerridae) that live on the surface film, buoyed by air trapped in the hairs on the body.

The subsurface layers of the lighted open sea contain many of the free-floating, or planktonic, forms, including many kinds of animal larvae. Apart from the blue coloration that is a most striking characteristic of plankton of the open ocean, there are few signs of adaptive characteristics special to this environment. The minute copepods of the family Pontellidae are capable of jumping out of the water to distances of about 15 centimetres (6 inches). The paper nautilus (*Argonauta*) is able to ride on the upper surface of the bells of certain jellyfish.

Seaweed buoyed by bladders is widespread in the oceans of the world and carries with it many associated animals that show adaptations of colour and form suitable to the habitat (interestingly, most are from near the shore, not oceanic).

Plankters such as diatoms and copepods migrate up and down in response to light. Illumination in the upper few metres is too bright for survival of most phytoplankton and zooplankton; hence, this vertical zone of distribution is sparsely populated. The zooplankton of the deeper but still lighted zone in which photosynthesis is active undergo diurnal migrations determined by food and light intensity. Vertical movements of plankton diatoms are related to variations in structure. Increase in spination (the distribution and arrangement of spines) increases the relative surface area and so retards sinking or flotation of the organism when its specific gravity differs from that of the water. Special types of plants adapted to a floating existence have evolved.

Diurnal migrations of zooplankton

In the open sea the chief grazers are the copepod crustaceans, semimicroscopic in size, that feed mainly on diatoms, dinoflagellates, and other microorganisms. Protozoans and the larval stages of larger invertebrates also graze upon the phytoplankton. Some fish, such as herring and menhaden, feed on phytoplankton but are mainly coastal forms.

Bathypelagic animals inhabit the deep, dark waters of the ocean above the bottom (Figure 17). The deep-sea plankton contains no algae except for a few so-called olive-green cells. At 50 metres (160 feet) or less, from 3,000 to 100,000 plankton forms occur per litre. The number of species in the plankton does not decline with depth as strikingly as its number of individuals. Unusual size characterizes some of the deep-sea plankton forms. Microscopic swimming forms such as crustaceans differ significantly from their relatives of shallow seas in usually being luminous and sometimes blind. Other forms have enlarged eyes and long appendages that assist in floating.

Life in dark open water

There is a great diversity of features and forms among the deep-sea fishes. Many of them are blind and small, from about 10 to 30 centimetres (4 to 12 inches) in length. One of the chief characteristics of this deep zone is food

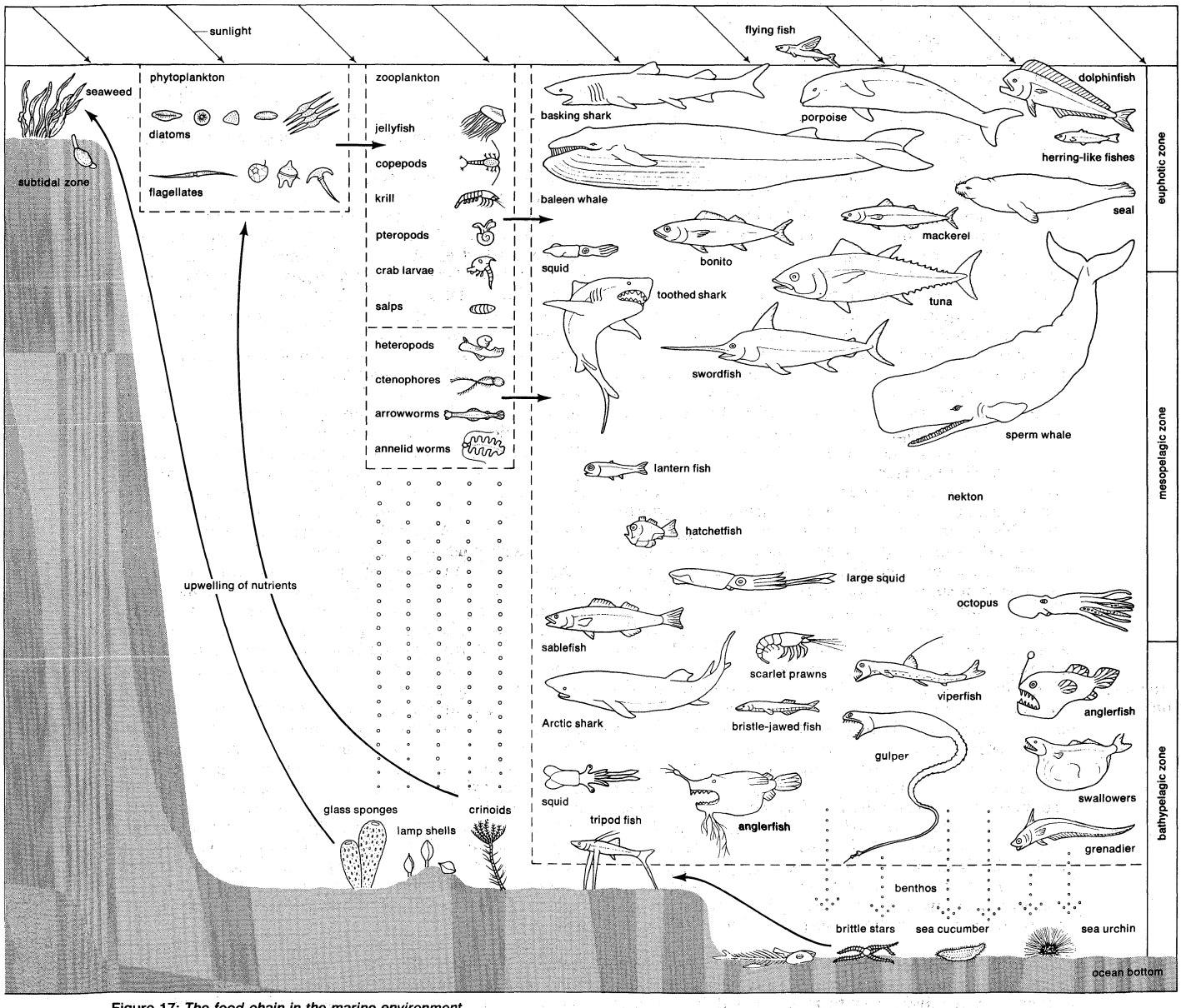


Figure 17: *The food chain in the marine environment.*
All life in the sea is dependent upon sunlight, which is directly converted into food by microscopic plantlike floating organisms (phytoplankton). Phytoplankton in turn is eaten by animal-like plankton (zooplankton). The plankton supports a succession of actively swimming predators (nekton). Organic debris rains down (dots) and provides food for animals at lower depths. Near-shore organisms are supported by land drainage. Coastal upwelling (upward arrows) provides the phytoplankton with nutrients released by decomposition of organic matter on the bottom. (Organisms not drawn to scale.)

scarcity. Another is the size and extension of locomotory appendages of crustaceans as well as fish. The scarcity of food near the sea bottom results in considerable modification of feeding structures: great teeth, enormous jaws, and uncannily extensible stomachs.

Organic material in various states of decomposition rains down from above and is consumed by mid-depth and abyssal detritus feeders. In fact, plant material as such probably never reaches the bottom but goes into solution to be reclaimed by bacteria and other consuming forms. The previously mentioned olive-green cells have maximum distribution below the euphotic zone and are believed to reclaim significant amounts of dissolved organic matter.

Deep-sea, or benthic, animal life shows special features. The abyssal region extends from 100 to 1,500 metres (about 330 to 4,920 feet) to the lowest depths. It is a zone of muddy bottom characterized by low temperatures (usually between 1° and 2.5° C, or 34° and 37° F) and by food scarcity. The population is sparse. Sessile animals (those that tend to remain in a given locale but that are not necessarily fixed) are characteristic and include sponges, stalked polyps, sea anemones, sea lilies, and rhizopod protozoans. Characteristic forms are the echinoderms, such as the sea urchins, some of which have abnormal body forms. There are giant isopods, 15 to 20 centimetres (6 to 8 inches) long, with large eyes. Some blind isopods have long legs and tactile feelers that enable them to move over the muddy ooze and feel for their prey. Deep-sea crabs often have long, slender extremities, sometimes armed with great spikes that keep them from sinking into the ooze.

ADAPTATIONS TO MARINE CONDITIONS

Structural adaptations. Structural adaptations pertain to the rigid conditions of existence at great depths. They relate to acquisition of food, absence of light, stillness of waters, low temperature, and great pressure. The uniformity of these factors is more marked in the deeps than elsewhere in the ocean. The enormous pressure experienced by deep-sea organisms is not countered by any special development. Because of the balancing of external and internal pressures, water pressure is apparently not felt. Pressure changes can be endured if they are gradual enough to permit adjustment. A swim bladder, the hydrostatic apparatus so widespread in surface-living fishes, enables them to float or, by varying the gas content, to rise or sink in the water. A number of deep-sea fishes possess swim bladders. At a depth of 1,000 metres (3,280 feet), with a pressure of 100 atmospheres, the gas content of the swim bladder is compressed to one-fifteenth of the volume it would have at a 10-metre (33-foot) depth.

Aids to buoyancy

The most widespread means of reducing specific gravity among pelagic forms is through the absorption of large amounts of water in, for example, connective tissue, thus producing transparent gellike tissue found in cnidarians (coelenterates), snails, annelid worms, and pelagic cephalopods. Certain decapods are so transparent because of such tissue that print may be read through their bodies. Plankton fishes and eel larvae are also watery and transparent.

More effective than the absorption of water is the storage of lighter materials, such as low-salinity water, fat, or even air. The fluid of vacuoles of radiolarians and certain ctenophores has a lower specific gravity than seawater. The accumulation of fat is widely distributed among pelagic animals, thus lowering the general specific gravity. It is present in protozoans, such as radiolarians that contain oil drops, in crustaceans such as cladocerans and copepods, in some mollusks, and in many fishes that store food in the form of fat in their livers and that have oil drops in their eggs.

The most effective buoyancy means is the inclusion of air or other gases in the body. Jellyfish have air sacs filled with gas from a gas-producing gland. Certain cephalopods, such as the nautilus, have a shell containing air in chambers. The air sacs of fishes have already been discussed.

The relative immobility of deep water favours special animal developments. Fragility of the skeleton and other structural elements of the body is an example. Occasional development of giant forms—sponges and coelenterates more than two metres high and sessile tunicates one metre high—are a product of perpetually still water. The deep-sea environment also produces the largest of the sea urchins, crabs, ostracods, isopods, and sea spiders. This unique environment also permits the degeneration of all supporting and strengthening elements of the body. Thus, the shells of unicellular radiolarians, skeletal spicules of sponges and echinoderms, shells of sea urchins, mussels, and snails, and bones of fish are often so thin and scarce that they could not provide support or protection in disturbed waters. Below 2,500 metres (8,200 feet), the water is so deficient in calcium that the element dissolves readily in it; hence, it is difficult for animals to build up and maintain calcareous supporting structures.

The common factor of all pelagic animals is their independence of the bottom. They have obvious means, already discussed, for maintaining themselves in open water without sinking. In addition, shelled pelagic protozoans have thinner shells than bottom species. The calcium carbonate content of the shells is higher and the size of the pores is larger in the pelagic forms, thus favouring suspension. The shells of pelagic crustaceans are less calcified than those of their benthic relatives; they also have a higher fat content. Pelagic snails have delicate shells or none at all. The pelagic bivalve mollusk *Planktomya* has uncalcified shells. The skeleton of many pelagic fishes is weak, little calcified, or significantly reduced. Reduction of weight in pelagic copepods is achieved by depositing their eggs singly instead of carrying the egg sacs with them.

Effects of light and oxygen content. Lack of light in oceanic depths has important implications for physiological vitamin balance. For example, vitamin D, needed for building bones of vertebrates, can form and function only in the presence of light. Perhaps the lack of this vitamin accounts for the slightness and delicacy of bones of deep-sea fishes and for the markedly distorted forms of many bottom-living fishes.

Lack of vitamin D

Pelagic forms of the deep usually have an adequate supply of oxygen. Because of the low temperatures, these animals have low oxygen requirements. But the benthic occupants of the deep-water sediments require adaptations to meet the problem of low oxygen content. Some bottom dwellers can survive for relatively long periods without oxygen in the liquid environment. Respiratory pigments may provide an oxygen reserve during short periods in oxygen-deficient depths. The shipworm (*Teredo*), which may be found at great depths, makes use of glycogen as an oxygen supply, the glycogen constituting half the dry weight of its tissues. The biology and physiology of inhabitants of such anaerobic sediments are specially adapted to their environment. Besides a biochemical adaptation through the breakdown of glycogen, some animals cease to move. Others have morphological features that enable them to supply aerated water to vital respiratory organs.

Bioluminescence, or the production of light by organisms, is especially characteristic of the deep-sea population. The bottom of the abyssal swarms with light producers, including coelenterates, echinoderms, and annelid worms. As many as 44 percent of the fishes of depths below 900 metres (3,000 feet) are light producers. The light organs of deep-sea fishes vary in size and form. Their source is functionally metamorphosed skin glands or trapped luminous bacteria. Bioluminescence serves to lure prey, to frighten enemies, to enable the sexes to find each other, and to enable species that move in schools to keep together.

Maximum diatom abundance in the open sea is at about 30 metres (100 feet). The daytime concentration of upper zooplankton is at about 125 metres (410 feet), the level varying with location and light intensity. Most of the zooplankton ascend to the surface at night. The echo phenomenon known as the scattering layer is caused by plankton or nekton or both throughout the oceans of the world at from about 400 to 800 metres (1,300 to 2,600 feet). Some of the animal components of this layer undergo diurnal migration of 400 metres or more.

The distribution of euphausid crustaceans has been explained in terms of temperature and salinity. In the Antarctic, there appears to be a regular cycle in which

the immatures and adults are carried into lower latitudes by shallow water movements, and the eggs are returned to high latitudes by deeper, southward-flowing water movements.

Associations. Mutualism, as discussed above in *The zone of life: an overview: Biotic interactions,* is illustrated by the widespread cleaning of fish surfaces of undesirable parasites by associated organisms in search of food. Many species of fish and shrimps, as well as other organisms, clean parasites from other animals. The host is relieved of irritation while the cleaning species gains food. An example is the Spanish hogfish (*Bodianus rufus*), which swims into the mouth of the barracuda (*Sphyraena*) and forages among its teeth for food.

Commensal advantages

The advantage derived by a commensal (species associated with the host) may be the provision of a resting place, shelter, or transport and often of food. Commensals in somewhat permanent contact with their hosts include many animals and plants that simply require physical support. Certain barnacles are found only on the backs of whales, where they benefit by being transported great distances. Pilot fish (*Naucrates ductor*) eat leftovers from sharks, which they follow at close range. Whiting (*Godus merlangus*) and man-of-war fish (*Nomeus gronovii*) shelter among the tentacles of jellyfish. Tiny gobies (suborder Gobioidei) hide under the gill covers of larger fish.

Fish serve as hosts to a large number of parasites, including protozoans and many kinds of worms. They are also plagued by fish lice or copepods. In many deep-sea angler fish, the male lives as a tiny permanent parasite upon the body of the female and obtains his entire nourishment from her blood supply. These species have solved the problem of finding mates in the inky blackness of the deep sea.

PRODUCTIVITY OF MARINE COMMUNITIES

The productivity of the oceans may be judged by the biological oxygen consumption of the water or by nutrient concentration, including soluble inorganic phosphates, nitrates, nitrites, ammonium salts, and silicates. These minerals are consumed in the upper lighted layers, the inorganic phosphorus and nitrogen are regenerated by bacterial decomposition or by dissolved organic debris, and silicates are reformed by the dissolution of planktonic tests. The return of nutrients to the upper photosynthetic layers is accomplished by upwelling and the action of currents.

There are vertical-distributional strata that may differ markedly in nutrient concentrations in any one ocean. Four layers are recognized: (1) the surface, well-mixed, euphotic layer of relatively low concentration that varies little with depth, (2) the layer in which the concentrations increase significantly with depth, (3) the layer of maximum concentration, which has a range of between about 500 and 1,600 metres (1,640 and 5,250 feet), and (4) the thick bottom layers of relatively uniform phosphate and nitrate concentrations and in which silicate content increases significantly with depth.

Differences in the nutrient concentrations of the oceans depend largely on the composition of the deep water masses at their origin and subsequent changes induced by circulation and biological processes. Meaningful comparisons of oceans are thus difficult to make. Available data for the Indian Ocean suggest annual biological oxygen-consumption levels of about 0.45 millilitre of oxygen per litre of seawater in the upper 400 metres (1,300 feet) in the Antarctic shelf area and corresponding levels in the equatorial region of about 1.5 millilitres per litre. The North Indian bottom water levels below 2,000 metres (6,600 feet) average about 0.04 millilitre per litre, and at depths between 600 and 1,200 metres (2,000 and 4,000 feet) biological oxygen consumption levels have been found to range from 1.5 to 2.0 millilitres per litre.

The standing crop, or abundance, of zooplankton in the Antarctic waters in summer has been found to be about 10 times greater than that in the tropical Atlantic waters. At low temperatures, a larger total number of organisms can be supported on the same amount of food; and, in colder waters, nutrients may be supplied more rapidly in relation to their use. In tropical waters with great upwelling of nutrients from the bottom layers, a large standing crop of plankton exists and can support a productive food chain.

Marine biomass

Probably the total marine biomass is far greater than the combined biomass of land and fresh water. Standing-crop data are commonly reported as milligrams of carbon (C) per cubic millimetre (or other unit of volume). The range is commonly 10 to 1,000 milligrams of carbon per cubic metre. Values are usually derived from measurement of phytoplankton pigments, especially chlorophyll *a*. Values for open-ocean areas may range around 5 to 150 milligrams of carbon per cubic metre.

The entire oceanic productivity falls in the range 1.6 to 15.5×10^{10} tons of carbon per year, compared with 1.9×10^{10} tons for the land. About 58×10^6 tons of fish and shellfish are produced annually by the oceans, an amount believed to represent only about 0.03 percent of the total amount of organic material estimated to be produced annually in the sea. Most of the fishery yield comes from waters over the continental shelves and within the 200-metre depth range where nutrient concentrations are high.

This productive area represents only about 3 percent of the total ocean surface. The deep, open-ocean areas are comparable to continental deserts. The reduction in biomass from shallow coastal waters to oceanic depths is about 10^6 for a benthos (organisms living on or near the bottom); and for the plankton it approaches 10^4. Undoubtedly, this low rate is associated with and indeed reflects the small amount of detritus and other food material reaching the bottom at great depths. The low temperatures of the ocean deeps reduces respiratory rates and growth rates. Maintenance demands in terms of food supply are correspondingly low. In summary, the pelagic and benthic zones of the deep open ocean are sparsely populated and the least productive of all oceanic areas.

Instances of productivity variations among the Atlantic, Pacific, Indian, and Arctic oceans are associated with latitudinal temperature differences and upwelling of deep water rich in nutrients. The maximum–minimum variations in standing crop of zooplankton during the year at a particular station in the open ocean may amount to a factor of two or three. With approach to continental regions, the productivity of the water is greater, and the seasonal variations in volume of dominant zooplankton forms is much greater, perhaps seven or eight times richer. Temperate areas of the Atlantic are much richer in macroplankton numbers than the sparsely populated waters of the Sargasso Sea, which are about four times less populated than continental-slope areas. The standing crop of zooplankton in cold northern waters is sometimes over eight times that of tropical waters, and there may be a possible tenfold to twentyfold increase in zooplankton from the winter minimum to the late summer maximum. (C.N./Ed.)

Inland waters and their communities

For purposes of convenience, ponds and lakes are described as lacustrine ecosystems and springs and rivers as riverine ecosystems.

LACUSTRINE ECOSYSTEMS

The lacustrine environment. Standing waters—ranging from days-old ponds to million-year-old lakes—despite differences in size and in age, share certain factors in common; these determine the kind of organisms that can live in such communities. Among the most important of these factors are light, temperature, and dissolved nutrients; wind waves are important in large lakes.

The influence of light. If a lake is turbid, photosynthesis is possible only in the uppermost lighted layers. In fact, when the sun is high in the sky, the illumination in the upper few centimetres of even a murky lake is likely to be too high, rather than too low, for optimum photosynthesis to occur. Deep in the lake, however, light is scarce. Attenuation of light is a logarithmic process; for example, if one-half of the incident sunlight reaches a depth of one metre (about three feet), only a quarter of it will reach two metres, an eighth of it three metres, and so on. Accordingly, the illumination and the amount of photo-

synthesis drop off extremely rapidly with increasing depth.

The food requirements of organisms, however, are nearly constant. Respiration of both plants and animals continues in the dark, so in the deeper levels of a lake respiration is likely to exceed photosynthesis. If the depth is great enough, plants will not be able to manufacture enough organic material to keep themselves alive, and the animals must depend upon food falling from the illuminated surface layers of the lake.

Lake stratification

It is convenient to think of a deep lake as divided into two layers (Figure 18), a food-generating, or trophogenic, zone, in which there is enough light for photosynthesis to exceed respiration over a 24-hour period, and a food-consuming, or tropholytic, zone, in which respiration exceeds photosynthesis. The boundary between trophogenic and tropholytic zones (the compensation level) commonly moves up and down with the seasons, particularly in polar lakes with their pronounced seasonal pulse of incident sunlight.

Temperature effects. Temperature is also important to life in lakes. It acts directly, of course; fluctuations in temperature are gradual and less extreme than they are in air. In the temperate zone during spring and summer and in the tropics during the dry season, solar heat is commonly delivered to the surface of a lake faster than the wind can stir it into the depths. Unless the lake is very shallow relative to its size, the warm surface water, being lighter, rides over the cool water beneath, and the lake is effectively divided into two compartments until either autumn or a rainy season cools the surface enough for complete mixing to occur.

The surface layer of a stratified lake is called the epilimnion and the deep layer the hypolimnion. The boundary zone between them, where temperature changes rapidly with depth, is usually known as the thermocline. Epilimnion and hypolimnion are not completely sealed off from each other. There is a continual rain of debris, largely corpses and feces, through the thermocline and some upward movement of gas bubbles, mostly methane produced by bacterial decomposition of organic sediment. Organisms with a broad enough range of temperature tolerance may move freely through the thermocline in either direction. In general, however, movement through the thermocline is slight.

The relative extent of the epilimnion and the trophogenic zone and of the hypolimnion and the tropholytic zone depends upon both the transparency of the water and the interplay of wind mixing and solar heating, but in general the epilimnion tends to be strongly trophogenic, the hypolimnion tropholytic. During stagnation, respiration by organisms in the deep water may use a large part of the hypolimnetic store of dissolved oxygen, so that only animals capable of respiring without oxygen can survive there. If the hypolimnetic volume is large or if it includes an appreciable part of the trophogenic zone, then an adequate supply of oxygen will be available throughout the period of stagnation, and the hypolimnion may provide a refuge for oxygen-demanding animals, such as lake trout or whitefish, which cannot tolerate summer surface temperatures (Figure 19).

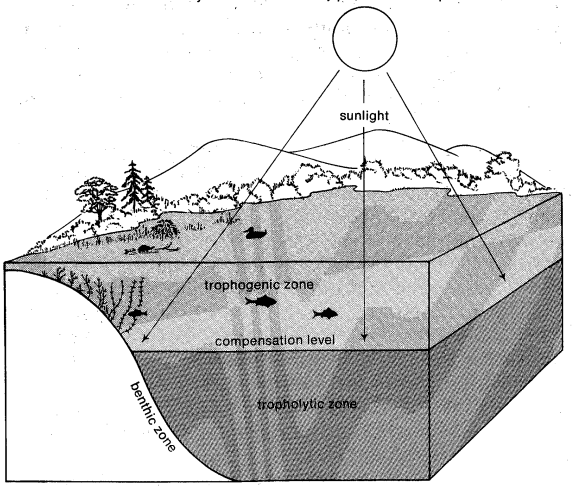


Figure 18: A lake in midsummer, showing zones of food production and consumption.

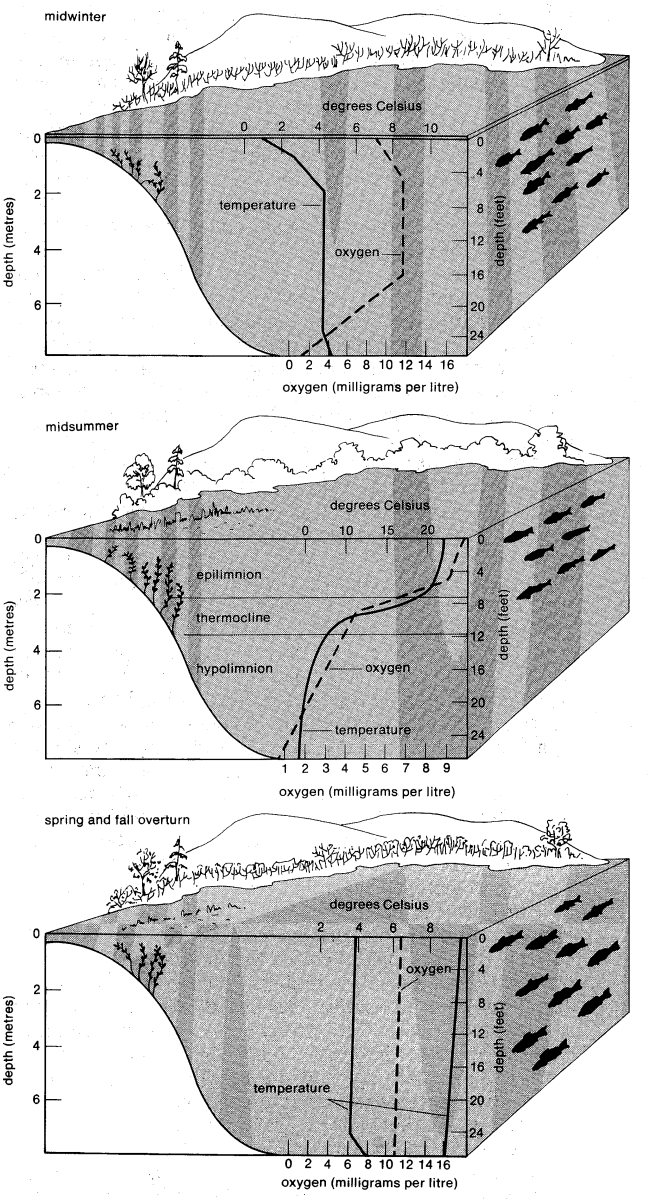


Figure 19: The influence of the seasons on lake stratification. The oxygen and temperature curves of the superimposed graphs are related to the depth of the water, as shown.

In most lakes temperature differences are the most important cause of stratification; of special interest, however, is the minority that are stratified because of chemical differences between deep and shallow water. In this condition, called meromixis, a pocket of dense water (the monimolimnion), rich in dissolved salts, lies under the fresher upper water of the lake.

Chemical stratification

The warmest lakes known are meromictic ones of such high transparency that an appreciable amount of solar heat penetrates to the monimolimnion, where it accumulates. The density changes in such chemically stratified lakes are much greater than those resulting from temperature changes in freshwaters and occasionally lead to the formation of a lake—such as Soap Lake, in Washington, U.S.—that in wintertime may have a surface layer of ice although its hypolimnion is almost hot enough to poach eggs.

Mineral factors. Life in calcium bicarbonate lake waters is determined less by changes in the major constituents—

calcium, magnesium, sodium, potassium, bicarbonate, sulfate, chloride, and silica—than by variations in biologically important substances that are commonly present in much smaller concentrations. The production of a lake commonly can be increased by adding phosphorus or compounds containing nitrogen; the addition of both increases production more than the addition of either alone. Changes in silica content have serious consequences for diatoms, algae with a siliceous cell wall; iron affects plant life directly and also chemically controls the availability of phosphorus.

In laboratory experiments it is possible to show that lake algae require numerous elements—nitrogen, phosphorus, sulfur, potassium, magnesium, silicon, sodium, calcium, iron, manganese, zinc, copper, boron, molybdenum, cobalt, and vanadium—and three vitamins. In most lakes, however, most of these materials are generally present in such adequate supply that it is usually difficult to demonstrate a deficiency.

Wind waves and seiche effects. Lunar tides are insignificant to life in lakes, being barely detectable on only a few of the world's largest lakes. Progressive wind waves, however, are a dominant influence on the plants and animals of the shore zone. In locales where the winds are strong and act over considerable surface area, waves more than 5 metres (about 16 feet) high can be produced—high enough to remove mud, clay, silt, and many organisms from beaches. Waves are probably the dominant influence on the distribution of shore plants and of the many smaller organisms of all sizes adapted to life among them.

The effect of an internal seiche

Less evident most of the time, though perhaps even more significant to life in many lakes, are standing waves called seiches, which are usually produced by the wind or by abrupt pressure disturbances. The initial disturbance tends to pile the water against one shore of the lake, leaving the opposite shore more exposed than normal. After the disturbance passes the lake, water rocks back and forth like soup in a bowl. Seiches only occasionally exceed one metre (three feet) in height. Biological significance is attached not so much to the surface seiche, however, as to the corresponding displacement of the thermocline, set up at the same time. Such an internal seiche is usually of greater amplitude than the surface seiche; in fact, all of the epilimnion may rock toward one shore, leaving hypolimnetic water exposed at the other. These large displacements of water are the main mechanism for generating turbulence in the hypolimnion of a stratified lake, and they play a vital role in the biological and chemical economy of such a lake.

Character of lacustrine communities. Suspended in the open water of a lake is a distinctive community of small—mostly microscopic—organisms, the plankton. Although many plankters can swim weakly, they do not use this ability to move from one part of the lake to another but only to maintain themselves at a desirable level in the water. They are thus at the mercy of the currents, and even turbulent eddies may overpower their slight capacity for directed movement; planktonic organisms were generally overlooked before the invention of fine silk nets and other devices that separated them from the water.

Sharing the open water with the plankton are the nekton; they are large, active swimmers able to move under their own power from one part of a lake to another. Fish, seals, turtles, hippopotamuses, and other large animals of lakes are all nektonic.

Often associated with the surface film of the lake is a well-developed community of plants and animals known collectively as the pleuston. Many of these organisms are small enough that the surface tension of the water supports them; microscopic members of the pleuston are sometimes distinguished as neuston.

More important in a quantitative sense is the benthos, the community of organisms inhabiting the lake bottom. If the water is shallow and transparent, the benthos may include an array of rooted vascular plants with microscopic algae living on them. Even deep or opaque lakes may have a dense population of benthic animals, sustained by organic matter descending from the trophogenic zone.

Lakes with waves adequate to maintain sand beaches are ringed with a psammon community, tiny organisms adapted to grow or move through the grains of sand. This community includes both plants and animals. Although sand grains reflect light in all directions and destroy visual images, a considerable amount of diffuse sunlight penetrates many centimetres into clean sand.

All the above-mentioned communities include microorganisms; some are only passive riders on the bodies of larger creatures. In addition, bacteria are found to depths of many metres in the mud under a lake—depths not considered part of the benthic habitat. Stratified lakes, especially meromictic ones, often have a layer of bacteria that live by reacting with sulfate compounds so as to form hydrogen sulfide at the interface between oxygenated and nonoxygenated (reduced) zones.

In the Arctic and in the temperate zone the abundance of plants of both major producing communities, the plankton and the shallow-water benthos, fluctuates with the seasons. This presents no problem as far as the higher plants ringing the shore are concerned: like the land vegetation, they sprout in the spring, grow more slowly during the summer, and die back in the fall in response to the same climatic controls that influence trees or grasses.

Plankton cycles

The fluctuation in abundance and activity of the planktonic algae, however, is more complex. A common pattern consists of two peaks of abundance and activity each year, one in the spring and one in the fall, with a period of relative quiescence during the summer. It appears that during the winter, when solar energy is scarce, photosynthesis is reduced, and many planktonic organisms die, releasing the nutrient salts of which they are composed. If the lake is covered with ice and snow, this process is so enhanced by the intensification of the darkness that oxygen may be sufficiently depleted to kill fish.

In spring the ice melts and freely circulating water, rich in nutrients, is exposed to bright sunlight. The result, predictably, is a spurt of photosynthetic activity and an increase in plankton abundance. During summer, thermal stratification occurs and with it a gradual depletion of nutrients in surface layers.

With the autumn overturn, however, nutrient-rich water is once again exposed to surface light intensities, which, though lower than those of spring, are still sufficient to sustain a burst of plankton production. The fall "bloom" of plankton, however, is usually somewhat less intense than the spring one.

As might be expected, many variants of this basic pattern occur. In cool, deep Alpine lakes the summer drop in production may be slight; in very cold Alpine lakes or in Arctic ones, there is likely to be no drop at all but only one burst of production—starting much later in the spring, ending much sooner in the fall, and peaking in midsummer.

In addition to these regular and predictable fluctuations in total plankton production, there are changes in the plantlike plankton (phytoplankton), with first one species, then another, appearing in a fairly regular seasonal succession, repeated with minor variations from year to year. There is also a succession in the animal-like plankton (zooplankton), with some species remaining throughout the year but others appearing only during the winter or only during the summer.

These changes probably could be understood in terms of seasonally changing chemical and physical conditions, but the exact way in which the environment acts to control a single species throughout the year has not been worked out in a single lake as yet.

It seems likely that the appearance of blue-green algae in many lakes during midsummer, when dissolved nutrients are at their lowest ebb, has to do with their capacity to utilize atmospheric nitrogen, giving them an advantage over spring and fall species that must have their nitrogen presented to them in the form of nitrates or ammonia. It is also possible that biochemical factors may be important: perhaps the blue-greens require organic growth factors released by the dominant algae of spring.

Adaptations to the lacustrine environment. Compared with rivers and oceans, lakes are rather transitory features of the Earth. A lake is likely to fill in or be drained, to

evaporate or freeze solid, in a time that is short compared with the rate of evolution. Many ponds are seasonal, some desert lakes (playas) hold water for only a few weeks after each rain. Lake Eyre, in Australia, contains water so seldom that its flooding is usually recorded in the scientific literature.

Responses to the changes in lakes

Dormancy and emigration. Because of the ephemeral nature of lakes, many aquatic organisms have mechanisms of resting in a dormant state or for moving from one lake to another. Aquatic vascular plants are commonly provided with large, thick-coated seeds, with underground storage rhizomes (underground stems), or with overwintering buds that will keep the population alive while most of the active individuals succumb to drought or cold. Daphnia, the water flea, normally reproduces parthenogenetically (without the need of mating), but, when food gets scarce or the pond begins to dry up or some other unfavourable circumstances develop, males appear in the population and fertilize the females. The offspring of this sexual union develop inside a thick-walled resting egg, which is shed into the environment at molting, instead of being placed in the mother's brood pouch. The resting egg is capable of withstanding adverse conditions of cold or drought and does not hatch until favourable circumstances have returned. Similar resting eggs are produced by many other planktonic crustaceans and by bryozoans. In the fall, in temperate-zone lakes, large, globular, jelly-filled colonies of bryozoans are blackened by thousands of small, round resting eggs, each armed with a ring of anchorlike hooks that increase the egg's chances of being carried to a surface suitable for growth.

Many planktonic algae produce resistant spores capable of withstanding drought, and the larvae of bivalve mollusks are adapted to cling for dispersal to the gills of fish or to the feet of ducks. The short adult phase in the life of aquatic insects is also a device for moving from lake to lake. A large part of the lacustrine biota seems to consist of species well-adapted to fleeing to more favourable areas should the lake in which they live become unsuitable. This is true even of such large, self-sufficient animals as the hippopotamus and the crocodile.

Planktonic adaptations. Plankters must stay suspended in the light water to survive. If they sink into the tropholytic zone, they die, either by poisoning or by starvation. Many plankters have overcome this threat by moving enough to keep themselves suspended. A few have developed gas bubbles or oil droplets to aid buoyancy. Other plankters have adapted by becoming as small as possible; thus, they sink very slowly through the water, relying on normal turbulence to hinder sinking farther. If the growth of the population is greater than the slow loss to the depths, then the species can survive. This is probably the commonest way out of the dilemma and may explain why so many plankters are microscopic in size, especially in fresh water. In the sea, with its higher salinity, the sinking problem is less acute, and larger plankters are found.

Many of the planktonic organisms in lakes are surrounded either with a jelly coat or with long spines sticking out in all directions. Both are probably adaptations for increasing effective diameter without greatly increasing mass, thus reducing the sinking rate. Many diatoms are slightly asymmetrical—sigmoid or S-shaped, for example—a physical advantage that also reduces the rate of sinking.

Vertical migration and cyclic change of form

Many species of lake plankton show one or another of two bizarre phenomena that may adapt them to open-water existence, although just how is still obscure. The first phenomenon is daily vertical migration: the plankters move up and down in the water, often considerable distances, according to the daily cycle of solar illumination. The second phenomenon is cyclic change of form. This occurs in animals such as daphnia and in some plants, all of which reproduce asexually as they change in form from one generation to another. Some populations show only slight change during the course of each year; others go through such variation that the extremes in form would certainly be taken for different species if the intermediate types were not known. These form changes are induced largely by changes in temperature during development, although turbulence also plays a role. Of the many ideas that have been advanced for the adaptive value of the changes—that they are adjustments to the properties of warm water, that they keep the animals swimming in strata in which food is densest, that they are defense mechanisms against seasonal predators—none seems a convincing explanation for the general phenomenon.

Physiological adaptations. The ancestors of most animals that live in lakes were either terrestrial or marine. Adapted to a situation in which water was scarce, these ancestors had to be able to conserve water and to excrete extra salts and waste products with minimal water loss. The water in most lakes is much more dilute than is the body fluid of the animals; there is thus a strong tendency for the animals to gain water. This problem has been solved in two fundamentally different ways. One is typified by the freshwater mussels, which have an enormous area of mantle (an extension of the body wall) and gill exposed to the water; it might seem that they would take up water at a high rate. The mussel, however, has developed extremely dilute body fluid—thus minimizing the tendency for water to cross its membranes—and a kidney that, because it is able to excrete very dilute urine, can excrete the excess water that does happen to enter the body. The problem has been solved in a different way by most insects in fresh water, which have a hard outer skeleton (or exoskeleton), covering as much as possible of the body; because the exoskeleton does not allow water to pass through (*i.e.*, it is impermeable), water uptake is reduced and can be handled adequately by an excretory apparatus designed for terrestrial performance.

Not all lake dwellers face the problem of an excess of water. The problem of those who live in waters many times more concentrated than the sea is to conserve water and get rid of excess salts. Relatively little is known about the excretory mechanisms of animals in such environments, however.

Water and salt balance are more easily studied in large nektonic and benthic animals than in microscopic organisms, but the latter, with their relatively thin walls and enormous surface-to-volume ratio, are exposed to severe stresses. In the case of single-celled protozoans, the problem is met by a structure called a contractile vacuole, by which dilute fluids are excreted.

Oxygen sources

Lake animals such as mussels, which have a large permeable surface exposed to the environment, are able to obtain their oxygen by diffusion from the water. Many insects, however, exploit the natural advantages of air as their respiratory medium.

Lake plants are commonly well provided with hollow tissues in their stems, through which they are able to pump oxygen down to roots and tubers buried deeply in mud. Mosquito larvae of the genus *Mansonia* tap into this gas supply with their respiratory organs and so free themselves from the usual repeated journeys to the surface to replenish their air.

Species interactions. Probably because most lakes are young and transitory, they have few of the elaborate and subtle interactions between species that are so conspicuous a part of the marine or terrestrial scene. A few cases are reminiscent of sea corals and the algae that live in them (zooxanthellae). Freshwater sponges, for example, are green because of the algae within them when they grow in the light; the same sponges are white when they grow in darkness. Some salamander eggs are usually green as a result of algae growing in the jelly masses.

Among the fishes of the oldest lakes are some interesting cases of egg predation; in one, a species of cichlid fish feeds exclusively on the eggs of mouth-brooding cichlids of other species. The predatory cichlid mimics the prey cichlid closely in colour pattern and general appearance, persuading the mouth brooder to spit out its incubating eggs.

Parasitic relations between lake dwellers are elaborate, probably because the cycle has evolved in rivers and been secondarily extended to lakes. Many parasitic life cycles involve both lakes and terrestrial organisms; for example, the schistosomes, a group of blood flukes that are widely distributed in the lakes of the world, pass several larval stages in freshwater snails but spend their adult lives in

the bloodstream of a terrestrial vertebrate. In most of the temperate zone the vertebrate host is usually a duck. Occasionally, certain of the larvae may penetrate the skin of humans, where they die, producing an annoying but usually short-lived dermatitis called swimmer's itch. In Japan and much of the tropics, however, the larvae of several species complete their development in man and cause a serious human disease.

Biological productivity. Generally, the productivity of lakes is lowest in the Arctic, intermediate in the temperate zone, and highest in the tropics. The dark Arctic winter provides little solar energy for photosynthesis, and most of the springtime period of bright sunshine and long days occurs when the lakes are still covered by a thick layer of ice. In the tropics, on the other hand, where there is an adequate supply of solar energy all year and no ice cover to absorb it, high temperatures speed decomposition of plants and animals, so that nutrients of which they are composed are returned rapidly to the water and support new growth.

Correlations with latitude

A second important influence on the productivity of lakes is the nature of the underlying bedrock. Rocks that are readily soluble and rich in nutrient elements, such as calcium, potassium, silica, and phosphorus, provide for a high and sustained rate of photosynthesis. Lakes resting on insoluble rocks, even though located in the favourable light and temperature environment of the tropics, have a productivity so low as to be unmeasurable by standard methods. It is in tropical regions of internal drainage, with much volcanic ash in the rocks—such as the rift valleys of tropical East Africa—that the most productive lakes in the world are found.

Presumably, an estimate of the total amount of each nutrient in the water and organisms of a lake would show a clear relation with productivity, but such an estimate is difficult to make. Even for phosphorus, which can be detected easily by radioisotope-tracer methods, the estimate is too cumbersome to be used on a widespread comparative basis.

There is, on the other hand, a fairly strong correlation between the bicarbonate-carbonate content of lakes and their primary productivity. This is likely to be at least partly the result of a correlation between bicarbonates and the generally unknown and unmeasurable total supply of nutrient salts, but it also may mean that plants in most lakes are chronically short of a convenient carbon source for photosynthesis.

If a lake is deep and transparent, it can sustain a respectable productivity per unit surface area while at the same time exhibiting a low productivity per unit volume of water. In a shallow lake, on the other hand, all the primary production takes place in a short water column, so that its effects, such as the production of algal blooms, are much more pronounced. This concentration of such growth is beneficial to organisms higher up on the food chain, such as zooplankton and fish, because large volumes of water are not required to obtain food. This is probably one reason the shore zone of a lake normally supports much denser populations of fish than does the open water.

The depth of mixing is important to the phytoplankton in another way. If it is so great that the phytoplankters spend an appreciable part of their time in the tropholytic zone, they may use up a large part of the food they manufacture in their own respiration. This is probably why the dense perennial blooms of highly productive tropical lakes develop only in shallow lakes or in meromictic ones, in which mixing is shallow and contained largely in the zone of trophogenesis.

Lake food chains

In general, only about 10 percent of the energy passing through one level in any food chain is consumed by the next trophic level above, which places a severe limit on the possible length of lake food chains. Few, if any, consist of more than six links, and in small lakes they may be only three links long. A high-level predator can increase its population by shortening the food chain on which it is dependent, just as human populations can (and do) by shifting from a diet rich in meat to one that is largely cereal. The resulting increase of food from a short food chain, however, is obtained at a considerable increase in the energy used to gather the food. (D.A.L./Ed.)

RIVERINE ECOSYSTEMS

The riverine environment. By volume, the rivers of the world constitute a small fraction (1 percent) of the total amount of water in aquatic systems, but their high rate of flow gives them great ecological significance. As active agents of mechanical and chemical erosion, rivers exert a major control upon the terrestrial ecosystems of their watersheds, as well as any lakes or oceans into which they discharge.

There are significant differences among the rivers of the major continents regarding chemical composition, concentration, and rate of erosion. Australia is so dry and Antarctica so cold that they make little contribution to the global budget for river transport. The streams of Asia carry almost three times as much suspended material per unit area as do those of any other continent. The river waters of South America are the most dilute; those of Africa, being similar in composition, are about twice as concentrated. North America, Asia, and especially Europe differ from the two more southern continents in carrying a much larger quantity of calcium and bicarbonate. The low salt concentration of South American rivers is offset by their great discharge.

Rivers, particularly in arid lands, show great annual fluctuations in size and other properties. Variations of four times the normal discharge and six times the quantity of suspended particles are not unusual and are accompanied by changes in current velocity, concentration of dissolved substances, and shape of the channel. The formulation of the properties of a riverine ecosystem is accordingly more difficult than it is in lakes and the ocean, and, although considerable attention has been devoted to the biology of streams, the results have tended to be descriptive and individual, with little applicability to streams in general.

Zones of erosion and deposition

A distinction is often made between a zone of erosion and a zone of deposition in rivers, but it must be remarked that, whatever the size of the particle that is moved by a flood, it will be deposited when the flow is less. It is preferable to make the distinction between zones where the bottom has stones large enough for the biggest invertebrate animals to cling to and where the bottom is of sand or mud whose particles are so small that the larger invertebrates must burrow. If a river with a sandy or muddy bottom is shallow enough, rooted plants will grow in it. Between the two zones there will be an intermediate one where fine particles settle during low flow and are washed away by a flood, and the slope, or river bottom grade, may be such that this intermediate zone is longer than either of the other two. A river from source to mouth is a continuous whole, and attempts to divide it into zones must be arbitrary. Any classification system may prove useful for a particular purpose, without being a concept with fundamental significance for biologists.

Water emerging from deep underground layers is generally cold, often with but a small range of temperature change during the course of a year, but it may be warm if it drains shallow soil. The smaller the volume of water the warmer it will become in the sun and the more heat it will lose at night. Larger water volumes have less daily temperature fluctuation and they are never far from average air temperature. If there is not a big difference in altitude, headstreams may therefore reach a higher maximum temperature than the main river. This condition, of course, has importance regarding the type of life found in the various parts of a stream.

A torrential stream is nearly always well oxygenated. A slow river is often well oxygenated by day because of the activities of photosynthetic plants, but it may be depleted of oxygen by night, when only oxygen-consuming and decomposition processes are active. Turbidity, the muddiness and cloudy condition of the water, generally increases steadily from source to mouth. Accumulated organic matter washed from the land is believed to be the main base of the food chain in running water, and therefore in hilly or mountainous regions the amount of material entering the stream that may become a source of

food tends to increase as the area drained increases with distance from the source. The amount may decrease farther downstream where the carrying capacity of the river drops with reduced flow.

An old and well-known method of dividing rivers into zones is based on the presence of one of four fish species commonly found in western Europe: *Salmo* (trout), *Thymallus* (grayling), *Barbus* (barbel), and *Abramis* (bream). Each zone as defined by these fish coincides with a given slope, though for any zone the wider the stream the more gentle the slope. Even in western Europe there is a certain inconvenience in designating the zones by species of fish, for some, notably grayling, are not widespread. Elsewhere other species must be sought. Running water systems have also been classified into zones called the rhithron, which is roughly the stony upstream region, having well-oxygenated water and with temperature rising to no more than 20° C (68° F), and the potamon, which is the downstream portion, having a sandy or muddy bottom and temperature exceeding 20° C at the warmest time of year. Both rhithron and potamon are subdivided. Within any one subdivision, however, the composition of the living community may be changed by some factor unconnected with flow or temperature, and the terms therefore have no precise biological significance though they may be useful in a general sense.

Upstream and downstream zones

Character of riverine communities. *Organisms in streams.* The slower the streamflow, the more the composition of the living community resembles that of still water, and, therefore, any discussion of peculiar features of running-water organisms must centre largely round those that have colonized the swifter-flowing waters.

Stagnant waters fill and disappear but a watercourse must flow as long as there is precipitation to supply it. It is, therefore, not surprising to find in running water certain groups of organisms that have changed little during the millions of years in which others have evolved and adapted. Most stone flies (order Plecoptera) and many mayflies (Ephemeroptera) inhabit running water, and other primitive groups confined to it include the crustacean sea hare and the caseless free-ranging caddis fly larvae of the insect family Rhyacophilidae. Moreover there are few groups of freshwater organisms that are not represented in running water. The exceptions are extreme specialists such as the phyllopod crustaceans (fairy shrimps and related forms), which must swim to feed and have survived only in temporary water where they are free from the predation to which this mode of life exposes them.

Rapidly flowing water is well oxygenated, and a moving medium brings to its inhabitants a constant supply of oxygen and also salts, at the same time removing their waste products and carrying a constant supply of food. The disadvantage of life in a moving medium is that any accidental displacement must always be in one direction. Even the organisms that live in the substratum (various layers of the stream bottom) are occasionally subject to this hazard when exceptional flow causes the substratum to shift. A further problem is the accomplishment by organisms of colonization of a biotope (uniform habitat occupied by a uniform community of organisms) that may be very long and very narrow.

Hazards of the environment

Plant communities. Stones and rock are covered by algae, many of which are single cells, though some are filamentous and trail in the water in tresslike tufts. There are many species of these, and they are often scoured off their supports by sand and gravel when the rate of streamflow increases; there are marked changes in the communities with the seasons even when this does not happen. Algae tend to occur in irregular patterns varying considerably in species composition and abundance. It has not, therefore, so far proved possible to discern definite algae communities associated with given stream conditions, though it is known that temperature, light, substratum, and dissolved substances in the water are factors that influence the occurrence of certain species.

A stone or boulder that survives for a long time without being overturned, and that is not scoured by particles of sand and gravel, will become densely covered by mosses and liverworts.

Rooted plants grow thickly where substratum and water depth are suitable, but attempts to define communities comparable with those in still water have been unsuccessful. This is partly because of the instability of the flowing water system. Once established, a tuft of vegetation impedes flow and facilitates the deposition of silt, which may alter the substratum until it becomes favourable for some other species. Eventually the vegetation may present so much resistance to streamflow that the whole mass is washed away. Many species of plants in temperate latitudes die down in winter and no longer protect silt that settled round them in summer. Copious plant growth interferes with water flow to such an extent that it often becomes necessary to cut and remove it from some streams to prevent flooding.

Animal adaptations. *Life in the streambed.* There is a zonation of animals in the substratum, as may be seen in Figure 20. If bottom particles are not too fine, there is a rich and varied animal community down to 1 metre (3.3 feet) or more below the surface of the streambed. Certain crustaceans, mites, and nematodes are restricted to these deep habitats but also abundant may be tiny nymphs or larvae, the immature insects that, when they are larger, inhabit higher regions where the particles are coarser.

Stratification of life in the streambed

Nearer the surface, in sand and gravel, may be found various animals, notably segmented worms (of class Oligochaeta) and the larvae of certain families of flies (of the insect order Diptera) that, having come in from a similar habitat elsewhere, have been able to colonize running water with little change of form and habit. Nearer the surface still, but protected from the current by the large stones above, are many animals that show no structural adaptation to life in running water, though obviously without some modification of behaviour they would not survive. Most of the stone flies, various caddis fly larvae (order Trichoptera) that make cases, and amphipods (shrimplike crustaceans) such as *Gammarus* are examples that are likely to be familiar over most of the world. Some of the stone fly nymphs are long and thin and consequently are well designed to live among the spaces between small stones and pieces of gravel. There is no rigid line of distinction between the inhabitants of the three zones mentioned.

Life at the surface of the streambed. Finally come the animals that colonize the large stones forming the surface of the streambed. Many are structurally modified for swift-water environments. The freshwater limpets (family Ancylidae) have a simple conical shell that offers little resistance to the current and a foot that forms an efficient sucker. Larvae of the net-winged midges (family Blepharoceridae) have six suckers provided with muscles that can create a powerful suction. Another similar adaptation is seen in several families of beetles and reaches its highest development in the water pennies (family Psephenidae). The body is flat and wide, and the margin is pliable and beset with spines and hairs. It can be adjusted to fit the irregularities of the surface sufficiently to prevent any water flow underneath the larva, and the pressure on the rounded upper surface tends to press it against the substratum even though there are no suction-producing muscles. In certain mayflies, modification of the nymph's gills produces the same effect. Such nymphs are frequently confined to running water because they cannot use their gills to create a current if necessary. For example, the mayfly genus *Rithrogena,* which has modified gills, respires less rapidly as current speed falls, whereas related genera that can move their gills consume as much oxygen in still as in flowing water. They occur on stony lakeshores, and species of *Rithrogena* do not.

Nymphs of the mayfly family Ecdyonuridae (or Heptageniidae) have broad flat heads with thin margins, and upper leg segments are of a similar shape though less acute at the margins. The body and legs are pressed against the surface; retaining its hold by means of sharp claws, the nymph can scuttle across a flat hard surface with great nimbleness. In the Southern Hemisphere this adaptation is found in spinners (family Leptophlebiidae), which replace the Ecdyonuridae.

Larvae of the blackflies (family Simuliidae) may perhaps

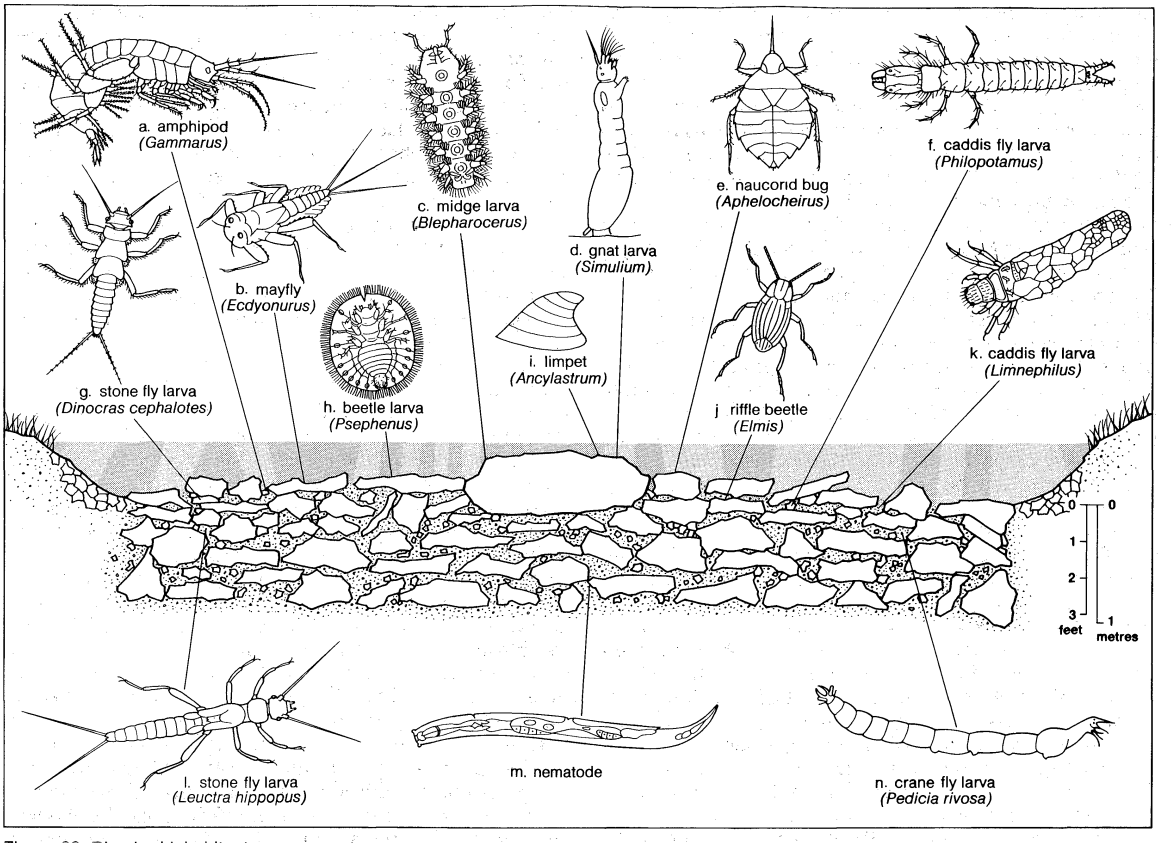


Figure 20: Riverbed inhabitants.

From (a,d–g,i,k–n) T. Macan, *A Guide to Freshwater Invertebrate Animals*, by permission of the author and Longman Group Ltd., (b) G. Pleskot, *Der Stand der Biologischen Fliesswasserforschung*, (c,h,j) C. Wesenberg-Lund, *Biologie der Susswasserinsekten*

Adaptation to the running-water environment

be said to be the most highly adapted of all running-water animals because they not only anchor themselves in swift currents but also make use of the flow to bring food. At the tip of the abdomen and on a proleg near the head are pads armed with concentric rows of hooks, of which the points can be directed outward by pressure of the body fluid and retracted by muscles. The larvae can produce threads of silk, and these they lay over the surface of a stone, often where the current is swift. At intervals, tangled masses of silk are laid down. The larva can move across its silken lattice with a looping action, taking hold with fore and aft pads of hooks alternately. Most of the time it trails in the water with the rear hooks embedded in one of the tangled blobs of silk. It extends a rakelike apparatus, and when this has strained a mass of fine particles from the water the appendage is retracted and the mass is eaten.

Three families of caddis flies utilize the current to bring food. None makes a case in the fashion of most caddis fly larvae but all spin nets. Members of two of these families feed on drifting organic matter, but the third is carnivorous and lives a life not unlike that of a spider on land. Members of this family occur in still water.

Adults of most beetles (order Coleoptera) and bugs (order Heteroptera) must come to the surface periodically to renew their store of air, a necessity that is plainly inconvenient in a medium flowing one way. It is not unexpected, therefore, that in swifter regions of rivers these two orders are represented mainly by families that have developed plastron respiration and can live permanently on the bottom (riffle beetles, family Elminthidae, and creeping water bugs, family Aphelochiridae).

Factors affecting communities in streams. Several factors influence the composition of the communities. It is impossible to separate the effects of streamflow and substratum. When current falls to a speed at which fine particles settle, the animals that inhabit the spaces between small stones and pieces of gravel are replaced by burrowing forms, of which the main groups are oligochaetes (worms), lamellibranchs (clams, mussels), and chironomid larvae (flies). Nymphs of the burrowing mayflies (family Ephemeridae) burrow in sand and mud, and some dragonfly nymphs have also taken to this way of life, particularly in Africa.

A varied animal community is found in vegetation, if this develops. Animals with adaptations for life on a hard flat surface, such as the mayflies of the family Ecdyonuridae or limpetlike mollusks, do not occur, but less specialized forms such as the shrimplike *Gammarus* may continue in great numbers. *Simulium* (gnats and blackflies) and *Baetis* (mayflies) are two genera that are also abundant in both stony and weedy zones, though the species are generally different in the two.

Within one zone, rate of flow is probably not of great importance for most species since they live among stones, avoiding places exposed to the full force of the current. It is important to the net-spinning caddis fly larvae, which must find a place where current speed lies within a narrow range. It has been shown experimentally that a change in the pattern of flow is quickly followed by a change in the distribution of their nets.

Relation between stone size and species colonization

Different kinds of rock break up into stones that differ in shape and size and texture, but whether this affects the communities that colonize them is not known. That the distribution of certain species that inhabit streambeds of fine particles is related to the size of the particles has been clearly established. In Europe the green drake mayfly (*Ephemera danica*) inhabits sand, and the common mayfly (*E. vulgata*) mud, and several similar examples have been described in North America. Stones that have been rounded by abrasion during transport by ice or water provide a difficult substratum colonized only by agile animals. Flat stones that move less easily harbour a more varied community, and it is richer still if, as generally happens, the stones become covered with moss.

Springs and the stretch immediately below them are often inhabited by species not found elsewhere, and some of these are confined there by their requirement of low temperature. One of the few species that has been thoroughly investigated, both in the field and in the laboratory, is the flatworm, *Crenobia* (*Planaria*) *alpina*. It is commonly

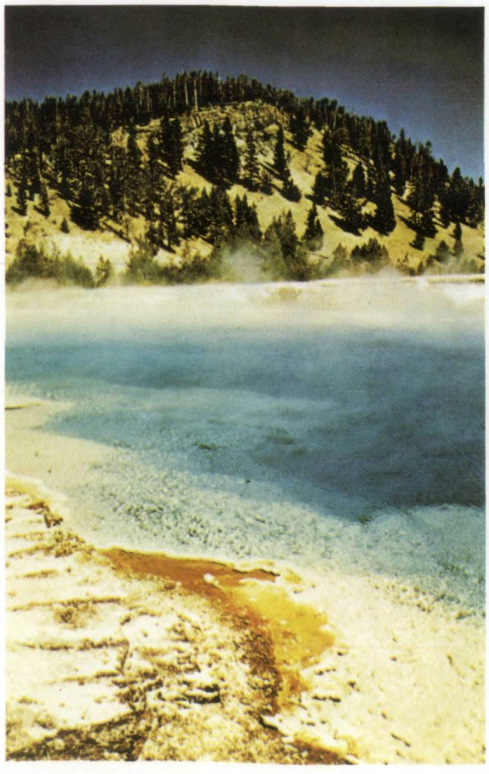
Polar bears crossing ice in Norwegian Bay, Northwest Territories, Canada.

Lake near Bergen, Norway, fed by a glacier.

Vegetation clothing banks of hot springs in Iceland.

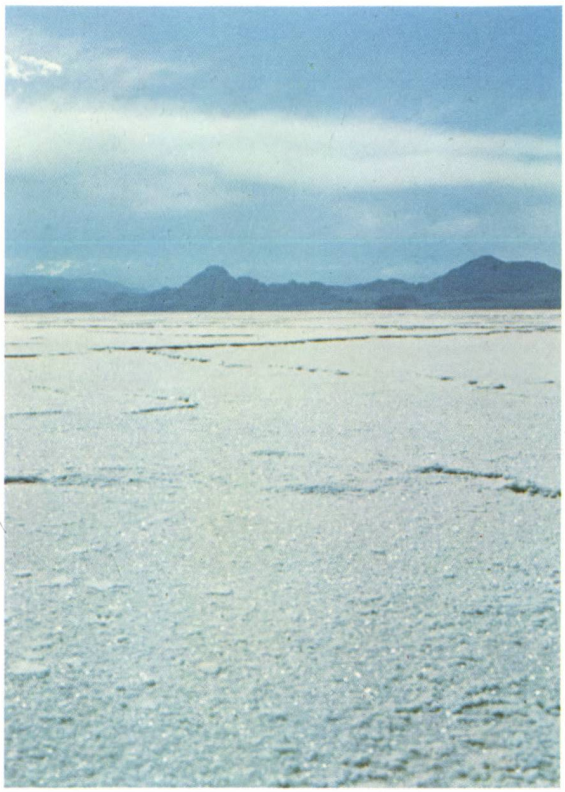
Extremes in aquatic environments

Mineral-rich thermal area, Yellowstone National Park, Wyoming.

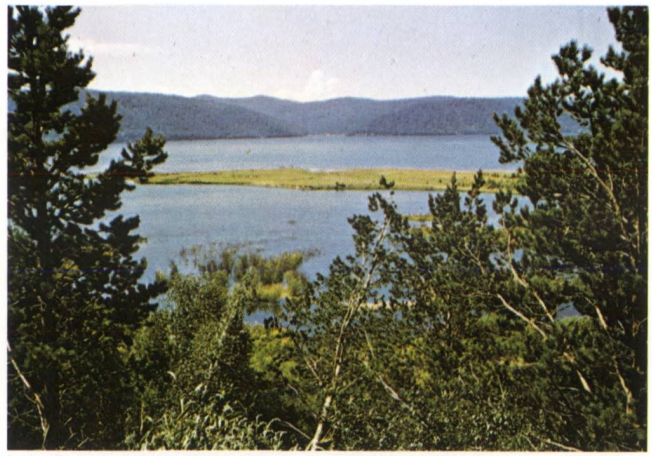


Algae-stained thermal pool, Yellowstone National Park, Wyoming.

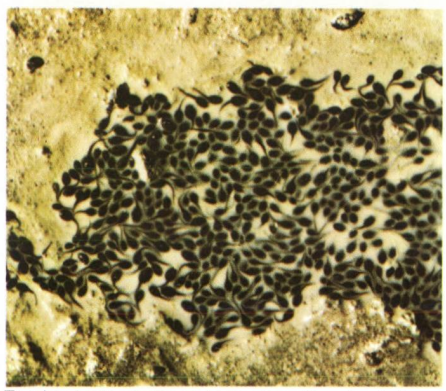
Plate 1: (Top left) Ted Grant—Information Canada Phototheque, (top right) Sven Samelius, (centre right) Peppe Traversa—FPG, (bottom left) Charles Steinhacker, (bottom right) Anthony Mercieca from Root Resources—EB Inc.



Salt-encrusted shore of Great Salt Lake, Utah.



Lake Baikal, Soviet Union, the world's deepest lake, rich in unique species.



Toad tadpoles massed in the remains of a drying puddle, Kenya.

Lake Manyara, Tanzania, a soda lake, containing only alkali-tolerant species.

Permanent pond, Loughrigg Tarn, Cumbria.

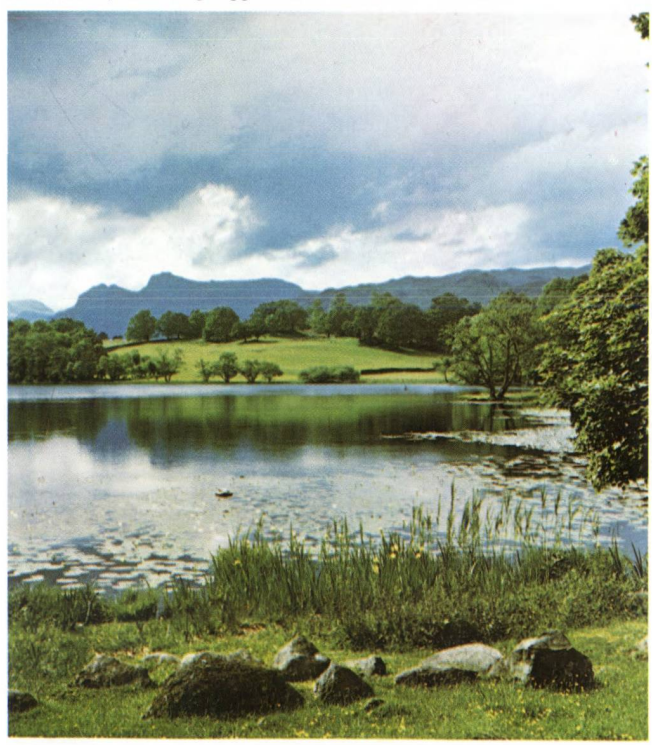


Plate 2: (Top left) Barbara Van Cleve, (top right) Harrison Forman, (centre right) Jane Burton—Bruce Coleman Ltd., (bottom left) from the book *The African Experience and The Tree Where Man Was Born*, text by Peter Matthiessen, pictures by Eliot Porter, published in 1972 by E.P. Dutton & Co., Inc., and used with their permission, (bottom right) Tourist Photo Library

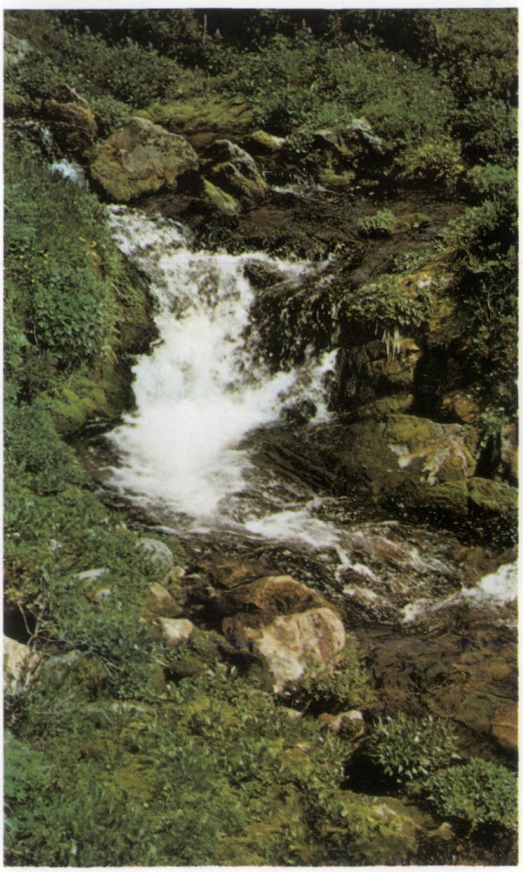
Plate 3: (Top) Laurence R. Lowry—Rapho/Photo Researchers, (centre right) G.R. Roberts, Nelson, New Zealand, (bottom left, bottom right) Bruce Coleman Inc., (bottom left) Mira Atkeson, (bottom right) James R. Simon

The Salinas River, California, emptying muddy water into Monterey Bay.

Delta of the Tongariro River, New Zealand, composed of accumulated silt washed downstream.

Inland water systems

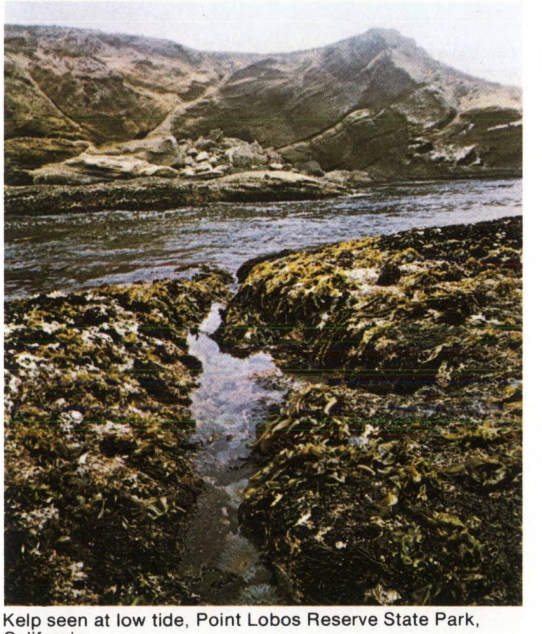
A clear, rapid-moving alpine stream, Goat Rocks Wilderness, Washington.



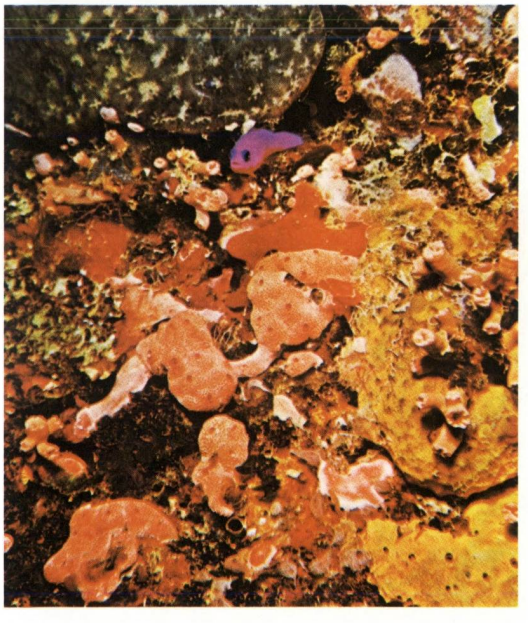
Trout Creek, a slow-moving, mature river, Yellowstone National Park, Wyoming.

Sand dunes encroaching on salt marshes bordering the entrance of Laguna Ojo de Liebre, Baja California.

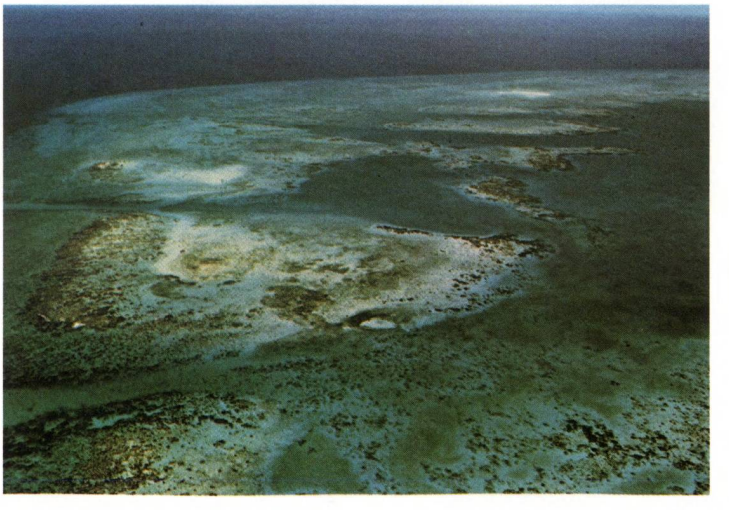
Marine life typical of that found in cold water, on a piling in Hodgkins Cove, Cape Ann, Massachusetts.



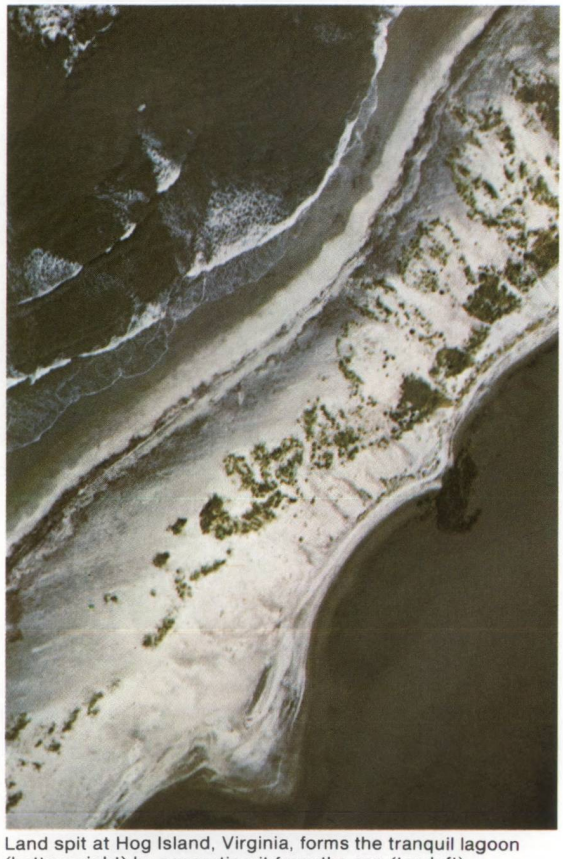
Kelp seen at low tide, Point Lobos Reserve State Park, California.



(Left) Marine life typical of that found in tropical waters, Palau Islands, Trust Territory of the Pacific Islands. (Below) Coral reef, Belize.



Sea bottom near Bonaire Island, Lesser Antilles.



Land spit at Hog Island, Virginia, forms the tranquil lagoon (bottom right) by separating it from the sea (top left).

Marine environments

Algal community, Sea of Japan, near Shimane.

Leaf barnacles and colony of tube worms exposed at low tide, Santa Cruz, California.

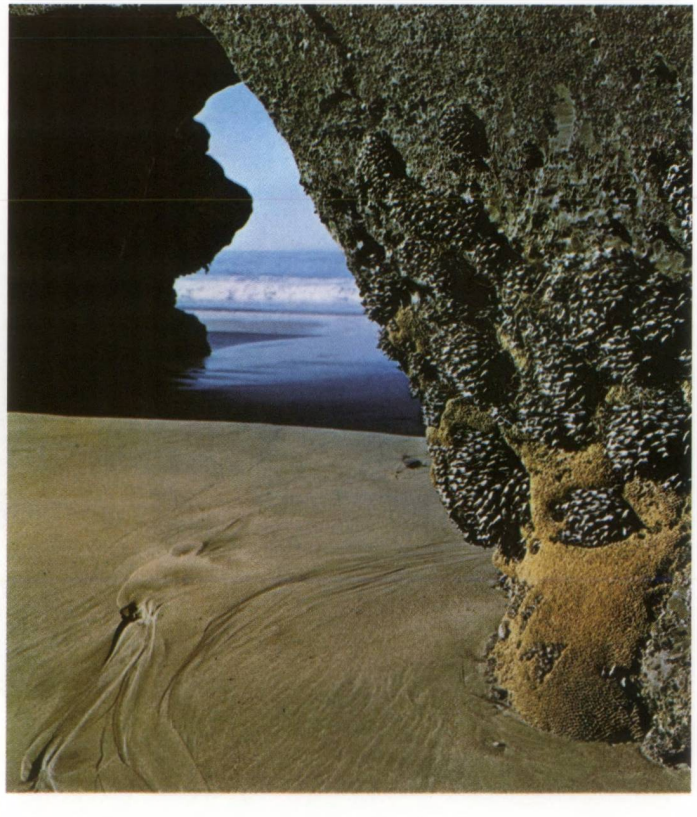
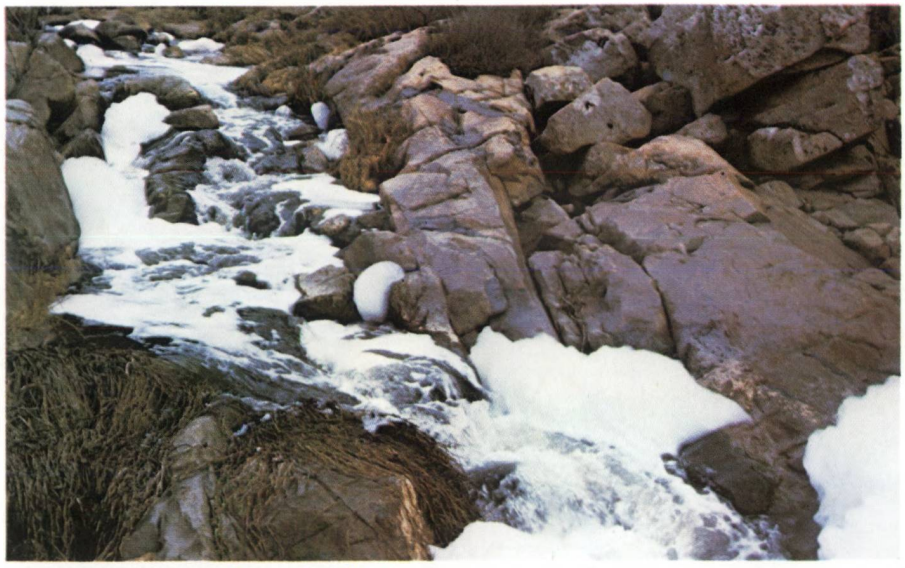


Plate 4: (Top) David Cavagnaro, (centre left, bottom left, bottom right) Douglas Faulkner, (centre right) Tom Myers

Plate 5: (Top left) Jeff Myer—Animals Animals, (top right) © Costa Manos—Magnum, (bottom left) W.H. Hodge, (bottom right) Al Lowry—Photo Researchers



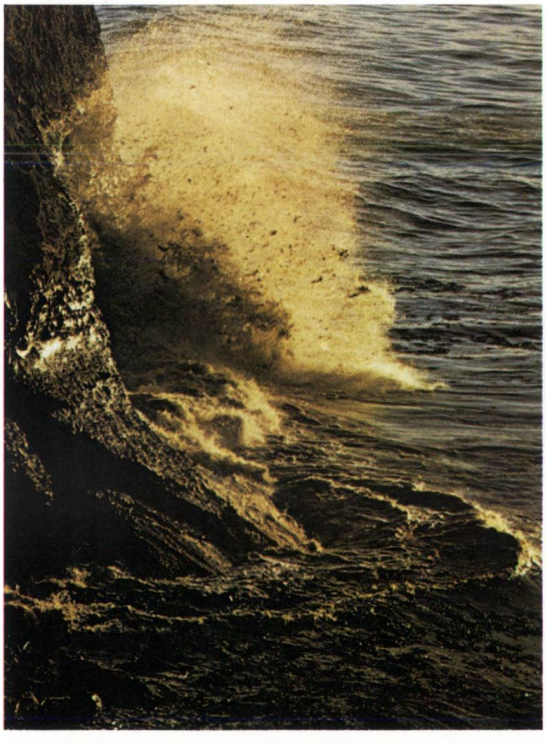
Detergent foam on Escondido Creek, California.

Duck feeding in refuse, Lake Mendota, Madison, Wisconsin.

Polluted water systems

Paper mill pollution, St. Catharines, Ontario, Canada.

Oil in surf at Santa Barbara, California.



Straw-covered tide pools soak up oil at Agate Beach, near San Francisco, California.

Plate 6: (Top) Barbara Adams from Root Resources—EB Inc., (centre left, centre right) The National Audubon Society Collection/Photo Researchers, (centre left) Raymond L. Nelson, (centre right) Lincoln Nutting, (bottom left, bottom right) Tom Myers

confined to water whose maximum temperature does not exceed 15° C (59° F), though it can tolerate 25° C (77° F), provided it is not exposed to it for too long. Reproduction generally takes place only in water colder than about 12° C (54° F). Other animal species are confined to cold water not by physiological limitations but by their inability to compete with a rival in warmer water. Britain for example lacks the amphipods *Gammarus fossarum* and *G. roeselii* and instead *G. pulex* ranges in British streams from source to mouth. In Europe *G. pulex* is confined to the middle reaches of flowing waters and is apparently kept out of the upper ones by *G. fossarum* and the lower ones by *G. roeselii*.

Animal populations tend to be densest in those places where the configuration of the bottom produces areas of still water with consequent settling of debris. In addition some of the most notable changes in the composition of the population can be related to food supply. Below lakes the water is rich in floating algae and tiny drifting animals as well as suspended organic matter, and here the numbers of net-spinning caddis flies and blackflies are much higher than elsewhere. They appear to exclude other species found under similar conditions of flow and substratum but without the augmented food supply. Slight enrichment by sewage leads to an increase in the number of flatworms and a decrease in the number of certain other species. It has been suggested but not proved that the flatworms eat them. In rivers enriched by the decomposition of sewage, great numbers of *Asellus* (a freshwater crustacean), leeches, mollusks, and alderflies (*Sialis*) are found. Below the zone of enrichment these animals become progressively scarcer and are replaced by the community associated with similar conditions of flow and substratum in unenriched water. Again the full interrelationships have not been worked out. Decomposition of sewage is accompanied by lowered oxygen concentration in the polluted waters, and the typical inhabitants of stony rivers require a high concentration. There is also evidence that they are adversely affected by the bacteria that coat the stones and indeed some of the animals themselves. In lakes the most plausible explanation at present is that predation by the species favoured by enrichment leads to the elimination of the species typical of areas where there is no enrichment, and it is reasonable to postulate that this factor operates in running water too.

Life cycles

Life cycles of running-water animals are unexpectedly diverse and offer a rich field for experimental investigation. Nearly all the observations have been made in temperate regions. A number of animals, notably the larger ones such as leeches, snails, bivalve mollusks, and crayfish, take more than a year to reach maturity. Many are univoltine; *i.e.*, they take one year to complete development. Some species in the genera *Baetis* (mayflies) and *Simulium* (blackflies) and in the caddis flies pass through two generations in a year. Quicker development is rare, *Gammarus pulex* being an outstanding example, completing its life cycle in less than two months.

Many of the univoltine species have a resting stage, which generally tides them over the warm period of the year. It is often the egg stage, sometimes the early nymphal stages, and sometimes, chiefly in caddis flies, the last larval molting stage. The common European spiny crawler mayfly (*Ephemerella ignita*), apparently spends some 10 months in the egg stage, at least in parts of its range; nymphs appear about midsummer, grow rapidly, and emerge in August or early September. In other species, the small minnow mayfly (*Baetis rhodani*) for example, some eggs remain unhatched for many months, whereas others hatch after a few weeks.

A net set so that water flows through it catches representatives of many of the species inhabiting the stream. Catches of most are much higher by night than by day and are often greatest just after dusk and sometimes again before dawn. This phenomenon, commonly referred to as drift, has been much studied and has provoked several controversies. The number of specimens drifting seems to bear some relation to the population in the stream and is at its highest during the period of most rapid growth. Therefore it probably represents to some extent a removal of surplus population. Drifting animals are likely to fall prey to fish, but there is evidence that those that escape this fate do not travel far before regaining the safety of the bottom. Some of the larger adult caddis flies have been observed flying upstream to lay their eggs, but there is evidence that other insects do not maintain the population in the upper reaches in this way. Many have been shown to move against the current during the aquatic stage of their life cycles, particularly when small, and it is likely that uniform colonization of all suitable stretches of a watercourse is maintained in this way.

Detritus, drifting organic matter originating from dead leaves and other vegetable fragments that are blown or washed into the water from the land, is the main primary source of food in rivers and streams. The limpets, the flat nymphs of the mayfly family Ecdyonuridae, and larvae of various caddis flies graze upon the algal felt that covers the upper surface of stones but probably subsist extensively on the detritus trapped in this. Detritus may also stick to the mucous trails left by flatworms or lodge in the irregularities of a rough surface. Little is known about the digestive powers of the various detritus feeders. That vegetable remains are not an easy source of food is indicated by the frequency with which they are passed twice through the same intestinal tract. It is likely that the actual source of food is the fungi and bacteria that are breaking down the vegetable tissue. The flatworms, the large stone flies, some of the caddis flies, and many fish are carnivores.

Detritus

Biological productivity. In a study of total production of flesh by trout (*Salmo*) per year in the Horokiwi Stream in New Zealand, there were about 50 grams per square metre (*i.e.*, 500 kilograms per hectare, which is approximately 450 pounds per acre). An important point that came out of this investigation was the large contribution to production made by young fish, whose size was very small but whose numbers were large. Fish are comparatively easy animals with which to work, mainly because of their large size and the fact that the eggs hatch within a short period. Moreover the young do not live in the substratum. The trout establishes territories whose sizes depend on the configuration of the bottom and not on food supply. Fish secure efficient exploitation of the resources of the environment by means of an indeterminate size at maturity. If food supply is poor, the fish remain small; if it is good, they grow much more rapidly; but sexual maturity is attained after a period that does not vary significantly with size, and the small fish reproduce just the same as normal adults. That calculations of production by populations of wild fish are of the right order can be confirmed by comparison with the extensive data obtained by commercial cultivators. Consequently there are reliable figures for the production by several species in the wild state.

Information about invertebrates is much less satisfactory. It was discovered how much the trout in the Horokiwi ate, and it turned out to be some 17 times greater than the number of invertebrates apparently available as food. This observation, known as Allen's paradox, has been confirmed also in other parts of the world. In other words, production is underestimated. It is possible that unaccounted-for populations among running-water animals are to be found well below the surface of the substratum, where it is known that tiny nymphs abound and where most sampling instruments do not penetrate. They may also be washed down from the smallest tributaries, which are too shallow or too swift for fish to enter. This explanation of Allen's paradox is speculation, but it is plausible. What can be stated with confidence is that, until an explanation has been provided and verified, calculations of production of invertebrates are premature.

Allen's paradox

Measurement of primary production by algae is difficult because any enclosing of them in containers, the technique that has been used extensively in lakes, alters the environment drastically by cutting off the current. Measurements of changes in the concentration of oxygen in a natural stream have been made, but the difficulty of ascertaining the amount of exchange between air and water has not been entirely overcome. Results suggest that primary production by algae is generally low. In contrast, that by rooted vegetation may be high. (T.T.M./Ed.)

Boundary ecosystems

Between the water and the land, and between one kind of water and another, a diversity of ecosystems exists. Small in extent, they differ from each other more, in some respects, than do standing and flowing waters; for convenience, they are here treated as a group.

BOUNDARY SYSTEMS BETWEEN WATERS

Estuaries. Boundary systems in which there is an appreciable current are most like oceans, lakes, and rivers. Estuaries, in which seawater is diluted by freshwater inflow, share many features with rivers and the ocean. An estuary has a net seaward flow equal to the discharge of its tributary rivers, but that flow is affected by the surge of tidal currents. It also has other distinctive features of circulation. As river water in an estuary reaches seawater, the river water, being lighter, rides on top and, having a considerable velocity, tends to drag the upper part of the seawater underlayer along with it. Under steady-state conditions, this salt water carried seaward by the fresh surface flow is continuously replaced by a landward flow of seawater at greater depths. Variations of this basic pattern are produced either by differences in size and shape of the estuary or by differences in the relative importance of river discharge and tidal mixing. The typical estuary may, however, be thought of as a two-way flow: the seaward-flowing water is lighter, at the surface, and of low salinity, and it courses along the right bank of the estuary (facing downstream) in the Northern Hemisphere and along the left bank in the Southern Hemisphere, because of the inertial effect of the Earth's rotation; the landward flow is heavy, deep, and saline and courses along the bank opposite from the seaward, surface flow.

Flow directions

A pollutant introduced into an estuary, therefore, is partitioned between the two flows and moves landward as well as seaward (see RIVERS: *Rivers as agents of landscape evolution: Estuaries*).

Lagoons. Coastal lagoons in arid regions have a current system that is opposite to that of estuaries. Seawater flows inward on the surface at the inlet to replace the water that is lost from the lagoon by evaporation, and dense water flows outward at depths and maintains the salt balance of the lagoon. In cases in which a lagoon is cut off from the sea, some parts of the lagoon may become two or three times saltier than the sea. The fauna is often that of brackish water—for example, nereides (polychaete worms), certain small clams, the barnacle *Balanus eburneus,* mullets, sand smelts, and cyprinodont fishes. Lagoons supplied with an inflow of fresh water that more than balances its loss due to evaporation may be much less salty than the sea and have a fauna with more fresh-water-tolerant, but related, species (see also OCEANS: *Features of the oceanic shoreline: Lagoons*).

BOUNDARY SYSTEMS BETWEEN WATER AND LAND

Shores. *The shoreline environment.* A great variety of plant and animal life is compressed into a narrow strip along the rim of the sea, stretching from the highest level wetted by waves at the highest tides down through low tide to a level where seaweeds cease to grow. This zone of coastal life can be as much as 100 metres (330 feet) in vertical extent but is usually much less. The upper limit of the zone depends on how high the waves can reach; cliffs exposed to frequent ocean storms may be wetted 30 metres (100 feet) or more above nominal high tide. The lower edge of coastal life may be the edge of the seaweed "forest" or the seaward slope of a coral reef, or it may merge imperceptibly with the benthic life of the seabed, as on many sand beaches. The lower limit of coastal life depends on extent of light penetration and varies from a few metres in muddy waters of temperate and Arctic regions to 100 metres in clear tropical seas.

A factor common to all coastal life is the continuous water movement: the daily or twice-daily rise and fall of the tide, the perpetual to-and-fro of alongshore currents, and the crash and wash of wind-formed waves. By these means the supply of oxygen and nutrients in the water is continuously renewed.

In most parts of the world, the tide is the most obvious environmental influence on the shore. The period of the tide and the extent of the rise vary significantly, and the part of the shore directly influenced—the intertidal zone—also varies greatly in extent and type.

Wave action along the shore also varies, but much more irregularly than the tide. Its influence is difficult to relate to the distribution of coastal life and almost impossible to separate completely from purely tidal factors.

It is obvious that physical forces mold coastal communities. Not so obvious, but just as certain, is that the organisms themselves affect the habitat. Gastropod mollusks such as winkles, limpets, and top shells (turban shells) contribute significantly to erosion of the rock while grazing on attached seaweeds. Erosion from such sources is uneven, resulting in the formation of pits and hollows that encourage other erosive forces. Along the open ocean coasts of Portugal, Spain, France, western Ireland, and the northwestern United States, sea urchins honeycomb the lower rocks. A group of bivalves found on most shores can tunnel into all but the hardest rock.

Organisms' effect on shore habitats

The animals living on sandy and muddy shores vary widely in response to the predominant size of particles making up the shore surface (*e.g.,* whether coarse sand, fine sand, silt, or mud). Many of the animals continually burrow in the deposit and turn it over much in the same way earthworms do on land. Where the deposit is sheltered from wave action, complex burrows and hollows may build up.

Many worms and some mollusks secrete tubes made of limestone or of sand grains cemented together. In some parts of the world, sizable aggregations of these tube-building creatures build up large reef structures near the low tide level of the shore. The greatest reef builders, of course, are the corals and other organisms associated with them in tropical seas.

The ecologist divides up the shore and shallow sea into a series of horizontal belts, or zones, according to the dominant plants or animals found in each (Figure 21). The widely used scheme of T.A. and Anne Stephenson recognizes three main zones and two fringing zones. The supralittoral zone, subject to salt air and some spray but not normally wetted by the highest tides, is inhabited by salt-tolerant lichens and maritime land plants, insects, and coastal birds. On sandy shores this is the seaward side of the dune formation, and in mangrove swamps the upper edge of the buttonwoods. On tropical shores there are animals of marine origin, including robber and ghost crabs. The supralittoral fringe extends from the highest level washed by the waves at high tide down to the upper limit of barnacles. The organisms in this fringing zone are wetted only by the higher tides each month and may thus remain dry for several days at a time. On a rocky shore the typical inhabitants are small winkles and the dark, encrusting marine lichens that form a black band on the

Habitat zonation

Adapted from *Ecology and Field Biology* by Robert Leo Smith, illustrated by Ned Smith (Harper & Row, 1966)

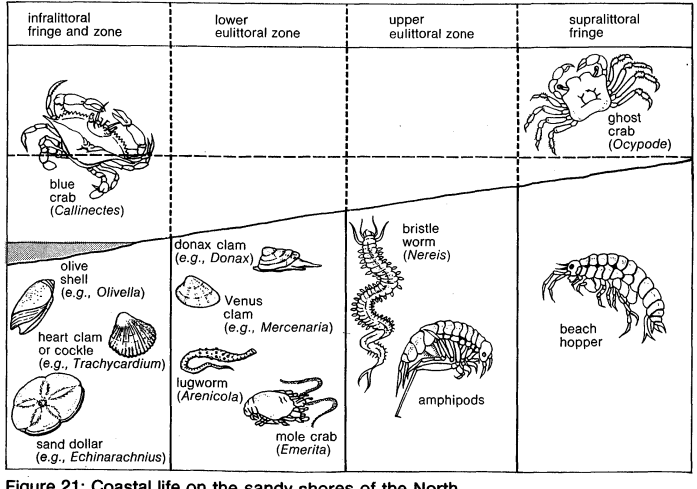


Figure 21: Coastal life on the sandy shores of the North American coast.

rocks (Figure 22). Some grapsid crabs and isopods may be found, though these all take refuge in crevices when disturbed. The red alga laver, or nori (*Porphyra*), and, in the North Atlantic region, two brown algae, *Pelvetia* and spiral wrack (*Fucus spiralis*), occur at the lower end of the zone. The eulittoral, or mid-littoral, zone can be divided into three subzones on most rocky shores. The uppermost part is usually dominated by barnacles (*Chthamalus, Balanus*) and by more of the winkles. In shelter, algae may take over from the barnacles—for example, in the North Atlantic and North Pacific, various species of brown seaweeds (fucoids) and the green alga *Enteromorpha.* The middle region of the eulittoral is often barnacle-dominated like the upper part, but limpets, mussels, and stalked barnacles (*Pollicipes*) may be present; in shelter from wave action, the barnacles tend to be overshadowed or replaced by algal growths. The lowest part of the eulittoral is the easiest to distinguish; it often lacks barnacles and may be dominated by a red algal turf or by an association of limpets and encrusting calcareous algae or by large brown seaweeds. Locally it may show reefs formed from the tubes of marine worms or mollusks or sheets of anemones. The

Infralittoral fringe

lowest part of the intertidal zone is called the infralittoral fringe and is often regarded as an upward extension of the infralittoral (sublittoral) zone. The fringe may be continuously submerged for days and is often characterized by growths of large brown seaweeds. Where the laminarian seaweeds are not present, the infralittoral fringe may take the form of a turf of small red seaweeds of many genera, without large limpets and with masses of encrusting calcareous algae, or in tropical regions it is part of the coral-reef formation. Other more local forms of the infralittoral fringe also can occur: groups of sea urchins, associations of anemones and starfish, and beds of marine grasses, which can mingle with the laminarian seaweeds present. In the Southern Hemisphere a fringe of sea squirts (class Ascidiacea) sometimes dominates this part of the shore.

Adapted from *Ecology and Field Biology* by Robert Leo Smith, illustrated by Ned Smith (Harper & Row, 1966)

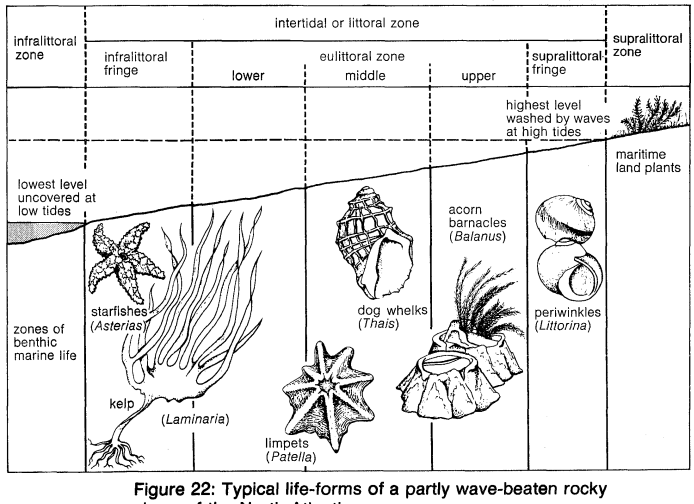


Figure 22: Typical life-forms of a partly wave-beaten rocky shore of the North Atlantic.

As already noted, the infralittoral or sublittoral zone itself is not unlike the infralittoral fringe, and many organisms are common to both. In temperate and cold climates it is preeminently the home of the large seaweeds (laminarians and giant kelps) that can reach up from as much as 30 metres depth to float part of their fronds at the surface. In warmer waters the corresponding algae are much smaller. As the influence of exposure to air and strong light decreases, the hardy forms typical of the intertidal zone become less abundant and their place is taken by more delicate forms; *e.g.,* smaller limpets, gastropod mollusks, sponges, long-spined sea urchins, starfish, and the finer red algae. In the tropics the infralittoral zone is often part of the coral reef formation.

Sand and mud shores

On sand and mud shores the same basic zones can be recognized, though the organisms are different from those described for rocky shores. The supralittoral fringe of a sandy shore is the semiarid region of loose sand near the high-water mark, with varying amounts of dried seaweed fragments and other detritus left by the tide; it is inhabited by sand hoppers, which live in burrows in the damp sand. The corresponding part of a mud shore may carry a typical salt marsh association (see below) or may be a shingle (pebbly) bank with sand hoppers among lines of rotting seaweed. The eulittoral of sand and mud shores is usually inhabited by several types of polychaete worm, the most typical being lugworms. Other worms, such as nereides and spionids, may be commonest near the upper edge, corresponding to the barnacle zone of rocky shores, while other nereides and *Nephtys* may predominate lower down. On wave-beaten shores, the worms are fewer but are accompanied by a large transient population comprising crustaceans, amphipods, isopods (which swim about in the sea at high tide), and the mole crab (*Emerita talpoida*). The fiddler crab (*Uca*) occurs widely on sand and mud shores. With increasing shelter from wave action, a beach becomes less sandy and more muddy and is inhabited by the smaller mollusks, which are common in the lower half of the eulittoral, and in full shelter the larger clams may be found. The infralittoral fringe of a sandy shore is usually inhabited by the same worms and mollusks found higher up, but in addition there are razor clams, cockles, heart urchins, and sand dollars. The infralittoral fringe is less marked than the rocky shores, and many of these organisms continue below the level of low tide. At locations where a few loose stones occur and where water movement is slight, brown seaweed occurs, while in muddy places the marine eelgrass (*Zostera*) forms dense beds.

Character of shoreline communities. One of the most characteristic features of life on the rocky shore is the tenacity with which the organisms attach themselves. The large brown seaweeds have strong disk-shaped or root-like extensions of the thallus by which they hold fast to the rock. Many of the animals (*e.g.,* the barnacles and tube worms) are permanently cemented to rock. Other animals, such as the limpets, have a large muscular foot capable of strong temporary attachment to the rock surface. Some shore animals, such as small crabs, amphipods, and isopods, resist detachment by the waves by fitting themselves into crevices, using some of their legs as anchors. Small crustaceans and mollusks take refuge among the fronds of seaweeds. Perhaps the most highly adapted clinging organ is found in the sucker fish (family Catostomidae), which has pelvic fins modified as vacuum cups.

Attachment methods

Many inhabitants of severely wave-beaten rocky shores also show adaptations that reduce the effects of waves; *e.g.,* the depressed conic shape of limpets and barnacles, the short, relatively unbranched yet flexible and mucus-coated thallus of the seaweeds.

The animals of the shore are ultimately dependent on seaweed and larger plants for food, supplemented by plankton from the ocean and by organic debris from the land. The main herbivores of the shore—the limpets, winkles, and top shells—are all grazers. They feed on the mature, large algae directly or on the smaller recently settled plants and other simple algae that grow on rock surfaces. Pieces of seaweed that accumulate after a period of rough seas also are eaten by winkles and top shells, which can thus act as scavengers.

Most of the crustaceans of rocky shores feed on a wider variety of organisms and also are scavengers. The barnacles that sweep the water passing over them with a net formed from six of their appendages can eat anything from microscopic algae up to crustaceans a few millimetres long. Other crustaceans stir up and sort over the silt that often collects in crevices. Bivalve and vermetid mollusks, as well as some tube worms and sea squirts, feed on floating particles that they remove from the water by filtering.

The predators of rocky shores include dog whelks (*Thais*), phyllodocid worms, crabs, and shore fishes. Many of these predators, especially fishes and starfishes, lurk in pools and crevices or below low-tide mark and come up to feed only at high tide. In contrast, rodents and seabirds feed on mollusks and crustaceans when the tide is low. On sand and mud shores, large plants occur only at the upper and lower levels (*e.g.,* grasses at and above high tide, eelgrass at

low tide), but on very sheltered mud flats some green algae (*Enteromorpha, Chaetomorpha*) form mats on the surface in the eulittoral zone, while in salt marshes there may be clumps or beds of reedlike flowering plants. There may also be films of microscopic algae lying free on the surface or between the particles of the shore. These diatoms and flagellates can show daily and seasonal changes and may be numerous enough to give the surface a greenish or yellowish tinge.

Burrowers and tube formers

Larger animals of sandy and muddy shores include burrowers, which form galleries in the substrate, and tube dwellers, which live in tubes formed from secretions and sand grains or both stuck together. The burrows and tubes are only a few centimetres deep in the case of small amphipods and spionid worms, but half a metre or more with large worms and clams. Many of the animals feed by ingesting sand or mud and sorting out and digesting the organic fraction inside their gut. Some carry out more complex sorting operations to explore the surface of the ooze and suck in lighter particles. A large proportion of the mollusks of beaches are plankton feeders, who lie in the substrate with extensible siphons raised above the surface of the bottom. The sabellid worms found at low-tide levels feed in much the same way.

On the whole, predators are less common in sandy or muddy beaches or are less easy to recognize. There may be predatory gastropods and various small crabs, but much of the predation occurs at high tide, when crustaceans and fishes move in and attack exposed siphons and tentacles of partly buried plankton feeders. At low tide wading birds probe wet places with their beaks or puddle shallow pools with their feet to get at crustaceans and other small animals buried shallowly. Some animals of muddy shores are more or less sedentary (*e.g.*, tube-living worms such as the clymenids and sabellids), but others, like the small clams, are mobile and search for food. Generally, plankton feeders are evenly spread out, while deposit feeders aggregate where the organic content is highest or the particle size easiest to sort.

Productivity of shoreline communities. The large seaweeds are the most obvious primary producers of energy-containing matter. In the northern, cool temperate regions, the fresh weight of brown seaweeds such as knobbed wrack (*Ascophyllum*) can reach 32 kilograms per square metre of shore, and one kilometre of shoreline may hold up to 40 metric tons of one species of laminarian seaweed. The annual production of new organic matter from these plants averages about 10 to 20 metric tons per hectare per year. This production on rocky shores may be compared with production values of 20 to 30 metric tons per hectare per year from the best cultivated crops on land, and 30 to 75 metric tons per hectare per year from salt marshes and tropical swamps.

Primary and secondary producers

The algal symbionts in animal tissue are a special case. Primary production in coral may be as high as 30 to 40 metric tons per hectare per year, though the algal biomass may be no more than 700 grams dry weight per square metre. Of this, the larger proportion belongs to filamentous algae growing within the reef structure below the animal tissue, and not to the symbiotic algae. The total animal tissues in a coral reef may reach 200 g/m^2, and it appears that the primary production just about balances the needs of this animal tissue.

The phytoplankton in the sea is the least productive in terms of unit area and is generally found to be about one to five metric tons per hectare per year. The relatively greater production of coastal communities is a matter of greater actual utilization of energy in the shallow-water habitat and a better supply of nutrients circulated by vigorous water movements.

The production of animal organic matter, the so-called secondary production, is always less than the primary production on which it depends for its existence. On rocky shores the animal biomass may include 200 to 300 grams of dry organic matter per square metre of barnacles, 200 to 2,000 g/m^2 of mussels, but only 60 to 100 g/m^2 of littorinids. The barnacle zone of the eulittoral zone of a Pacific shore has been shown to contain 315 g/m^2 of algal tissue, 2.2 kg/m^2 of barnacles and other filter feeders, and 130 g/m^2 of limpets, winkles, top shells, and other herbivores. In limpet-dominated communities, there may be only from 20 to 50 g/m^2 of these herbivores, and since many of them live a long time and grow slowly, the annual production will be much less.

In sand and mud communities the biomass can be even smaller. For example, high values of the order of 20 to 80 g/m^2 of clams and 4 to 12 g/m^2 of lugworms (*Arenicola*) have been noted in favourable places, but average shores of muddy sand show a total biomass of only 1 to 15 g/m^2, of which the largest component are lugworms. There are as yet no estimates of production of animal matter in the entire community of a sand or mud shore, but in warm waters the small clams may grow fast enough to show an annual production of two to three times the biomass, which is, however, very small (0.2 g/m^2).

It seems that, of the high primary production of shore communities, only a small proportion, less than one-tenth, is turned into animal matter, even less if allowance is made for the contribution of organic matter from the land and the ocean. The shore is less efficient overall than the ocean in spite of having higher primary production, and much of the organic matter of the shore is lost to other communities—*e.g.*, to coastal plankton, to benthic life below the level at which seaweeds grow, and to land animals.

(A.J.So./Ed.)

Other boundary systems between water and land. Coral atolls and salt marshes are both systems in which a solid surface is bathed by a continuously renewed supply of water. In the case of the atoll, it is full-strength seawater, swept over the reef by the trade winds. In the case of the salt marsh, it is estuarine water flushed in and out by the tides.

The coral community is diverse, with many species of organisms. The salt marsh community, however, is simpler, with many individuals of a relatively small number of species. Both systems provide a surface to which producers can attach as they extract nutrient salts from the flow of water past them.

Groundwater is not without life. A sparse fauna lives in subterranean streams, feeding on the detritus washed down from the surface or carried into caves by other animals. Blanched and eyeless groundwater fishes and crustaceans are found chiefly in caves or where wells pierce the water-bearing layer in which they live.

Swamps, marshes, and bogs

In addition to salt marshes, with their abundant tidal flow, there are many kinds of wetlands that have so little flow that they are best considered as semiterrestrial analogues of standing waters. The principal distinction of importance is between swamps and marshes, on the one hand, which have access to mineral-rich groundwater; and bogs, on the other, which are insulated from groundwater by a thick layer of peat and receive most of their water and mineral salts through the atmosphere. The terminology of wetland types is confused by the use of many names, such as fen, mire, and carr, all of which, like swamp, marsh, and bog, are, by origin, simply local names meaning "wet land." For ecological purposes it seems sufficient to distinguish wetlands with trees as swamps; wetlands without them as marshes; and wetlands receiving their water and salts from rain as bogs.

Because of reactions between peat and bog water, the supply of dissolved salts is decreased below that in rain. As a result, the nutrients needed by plants for growth are in short supply, and many bog plants—such as sundews, pitcher plants, and bladderworts—have evolved the habit of trapping insects and other small animals to augment the meagre supply of mineral food in the water.

The great accumulation of acidic peat under a bog—often raised some metres above the level of the surrounding countryside—is an extreme case of one of the general puzzles of aquatic ecosystems: in few of them is all the organic matter produced in photosynthesis broken down completely after the death of organisms. A certain residue of organic matter remains and is preserved in a basin of sediment. Evidently, the sedimentary bacterial community is incapable of anything approaching complete decomposition of the organic matrix in which it lives. It is a curious fact that no microorganisms make use of this undecom-

posed residue—a fact to which human beings owe their supply of fossil fuels, such as coal, gas, and oil.

The best basis for comparison among ecosystems is probably photosynthesis by green plants. There are boundary systems without photosynthesizers—a cave stream, for example—but they have no primary production. The food that sustains them is imported from outside the system. They may have a productivity of blind crayfish, but their net production of organic material is negative.

Unfortunately, there is no generally applicable method for measuring primary productivity in all boundary systems. In places with a strong seasonal pulse of growth—as in a temperate cattail marsh—the maximal amount of plant material present after a growing season can be harvested in order to estimate net primary production, which is somewhat smaller than the gross rate because it takes no account of food consumed by the cattails (family Typhaceae) to maintain themselves over the season of growth or of cattail substance that may have been consumed by animals or bacteria during the same span of time. Other methods depend on measuring changes in the quantities of various substances, commonly oxygen or carbon dioxide, in the medium in which the plants are growing.

(D.A.L./Ed.)

TERRESTRIAL ECOSYSTEMS

A terrestrial ecosystem is a landscape that supports life; its components include earth materials at least as deep as roots extend, boundary layers of air, films of water, and interacting organisms. It is these components and the interactions and diverse relations among them that fulfill the concept of an ecological system.

Land as a medium for life. The complexes of terrestrial ecosystems in Table 9 indicate the wide range of temperatures and precipitation encountered in different communities throughout the world. If monthly or hourly extremes of these and other climatic variables are considered, as well as the conditions at different levels (microenvironments) above or below the ground, some idea may be gained of the constraints that land organisms had to overcome, during their evolution, after leaving the confines of water. Organisms that are available to colonize a given area and the local physical factors of soil, topography, and moisture availability limit what can arrive, survive, and thrive.

Pioneer organisms

Pioneer organisms may change a bare land surface, thus making it suitable for other species, which, in turn, further modify the local ecosystem and prepare the way for still later generations. Some combinations of species persist for hundreds or thousands of years if average rates of change of soil and other internal system conditions are slow enough. For intervals of more than a few thousand years, however, gradual shifts in the boundary conditions are imposed from larger systems (*e.g.*, climatic changes bringing on glaciers and major relocation of populations and life zones; gradual shifts and extinctions of species available over wide regions). Undisturbed natural systems adjust accordingly on a time scale that is long by comparison with the recovery times of communities resulting from perturbations of both human and natural impacts. Organisms along the evaporating fringes of the ice cap creeping into Antarctica's dry valleys may encounter water during only a few days in summer. Mountains—from Antarctica, along many Andean summits, around the Plateau of Tibet, and locally elsewhere—reach altitudes above which only wind-borne, wind-nourished life (with few or no vascular plants) persists. These and other extreme outposts of Earth's life were colonized from places where the obstacles to survival were not so severe.

The use of studies of successively smaller ecosystems and subsystems—not in isolation from one another but as nested in the progressively large systems—provides the breadth of data needed to comprehend fully the workings of the global ecosystem.

Smaller systems are dependent on larger ones in many ways. Like space vehicles or stations, some fascinating natural ecosystems depend on the importation of outside energy and some nutrients.

Fewer than 133 million square kilometres (51 million square miles) are accounted for by glacier-free land, compared with 362 million square kilometres (140 million square miles) by the oceans and adjacent waters. Small, but important, fractions of these areas lie in shifting boundary zones between land and water. Floodplain, swamp, and delta complexes include some of the world's most productive landscapes. Tides and currents, however, wash away some nutrients from these complexes—and from grassy salt marshes—fertilizing shallow marine waters and stimulating the aquatic food chain growth.

Even regions so dry that they have no exterior outlets of water to the sea gather saline waters in some rare storms or during moist years. Adjacent flatlands of sodium chloride and even sodium carbonate may be mostly barren but yet support salt-tolerant plants at certain times and places after water flows in from surrounding highlands.

As noted previously, such cases of shifting boundary conditions of land and water are of interest partly because special displays of clear-cut environmental control of life can be sorted out in the extremes. Similarly, bogs and mires receive widespread interest as wetlands dispersed widely, from the vast lowland Arctic tundras through boreal forest (taiga) belts to a few places farther south.

The ancestors of present land life evolved in fresh and salt waters with the appropriate membranes and metabolic processes for selecting the scarce elements and compounds that each cell requires for survival, possible genetic change (mutation), and further evolution.

Adaptation to land living

Limiting factors to living on land. Many adaptations were required before organisms could leave the seas and live on the land. Most green plants evolved from algae that had fluid sacs (vacuoles) inside cells which helped maintain water and chemical balances in the living cell substance. Vascular conducting systems help pipe water to all tissues. Impermeable outer coverings control outgoing water and incoming carbon dioxide through special openings called stomates, which are regulated by guard cells. Specialized adaptations have allowed desert succulents to take in carbon dioxide at night, when temperature and stress of moisture loss are low. Certain organisms have drought-resistant cytoplasm. Lichens, fungi that enclose algal cells in a biological partnership, survive far into the polar regions and into temperate and tropical deserts. The right species can photosynthesize for brief periods under favourable conditions of moisture or humidity, while resisting attrition of their accumulated assets of water and organic compounds during the long times of intervening aridity. In the dry as well as in fairly wet areas of tundra and taiga and especially in the rocky parts of mountain landscapes, other lichens and many mosses illustrate the patchwork patterns of colonization such as might have occurred as the continents first became populated by green plants. Microbes and animals were ready to follow, consuming dead or live organic matter, or both.

Major life forms and classes. *Growth habits and indicator organisms.* A classic ecological approach to interpreting environmental constraints on regional plant formations or local community and site types within regions is based on the kind of plant group, or life-form, that predominates (see Table 8). Anglo-American and Soviet forestry traditions have tended to emphasize the indicator value of such life-forms and, where possible, the genus and species of the main community or perhaps each of several layers (strata) approximating a class. The preliminary indications of community change are based on the prevailing heights of plants constituting the class and include the appearance of different species combinations in seedling or other young strata as compared with the vegetation that currently dominates.

In a complex ecosystem, however, the species of greatest abundance is frequently not such a sensitive indicator of habitat conditions as are others having lesser

Table 8: Life-Forms of Primary Producers

plant life-form group	height	Raunkaier categories*
Epiphytes (growing on plants of the following groups)		
Woody or evergreen perennials (bud-bearing shoots above ground)	buds more than 0.25 metre above ground	Phanerophytes
Trees (one to few stems; self-pruning)	taller than 30 metres	Megaphanerophytes
	10–30 metres	Mesophanerophytes
	less than 10 metres	Microphanerophytes
Shrubs (smaller woody plants, often with many stems and with shoot replacement)	taller than 0.5 metre	Microphanerophytes
	less than 0.5 metre	Nanophanerophytes
Subshrubs	less than 0.25 metre	Chamaephytes
Semishrubs	partly dying annually	Chamaephytes
Lichen–moss layer(s)	buds near ground surface	bryoid-chamaephyte
Perennial (or biennial) herbs		
Grasslike (including sedges, rushes, etc.)	buds near ground surface	Hemicryptophytes
Forbs (including non-grassy herbaceous plants)	buds below ground	Geophytes
Water plants	buds or bulbs below water	Hydrophytes
Marsh plants	buds or bulbs in wet ground	Hydrophytes
Annuals		
Summer, winter, and facultative perennial	"buds" in seed	Therophytes
Special forms†		

*A traditional standard terminology for life-forms as devised about 1900 by C. Raunkaier, a Danish botanist. †Including bamboos (woody grasses), tuft plants, climbers (lianas), succulents (stem and leafy).
Source: After H. Ellenberg, D. Mueller-Dombois, A. Kuchler, W. McGinnies, and others.

Eurasian ecosystem classification

abundance but greater fidelity or differential occurrence. Several Eurasian schools of ecosystem classification and newer techniques of statistical analysis have introduced the relative importance of various combinations of factors distinguishing one ecosystem type from another.

Ecologically similar species within major genera are recognizable among distant areas—especially in the Northern Hemisphere. On the other hand, the physical constraints of water relations and energy balance have led to convergent evolution of similar life-forms from genetically distant stock. Notable examples among plants include the hard-leaved (sclerophyll), mostly evergreen scrub and woodland or forest in the winter-rain, summer-drought (Mediterranean-type) climates of Chile, Australia, and California, as well as southernmost Europe, northernmost and southwest (Cape) Africa, and areas of southwest Asia.

The evolution, successional change, and seasonal movements of animals are related directly to plants that provide them food and shelter. In the cold or dry regions noted earlier, herbivorous and especially predatory and scavenging animals may have to range widely to survive and reproduce. Warmer, humid regions include migrants as well, but they also allow many other organisms within almost any habitat.

Classification by habitat. The habitat—where an organism lives—is frequently used to categorize the living things in an ecosystem. The range of such categories within terrestrial ecosystems may be divided broadly into soil organisms, rooted plants, organisms attached to plants, and free-moving animals.

Soil organisms include a size range from microbiota (microorganisms and minute animals) through mesobiota (arthropods and smaller larvae) to macrobiota (roots of plants, larger invertebrates and burrowing vertebrates). Plants provide an aboveground layer consisting of prostrate herbaceous plants, shrubs, vines, and trees. The organisms that attach to such plants are called epiphytes; those that cling but can move are often called periphytes. The freely moving animals—both walking and flying—are referred to as permeants.

Classification by niche. Not only the habitats but the niches (operating roles of organisms throughout their life) are involved in a careful consideration of ecosystems and their evolving populations. Just as people "find their own niches" in a complex society and adapt, through learning, to their own special vocations, an indefinitely large number of species possibilities have been narrowed down during the course of evolution to those especially adapted to definite and perhaps discrete "vocations" in a natural ecosystem.

Considering first the environment as conditioning a local ecosystem through climate, the presence of other organisms, preexisting surface, surrounding relief, and related moisture factors, then within a specified local area, or volume, on the Earth's surface, the range of variables is seen to be much smaller than in the entire biosphere.

Measuring species' dimensions

If each available species in neighbouring ecosystems is allotted a dimension, for example, those that never become established in a local ecosystem in question measure zero. Whatever the measure used, pioneer species are those scoring high soon after a bare or disturbed area has opportunity for recovery but declining—usually to zero—as waves of replacement species arrive. Self-perpetuating (climax) species may have more dimensions than the pioneer species or those that typically rise and disappear in stages of succession.

All of the organisms with which a species coexists at one time or another (a kind of intersection set) constitute part of its niche. Microclimatic variables within the local habitat occupied by the species limit the range of favourable conditions. Soil conditions, microtopography, and moisture conditions, all of which change with the other variables, are additional ecosystem properties that influence most of its inhabitants directly or indirectly. Also upper and lower limits of physiological tolerance influence where and how long a species can persist—either locally, on the time scale of a species life cycle, or globally, on the time scale of evolution.

Not only mere survival but also the probable rates of intake of resources, of elimination, of growth and reproduction, and of death for each kind of organism are influenced by the other variables. An organism can fill any of three niches at any given time in its life: it can be a producer of goods, like green plants; a consumer of goods, like most animals; or a decomposer of goods, like many microorganisms. The entire complex constitutes a food web or network. Energy passes along this net as in a power grid and is dispersed as heat along the way (see above *The zone of life: an overview: Energy flow and material cycling*).

Major divisions of the land environment. *Polar regions and tundra.* The colder divisions of the land environment can be grouped broadly as tundra and polar barrens, which include cold "deserts" and spotty tundra (Table 9). Bogs and mires accumulate dead carbon wherever decomposition fails to match production over a long period. Altitudinal and local mixtures of tundra dwarf scrub and meadow extend to many boreal mountains (and also to Alpine temperate and tropical areas tabulated under grassland in Table 9). Subpolar areas also include cold maritime herb meadow and scrub, bogs with dwarf woods, and open-wooded or tall-shrub "tundra."

Boreal and temperate forest complexes. The boreal forest (taiga) is sometimes divided into northern (transitional to tundra), middle (open woodland to closed forest), and southern (denser forest) zones. Estimates of typical forest area (and total carbon in living organisms) are separated from figures for unwooded and sparsely wooded bogs on peaty soils. Although "boreal" is frequently used in the general sense to refer to anything northern, it is best distinguished from several other predominantly coniferous zones called semiboreal. These include the subalpine and montane zones of mountains, as in the North American Western Cordillera. A mixed-forest zone in eastern and western Eurasia has some broad-leaved species besides the usual poplar and birch that are common in the taiga and its grassland-fringe zone of western Siberia and south central Canada. Except in these and a few other areas in which aridity and local relief complicate the pattern, a broad climatic outline around most boreal and semiboreal ecosystems is bounded on the north by the farthest extent

Boundaries of boreal and semiboreal ecosystems

of the 10° C (50° F) July average temperature and on the south by areas with more than four months above this temperature.

Temperate areas include some predominantly coniferous forests—*e.g.*, from the central British Columbian interior to lower elevations around Alpine-subalpine-montane complexes farther inland in North America. Except in parts of a cool coastal belt and some other local rain forests, conifer growth is usually limited by droughts as well as by sporadic insect outbreaks. These repeated disturbances, plus sporadic lightning fires, have influenced the extent of most conifer stands. Giant evergreen forests locally attain much greater than average carbon mass. Evergreens of lesser stature, dwarfed in areas of salt-wind exposure, may line these and other vegetation belts in coastal strips too narrow to map on a global scale.

Temperate summer-green or cold-deciduous forests, in contrast to the drought-deciduous forests of tropical monsoon regions, prevailed originally in the now industrialized areas of the northeastern United States, mid-latitude Europe, central eastern Asia, and locally elsewhere. Conifers peculiar to special habitats, especially in mountains and sand plains, also may have been part of the original vegetation and certainly have been extended in plantings in much of the cool temperate zone. In North America, for example, below the latitudes or altitudes where spruce fir was common, pines were important pioneer species on burned-over forestland and abandoned clearings; eastern hemlock, which is more shade-tolerant, persists or even expands in areas protected from fire, especially in the Great Lakes and Appalachian regions.

Warm temperate zone

The warm temperate zone, as generally understood in English, is referred to as subtropical in Soviet atlases and literature (which provided many of the preliminary estimates in the tables included in this section). Broad-leaved forests are partly deciduous but increasingly evergreen near the southern coasts of the United States, Japan, and China. Only a few conifers are important.

Mostly evergreen broad-leaved forests prevailed in humid portions of the Southern Hemisphere areas: eucalyptus in Australia, false beeches (*Nothofagus*) in New Zealand and the southern Andes. Here oceanic climates moderate winter cold. Wet hard-leaved (sclerophyll) forest and temperate rain forest ecosystems are localized on favourable sites; dry sclerophyll communities of varied height and tree spacing are especially prominent in both temperate and subtropical parts of Australia. Monterey pine (*Pinus radiata*) outgrows other planted conifers and the several native evergreen genera.

Scrublands. Forest and woodland areas include also maquis and garigue scrub in the Mediterranean area, chaparral in California, and similar vegetation capable of maintaining its leaves through dry summers or wet winters or spring seasons. Such scrublands merge into steppes toward regions of decreasing precipitation.

Grasslands. As the cool temperate forests interfinger with grassland in the Soviet Union and the northern United States, so the warm temperate forest alternates with grassland on different substrates west of 95° W longitude in the south central United States. As in tropical savannas, burned fringes of all the foregoing forests included open woodlands and sparser groupings of trees. Semiarid grassland is here typical throughout, but gradations to cold Alpine vegetation and semidesert are also evident.

Tall grassland includes the belt of meadow steppe south of the Eurasian forest and the North American prairie. Short or mixed grassland prevails in the next drier belt. From Mongolia westward, barely interrupted by the mountain meadows and terraced conifer forests of the Altai and other mountains, the shortgrass steppes offered a corridor

Table 9: Approximate Zonal Distribution of Major Preagricultural Ecosystems

major bioclimatic zones, temperature, relief	moisture	ecosystem type	area* (000,000 sq km)
Polar to subpolar area			
Polar	arid to humid	tundra, barren	7.86
Polar to boreal	wet peat	open bog and mire	1.30
Boreal Alpine	mostly humid	"tundra" meadow	1.58
Subpolar	mostly humid	scrub, herb, bog-woods	2.80
Total			13.54
Boreal and semiboreal area			
Boreal	mostly humid	taiga	10.10
Semiboreal or montane	mostly humid	mostly coniferous	6.91
Total			17.01
Temperate forest complexes			
Cool or montane	mostly humid	mostly coniferous	3.77
Cool	humid	mostly deciduous	3.76
Warm	humid	evergreen, deciduous	5.76
Warm or montane	mostly semiarid	woodland, scrub	3.83
Warm (moist site)	arid to semiarid	herb-woods mix	1.07
Total			18.19
Grassland area			
Alpine	mostly humid	mountain meadow	1.81
Temperate	humid	tallgrass, marsh	1.05
Temperate or montane	semiarid	grass, scrub, puna	9.43
Cold tropical montane	humid	paramo, scrub	0.94
Tropical, subtropical	mostly semiarid	grassy savanna	9.73
Total			22.96
Desert and semidesert area			
Temperate and tropical	semiarid to arid	sandscapes	5.77
Cold montane	semiarid to arid	mountain desert	3.20
Other temperate	arid	desert	10.45
Tropical and subtropical	very arid to arid	desert	9.93
Total			29.35
Tropical forest complex			
Hot (moist site)	humid	lowland rain forest	4.56
Hot	mostly humid	semievergreen	8.83
Hot to montane	humid-dry	monsoon deciduous	1.18
Warm to montane	humid	mostly evergreen	2.42
Total			16.99
Other tropical woodland-savanna complexes			
Hot (moist site)	arid to semiarid	herb-woods mix	0.32
Hot	humid-dry	woodland, savanna	8.93
Hot	dry-humid	scrub, wood, savanna	5.12
Total			14.37

*Areas regrouped after N.I. Bazilevich, L.E. Rodin, and Rozov.

for Asian horsemen to the Caucasus, the Crimea, and central Europe.

In the Southern Hemisphere occur the pampas of Uruguay and Argentina and the grassy parts of the South African veld, which share the scattered parklike tree growth of the subtropical savannas. With woodland or scrub mixed in, these latter are marked by seasonal moisture and fire.

Deserts and semidesert areas. Arid lands, climatically defined, include mixtures of grassy and open scrub or special life-forms. Cool semidesert steppes typically are marked in the north by silvery sagebrushes (*Artemisia*). Extremes of hot and cool temperatures, in addition to low moisture, further constrain plant growth in the warm semidesert. Saline deserts are interspersed among both of the above arid lands and even in some grasslands but occur in greatest extent in central Asia.

In Table 9, sandy landscapes, especially those with dunes, are probably estimated conservatively, but biomass is not great even in semiarid, grassy sand hills or in the saxaul (*Haloxylon*) sand woodlands actively cultivated in the Kara-Kum Desert. Almost 10 million square kilometres (4 million square miles), from semidesert savanna to absolute barren deserts of several kinds, are included as tropical desert.

Jungles and rain forests. When these and transitional semideciduous and semievergreen forest or woodland are taken into account, the area of completely evergreen tropical forest is more limited than many maps suggest. Substantial tropical or equatorial rain forests occupy only a fraction of this total area. Sites that are enriched by nutrients from upstream or upslope yield exceptionally high production, but only over still more limited areas, which have long been the first to attract shifting cultivation.

Other tropical-subtropical woodland complexes. Below the highest snowfields are equatorial mountains that include treeless zones with temperatures near freezing on most days of the year. These areas, with a peculiar combination of low and exceptionally tall cold-resistant herbs, are called paramo in Colombia and Ecuador. Far less localized are elfin woodlands and scrub vegetation, widespread near the summit of mountains in both the tropics and subtropics. Evergreen subtropical mountain forests are typical of monsoon regions that have drought-deciduous forests in the lowlands.

Development of savanna and scrub

Natural deterioration by leaching of nutrients and by clearing and erosion have already created much savanna and scrub in the coastal rain forest belt of eastern Brazil and in many areas of the Amazon basin. In West Africa, rain forest and moist evergreen forest appear to be more vulnerable to conversion to grassy savanna than are drier types of ecosystems because fire-resistant trees are not already present. Moist semideciduous forest, dry deciduous forest, and woodland are considered more likely for conversion to savanna woodlands through selection favouring resprouting of resistant trees already mixed in the vegetation or available for invasion from the vegetation of surrounding regions. Original savanna woodland maintains resistant trees but has opened up to allow more grass or scrub since burning has become an annual management tool—especially where fires are mostly started late in the dry season. (J.S.O./Ed.)

Polar barrens and tundra

The polar regions comprise the Arctic, in the north, and the Antarctic, in the south, and although these two are often confused they are indeed poles apart in their systems of animal and plant life, or biomes. The few similarities that exist result chiefly from the coldness, widespread aridity, and level of latitudes. (Because of the latitude, day and night practically divide the year: in both the Arctic and Antarctic the sun is continuously above the horizon in summer and below it in winter.) Even the criterion of coldness is useful mainly in contrast with more hospitable regions; although in the Arctic there is probably no lowland area that does not enjoy for at least one month in each year a mean temperature appreciably above the freezing point, in the Antarctic only the offshore islands and the coast of the Antarctic Peninsula are favoured with this warming.

Because the arbitrary latitudinal Arctic and Antarctic circles have no meaning as natural boundaries for plants and animals, biologists regard the polar regions in general as lying poleward of the limits of tree growth. One generally recognized natural boundary of the Antarctic is the Antarctic Convergence—a delimitation in the sea, where cold surface-water layers spreading northward from the Antarctic continent meet with and sink below warmer, mixed subantarctic waters. This boundary embraces to its south some of the subantarctic islands, such as Heard Island and South Georgia.

THE ENVIRONMENTAL SETTING

Fossil evidence indicates that in earlier geologic ages many parts of the Arctic and Antarctic had temperate or warmer climates favouring luxuriant tree growth. Over millions of years the change to relative barrenness has occurred probably both through variations in world climate (the ice caps are no more than 4.5 million years old) and through the drifting of landmasses poleward from lower latitudes. In any case the polar regions today—especially the Antarctic—are characterized by extremely harsh climates. Precipitation is generally so low—often only a few centimetres a year—in the highest latitudes that, coupled with the low temperatures and permanently frozen subsoil, there prevails a desertlike barrenness. Nevertheless, the availability of snow meltwater, often throughout the short growing season, and the long-lasting summer daylight favour some limited plant growth.

Desertlike barrenness

The Arctic. The highest latitudes in the north are occupied by the Arctic Ocean, the land farthest north being in Greenland and the Arctic Islands of Canada. To the south stretch islands or major landmasses, which in many sectors of the northern polar cap extend into the subarctic and then temperate zones, thus affording more or less continuous dispersal routes for plants and animals from the south. The Arctic Ocean is deep, but the surface is largely covered by floating ice. The land area is commonly rugged, with widespread mountain ranges and glaciers descending to the sea—especially in Greenland, occupied by the world's second largest ice cap (see ARCTIC, THE).

The Antarctic. The highest latitudes in the south are occupied by the continent of Antarctica, which is largely covered under the world's greatest ice cap. Only on projecting mountains (nunataks) and limited ice-free tracts near the shores can land plants grow. The continent is far more isolated than is any Arctic land, being fully 800 kilometres (500 miles) from the nearest continental landmass, though scattered and usually small subantarctic islands exist and probably act as "stepping stones" for some limited plant and animal dispersal. Nevertheless, for practical purposes the Antarctic is isolated by formidable sea barriers that have made it extremely difficult for colonization by plants and animals since the last glacial period (see ANTARCTICA).

THE BIOTIC COMPONENT

Arctic and Antarctic plants and animals are well adapted to their harsh and inhospitable environments. The form of adaptation is often similar in both polar regions, but few species are common to the two. While their terrestrial and freshwater ecosystems are widely different, the far more prolific marine ecosystems are comparable; indeed, in both the Arctic and the Antarctic, it is chiefly in the seas that life abounds.

Vegetation. Polar food cycles, from plant producers through various animal consumers, are relatively simple, with few organisms linked together. Unfavourable climatic and soil conditions combine to limit severely plant growth in the Antarctic. But in the Arctic, vegetation is often fairly luxuriant in the southern tundra ranges, and its component plants are diverse—as also are the dependent animal species. Of Arctic flowering plants and ferns, about 1,000 different species are known from regions north of the tree line. Some of these are confined to the Arctic, but the vast majority range far southward—especially on the mountain slopes above the tree line in temperate regions.

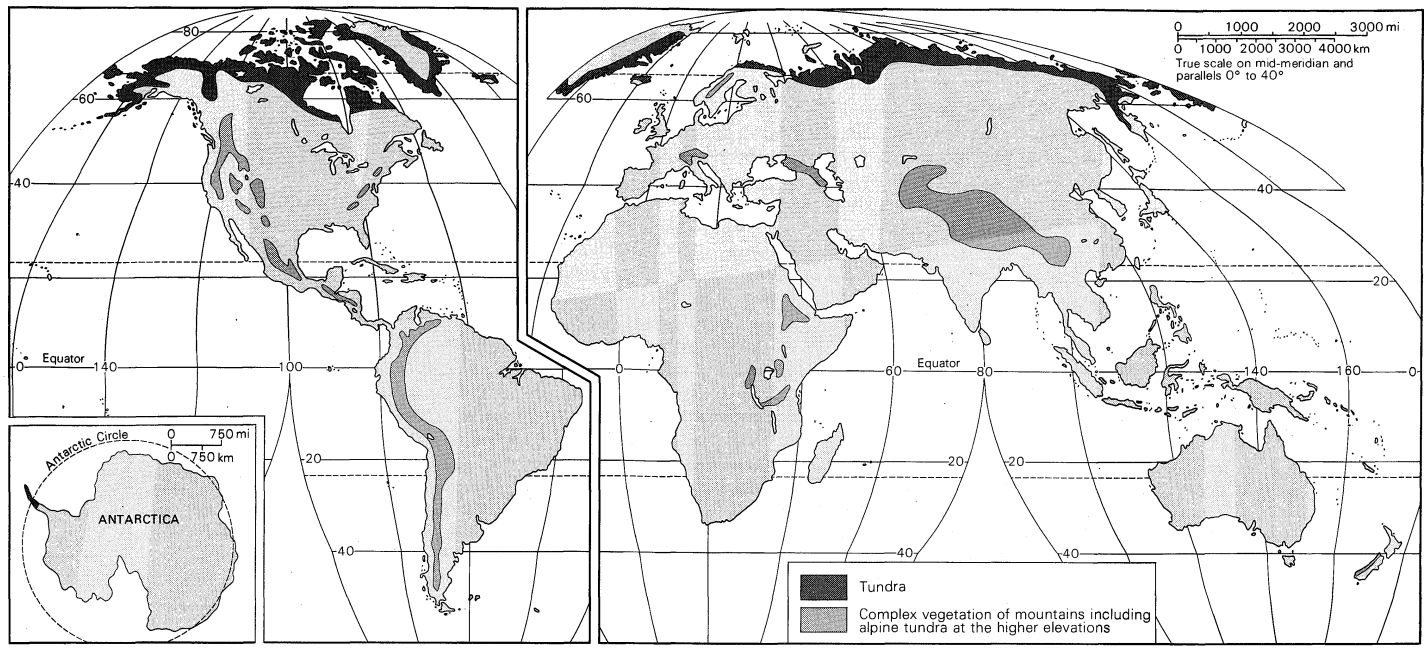


Figure 23: World distribution of tundra and mountain vegetation.
Adapted from *Biological Sciences Curriculum Study Green Version High School Biology*, 2nd ed.; Chicago: Rand McNally & Co., 1968

Such vegetation is known as Alpine tundra. Over much of the Arctic tundra and in many Alpine tundras the general appearance of the vegetation is that of a greenish brown, low grassland. While plants do not remain in flower for more than a few days or weeks, the blossoms are generally large in relation to plant size and are colourful, especially in Alpine habitats.

Coastal plain areas

Across the southerly Arctic tundra, with vast areas of low relief, boggy peat soils with an abundance of lakes and meandering rivers prevail. These coastal plain areas are dominated by sedges and cotton grass, and mosses including peat moss (*Sphagnum*) are common. On slightly elevated sites, often only 15 to 60 centimetres (6 to 24 inches) above the wet peaty soils, low willows, grasses, and rushes occur. Taller willows, grasses, and plants in the sunflower and legume families are common on the sands and gravels of riverbanks.

In marked contrast to the Arctic, the Antarctic proper possesses only three native flowering plants: two grasses of the hair grass genus (*Deschampsia*) and a relative of the pinks, *Colobanthus quitensis.* Although they are limited to favourable exposures in the peninsula sector, they also occur on the subantarctic islands and in southernmost South America. Apart from some algae and other relatively simple organisms, terrestrial life is restricted to the 4 percent of land area that is ice-free. There occur more than 70 species of mosses but relatively few liverworts. Lichens are the most successful and widespread group of plants in Antarctica because they have the ability to colonize bare rock surfaces and they have physiologically adapted to low temperatures and arid conditions; about 400 species have been recorded from the continent, although this number may be reduced by critical taxonomic study in the future. Lichens extend to slightly over 86° S and include *Verrucaria serpuloides,* the only known marine lichen, which passes its entire existence underwater. Of algae, there are approximately 360 nonmarine species that have been recorded from Antarctic ice, snow, soils, and temporary summer ponds. The snow algae, which give rise to reddish, yellow, or green patches of snow, are found only in the coastal areas where permanent snowfields undergo extensive annual thawing.

Fungi and bacteria

Compared with the Arctic, fungi and bacteria are few in the Antarctic. Bacteria that were isolated from lakes, and candida yeasts isolated from rock detritus and snow, proved remarkably well adapted to low temperatures. In the seas more than 200 species of algae have been recorded, the majority being diatoms; bottom-dwelling algae in shallow waters sometimes form extensive miniature "forests."

Animal life. Arctic animals tend to be far less numerous than plants, in both species and individuals. In North America the most notable among Arctic mammals on land are the Arctic hare, wolf, fox, collared lemming, polar bear (chiefly along coasts), short-tailed weasel, caribou, and musk-ox. Except for the musk-ox, these animals are all represented in Arctic Eurasia by the same or related species. All these mammals reach the far north, most of them practically to the northernmost limits of land, but some others, including the wolverine, shrews, ground squirrels, and some additional small rodents and weasels, range well into the Arctic without reaching so far north. This is also the case with the brown bear, red fox, and ermine.

Alpine tundras are similarly limited in the number of animal species and diversity. Accumulating evidence indicates that Alpine animal life in the Northern Hemisphere evolved in the central Asian highlands and spread to Europe via mountain systems and to North America via the Bering Strait.

Many animals found in Alpine tundras are not especially adapted for year-round Alpine life and undergo vertical migrations, descending into the less severe forest environments in winter and returning to the heights in summer. Such mammals include the mountain sheep, ibex, chamois, several wildcats, and many birds. Moun-

Leonard Lee Rue III

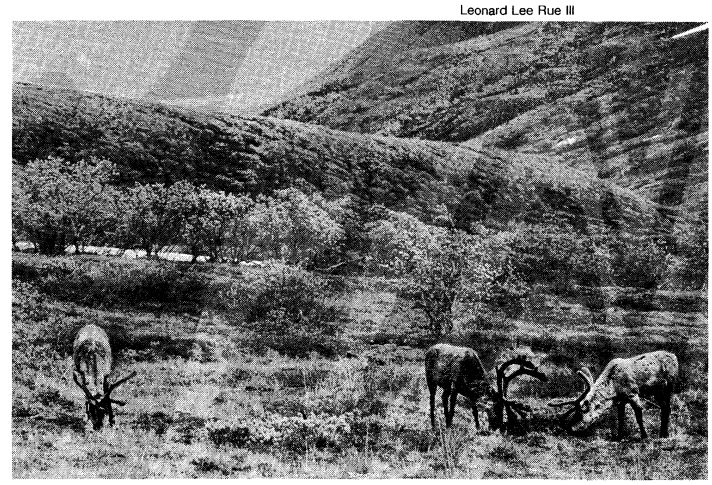


Figure 24: Caribou (*Rangifer*) feeding on mosses, lichens, and dwarf trees common to the tundra.

tain goats spend more time in winter at higher elevations than do ptarmigan.

In contrast with Arctic tundra mammals, some Alpine ones hibernate, such as marmots, ground squirrels, and zapodids. These animals consume large amounts of vegetation in summer and early fall before hibernation begins. Other small mammals, such as the pika and voles, cache hay in the fall for winter feeding, while rabbits and others forage as they can in winter. Foxes range over large areas of Alpine habitats in winter.

Many tundra animals sport white coats in winter, among them the foxes, wolves, and ptarmigan. This camouflage helps both predator and prey: predators can steal up without detection and prey can hide easily in the snow.

Arctic insects

Although the number of species of Arctic insects is small compared with that of temperate regions, those that are present are successful. Arctic insects resist freezing winter temperatures. In some species a high glycerol content acts as an "antifreeze" to lower the temperature at which freezing occurs. Many tundra insects and spiders are dark in colour, as a result of which they absorb more sunlight and maintain higher body temperatures. Some of the tundra species of blackflies and mosquitoes do not require a blood meal before depositing their eggs, in contrast to their temperate-region counterparts.

The Arctic seas are the habitat of various seals, whales, walrus, and narwhals, while belugas frequent narrow inlets and the mouths of rivers. (Humans in ever-increasing numbers threaten to change the fragile ecosystems of the Arctic as never before.)

Although a few marine birds range over the Arctic Ocean north of the northernmost coasts, most Arctic birds are limited to the vicinity of land and southward. Almost all of them migrate south to spend the winter, but Arctic ptarmigan (*Legopus*) regularly and snowy owls (*Nyetea scandiaca*) sometimes remain in the far north. On land, numerous ducks, geese, and waders, particularly, come to nest and pass the summer, as do some swans, while pipits and ravens are widespread, as are various falcons and other birds of prey. Often nesting in vast numbers collectively on birdcliffs or on the ground are various kinds of guillemots (murres) and terns, as well as puffins, little auks, and fulmars. Loons, or divers, call eerily from the numerous shallow lakes, and various kinds of gulls and skuas, or jaegers, screech from above. The Arctic, in fact, seems to abound in bird life in summer. At this time of year also the air inland (except in the far north) teems with insects, especially mosquitoes. Still farther south the insects probably contribute the greatest number of species to the total fauna; they diminish drastically in the far reaches of the north, except for the wingless springtails (order Collembola), which occur even on snow and ice.

Scarcity of animal life in the Antarctic

Like the land vegetation, the animal life of Antarctica is extremely poor. Truly terrestrial vertebrates are lacking, though various sea-feeding penguins and fulmars and other petrels come ashore to breed—as do also the great skua, Dominican gull, Antarctic tern, and Antarctic shag. There are four kinds of truly Antarctic seals—the Weddell seal (*Leptonychotes weddelli*), the leopard seal (*Hydrurga leptonyx*), the Ross seal (*Ommatophoca rossi*), and the crabeater seal (*Lobodon carcinophagus*), of which the last-named is the most abundant seal in the world, numbering an estimated five to eight million individuals. Several of the largest species of whale (order Cetacea) are still found in the open seas around Antarctica, though their populations have been drastically depleted by excessive hunting. By contrast the largest terrestrial animals of these rigorous regions are two species of small midges (family Chironomidae), one of which is winged and the other wingless; these, in addition to a limited array of mites, springtails, microscopic rotifers, tardigrades, protozoans, nematodes, and rare enchytraeid worms, constitute the native land fauna.

COMMUNITY STRUCTURE

Unlike the situation in most parts of the world, where plants form the main embellishment on land, the Arctic and, to a far greater extent, the Antarctic tend to have the framework of their terrestrial ecosystems dominated by the physical structures that locally make up the Earth's surface. In the Antarctic, nesting penguins, because of their accumulated mineral-rich guano (excrement), may create barrens on which no plant can grow. Thus on land it is chiefly in the low-Arctic and the subantarctic islands that plants really affect the landscape.

In the polar regions the struggle of the totality of organisms tends to be more with the exacting environments than with one another, though some competition for space and food does exist. This occurs among both animals and plants, but, in general, what one immediately sees of living materials in these regions consists almost entirely of plants, the greatest development of which is the Arctic tundra of low-growing species interspersed with lichen-covered rock fields and numberless ponds.

Besides marshy and other seaside vegetational types of limited extent, there may be extensive areas of sedges and cotton grasses and of true grasses, such as *Arctophila* and *Arctagrostis,* around lakes and ponds in the Arctic. Luxuriant "patchwork quilt" communities of mixed mosses and lichens are often found around and above bird-cliffs, nourished by guano. Where the snow drifts so deeply in winter that the growing season is drastically reduced by its late melting, zoned series of vegetation types develop. The seas abound in tiny free-drifting organisms, or plankton, of which the plant component, or phytoplankton, occur as deep as light penetration allows; these last support a wealth of crustacean and other zooplankton and, ultimately, fish and mammals (*e.g.,* seals and whales). Even floating ice bears a considerable population of diatoms on both its upper and lower surfaces, making the ice appear a dirty brown in summer.

Antarctic habitats

The Antarctic, by contrast, is impoverished. There is no true tundra on the continent, but only occasional thin tracts of sparse mossy vegetation and algal and lichen patches in especially favoured places. These tracts are found along coasts where the climate is mildest; they contribute to the peat formations on the west side of the Antarctic Peninsula. Other terrestrial habitats include rocky coastlines, the slopes of mountains projecting above the continental ice cap, glacial moraines, summer meltwater channels, and warmed areas around volcanically active fumaroles (holes emitting hot gases and vapour). There are also some lakes and numerous freshwater ponds and puddles. In all such areas on land and in fresh waters, however, the biomass is such that vegetation appears to be virtually lacking except in especially favoured areas that are merely tinted by it. By contrast, in shallow marine waters of the Antarctic the biomass of living matter may be considerable, being often greater than in the Arctic. This richness of sea life results in part from the upwelling of nutrient-rich waters in the Antarctic Ocean.

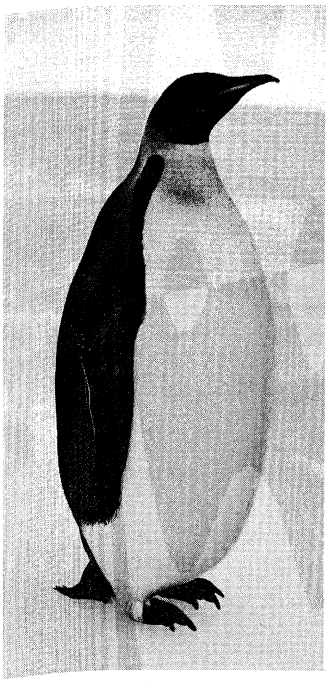
The subantarctic islands south of the Antarctic Convergence show polar affinities and are commonly included as part of the Antarctic in the broad sense, though on the islands of Kerguelen, Heard, and South Georgia, there occur coastal tussocky grasslands of meadow grass or fescue and inland fields of herbaceous plants, with mainly lichens and mosses on higher land. Especially notable is the Kerguelen cabbage (*Pringlea antiscorbutica*), the sole representative of its genus. Some ferns and a club moss occur in protected habitats. Animal life on these islands is more prolific than on the Antarctic continent; insects and birds are numerous.

BIOLOGICAL PRODUCTIVITY

In polar regions the greatest biological production occurs in marine waters rather than on land, with higher production occurring in the Antarctic than in the Arctic Ocean.

Production studies of Arctic tundra lakes indicate that there are many species of algae and even aquatic mosses living in the higher latitudes of the Arctic. These support a limited number of small crustaceans, worms, and insect larvae, which in turn support the Arctic char (*Salvelinus alpinus*), a fish related to salmon.

Plant production in the tundra appears to range from 3 to 10 grams of vegetable matter per square metre of area in willow–dryas barrens of the higher latitudes of the Arctic to values of 100 to 250 g/m^2 in lower-latitude



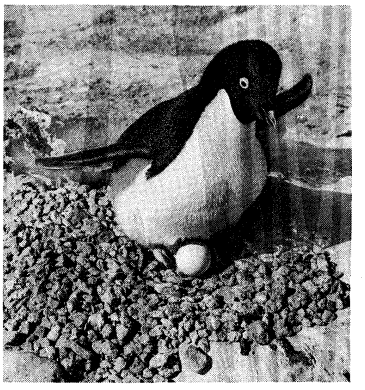


Figure 25: *Animal life in Antarctica.* (Top left) Emperor penguin (*Aptenodytes forsteri*), the only Antarctic bird that breeds in winter. (Top right) Adélie penguin (*Pygoscelis adeliae*) and (bottom) Weddell seal (*Leptonychotes weddelli*), two of the few types of animals native to Antarctica.

By courtesy of (top right and bottom) the U.S. Navy; photograph (top left), Popperfoto

sedge-dominated wet sites. The widespread cotton-grass–dwarf-shrub heath communities produce about 50 to 75 g/m^2 in new shoots per year. Values for temperate region grasslands and forests are about four to six times as great as the maximum rates for the Arctic tundra.

The accumulating data for Alpine tundras show 50 to 100 g/m^2 in windswept habitats, 100 to 200 g/m^2 in meadows, and values of 250 to 300 g/m^2 in dwarf-shrub heath communities. These values are higher than those in the Arctic because the growing season is often longer by several weeks.

These data do not tell the entire story, for the amount of plant material (standing crop) is often 10 to 25 times greater below ground than above ground, which indicates that the soil environment may be more favourable for growth than is the air.

Herbivore consumption

In many tundras, the harvest by plant-eating animals amounts to no more than 0.1 to 2.0 percent of the live plants per year. This means that the great majority of the plants produced fall to decay and are decomposed by microorganisms.

With Alpine and Arctic vegetation so sparse, it is no accident that humans have harvested herbivorous animals, including caribou, reindeer, ducks, and geese, rather than harvesting native or cultivated plants as they do in other major vegetation complexes (forests, grasslands, and deserts). (N.Po./L.C.B./Ed.)

Boreal and temperate forests

Forests are complex ecosystems dominated by trees, which form a buffer for the Earth against the full impact of the sun, wind, and precipitation. Whatever the type of forest—whether evergreen or deciduous, boreal or temperate—the trees that constitute it provide special environments, which in turn affect the kinds of plants and animals that can live within the forest.

THE ENVIRONMENTAL SETTING

Climate. Forests may develop wherever there is an average temperature greater than 10° C (50° F) in the warmest months and an annual rainfall in excess of 200 millimetres (about 8 inches). Above these limits there exists an infinite variety of tree species grouped into a number of stable forest types that are determined by the particular conditions of the environment. Although boundaries between these types are difficult to determine precisely, the major forest communities can be linked to broad latitudinal zones (see Figure 26).

The forests of the high latitude subpolar regions constitute the taiga and are dominated by stands of such conifers as pines, spruces, and larches. Taiga forests are found only in the Northern Hemisphere, where winters are long, with six months of the year showing an average maximum temperature of less than 0° C (32° F), there is a growing season of one to three months, and precipitation varies between 250 millimetres (about 10 inches) and 500 millimetres (20 inches), evenly distributed throughout the year. Broad-leaved trees predominate in the middle latitudes, where there are six months with an average maximum temperature above 10° C (50° F), precipitation in excess of 400 millimetres (about 16 inches), and a growing season of between 100 and 200 days. Broad-leaved deciduous forests occupy the cooler regions and are dominated by associations of several species including oaks, beeches, birches, aspens, elms, and maples.

Mixed forests form a transition between the coniferous taiga and the deciduous broad-leaved forest. Farther south, as frost becomes infrequent, the composition again changes, and a temperate rain forest of evergreen broad-leaved species develops. In the Southern Hemisphere the forest type contains conifers similar to the true pines: Kauri pine, Chile pine, and white pine. The temperate rain forest, also known as temperate evergreen forest and laurel forest, has fewer tree species than other types of rain forests and is lower and less dense.

Determination of forest type

Precipitation and temperature play important roles in the determination of forest type. The combined effect controls the amount of soil water available for tree growth. Water is lost from the forest by evaporation from the tree crowns and by transpiration from the leaves, both processes in large measure controlled by air temperature.

The principal influence of sunlight is on the air temperature. At high latitudes the sun's rays strike the Earth at a larger angle and therefore give less heat than at low latitudes. Another influence of solar radiation is through the day length (photoperiodism), which varies according to the season and affects flowering, bud opening, and a range of other physiological activities.

Topography and soil. Topography is also an important influence on forest vegetation; an increase of 300 metres (984 feet) in altitude is equivalent to a movement of 480 kilometres (300 miles) closer to the pole at sea level, so that a broad correlation can be drawn between montane forests and those at higher latitudes. Exact duplication of forest type is not expected, however, since the montane environment differs in important respects from that at an equivalent lowland station; notably, it has a shorter day length, greater precipitation, and higher maximum temperatures. The altitude at which the lowland forest passes into an Alpine type depends chiefly on the length of the growing season and the frequency of frosts. The zonation of forest types with increasing altitude is seen in the subalpine forest formations of the middle latitudes and the montane forests of the tropics. In the middle latitudes the subalpine forest is dominated by the more tolerant conifers (pines and larches), with poorly formed deciduous broad-

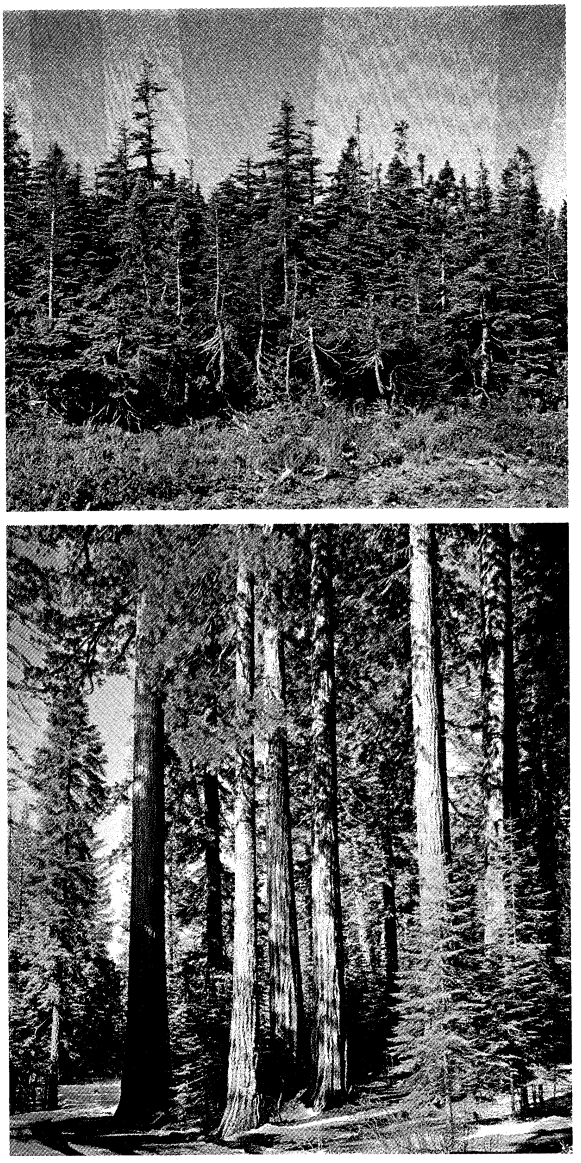


Figure 26: (Top) Northern coniferous forest in Nova Scotia, of red spruce (*Picea rubens*). (Bottom) Moist temperate coniferous forest in California, of redwood (*Sequoia sempervirens*).
(Top) Ken Brate—Photo Researchers, (bottom) A.C. Shelton—Publix

leaved species (beeches, birches, and willows) occupying the tree line, the upper limit.

Another important physical influence on forest type is the soil. The most important feature is soil depth, because this determines not only the capacity of the tree roots to reach for water and nutrients but also the stability of the trees. Forest soil shows considerable variability, but common features are depth, relative infertility, an undisturbed condition, a large quantity of woody perennial roots and characteristic plant and animal life. Quick-draining, highly stratified, sandy soils (podzols) are characteristic of the taiga. Brown forest soils, which are richer in nutrients, less porous, and less stratified than podzols, are characteristic of deciduous broad-leaved forests. Variation in soil type may be considerable even in small areas, *e.g.,* in regions subject to past glaciation; this variation is often reflected in species preference. In the mixed broad-leaved–conifer forests of the middle latitudes, conifers are usually found in poorer sandy soils of low alkalinity, whereas broad-leaved trees prefer richer clayey soils.

Wind and fire. Wind plays a primary role in the creation of the forest, by assisting in pollination and seed dispersal, but has only a secondary role in its maintenance. Air movement affects transpiration and may be a significant ecological factor in determining some forest types. Violent winds destroy forests by either uprooting or breaking stems; prevailing winds affect tree growth by producing eccentric stems and deformed crowns. Prominent among the physical destructive agents in the forest is fire, which usually results from human encroachment. Infrequent natural fires arising from lightning may be responsible for the development of stable fire forest types, such as Douglas fir (*Pseudotsuga menziesii*) stands in North America.

Prevailing winds

BIOLOGICAL FACTORS

Forests are distinguished from each other primarily by their composition. The number and variety of species present in the forest community depends on the age and density of the tree cover, the type of climate and soil, and the geologic history of the region. The richest forests in terms of total species composition are those having older and more open stands of trees, situated in favourable climates, and on fertile soils that have been relatively undisturbed in recent geologic time. Thus, within a given forest community there exists a complex interaction between the vegetation and the environment, which leads to the development of microenvironments with many physical properties.

Plant stratification and its influence on the forest environment. The forest community shows an extensive vertical stratification, from the tops of the tallest trees to the tips of the deepest roots. The effect on the stratification above ground is to moderate climatic extremes progressively downward from the treetops to the soil surface.

The pattern of vertical stratification varies with forest type. The conifer-dominated forest has the simplest structure: the tree layer is continuous up to 30 metres (100 feet) high, the shrub layer is absent or spotty, the field layer of herbaceous plants is best developed in more southerly forests, and the ground layer of liverworts, mosses, and lichens is almost uniform, particularly in the northerly forests. In the deciduous broad-leaved forests stratification becomes more complex. The continuous canopy of trees is between 10 and 30 metres (33 and 100 feet) tall but is subdivided into distinct upper and lower stories, in addition to the shrub field and ground layers.

Considerable lateral variability exists within each layer, or stratum, resulting from local differences in environmental conditions. Forests thus provide a great range of habitats for both other plants and animals. The most important physical factors influenced by forest structure are air temperature, precipitation, and soil characteristics; of secondary importance are sunlight and wind.

The forest environment exercises a moderating influence on air temperature, which is especially marked at the extremes of the temperature range. The forest is cooler in summer and warmer in winter, when compared with surrounding areas. The effect is most marked where seasonal temperature variation is greatest; *i.e.,* the farther removed from the equator, the greater the effect.

Vertical stratification and air temperature

The ameliorating effect of the forest on air temperature results from the vertical stratification of the forest vegetation and is less marked as the canopy cover decreases; thus, winter temperatures are little influenced in the leafless deciduous forest, although in the summer the same forest, in full leaf, shows not only differences in temperature compared with outside conditions but also a temperature gradient within the various strata of the forest. Daily fluctuations of temperature within the forest also depend on the nature of the vertical stratification. In the single-storied temperate conifer forest, temperatures are highest nearest the tree crowns. At night, heat is radiated from the crowns and the surrounding air is cooled, increases in density, and sinks to the ground. Subcanopy temperatures are therefore lower than canopy temperatures. During the day the temperature gradient is broadened; the crowns of the trees are warmed by the sun but at the same time shade the subcanopy region. As the canopy structure becomes more complex, more than one form of daily fluctuation in temperature can exist. Forest vegetation also has a marked effect on soil temperature. Forest soils are cooler in the day and warmer at night and show less seasonal fluctuation than similar soils outside the forest.

As precipitation falls on the forest, its movement is significantly influenced by the forest cover. The dense tree layer initially intercepts precipitations, some of which is immediately evaporated back to the atmosphere. Conifers tend to hold more water than broad-leaved trees (there is, of course, little or no interception from dormant leafless trees). Water penetrates to the soil from the canopy either through stem flow down the main trunks and thence into the soil along the root channels or by direct absorption in the litter of the forest floor. In either case the velocity of the fall is reduced and there is little soil compaction or surface runoff and consequently a reduced risk of soil erosion. Compared with stations outside the forest, the moisture content of forest soil, particularly at the surface, tends to be high. Water is lost through surface evaporation, which is low in the forest because of the ameliorating effect of the tree cover, and by transpiration, which is high and may be a limiting factor in tree growth. The water budget for the forest will vary significantly with type and situation, but the ability of the forest cover to reduce erosion and to even out the yearly fluctuations of stream flow are recognized advantages to water catchment areas. Although there is no evidence that forest conditions affect the amount of precipitation, the forest atmosphere is always more humid than surrounding areas.

Types of organic matter

The forest floor consists of a layer of organic matter overlying mineral soil. In the temperate regions two major types of organic matter predominate, the mull and the mor. The mull is associated with broad-leaved deciduous forests, especially those growing on relatively rich soils. It is less acid than the mor type, of a granular structure, and contains a considerable admixture of mineral soil. There is an indistinct boundary with the upper horizons, owing to the activity of earthworms moving freely between the decomposed layer of litter and the mineral soil. Mor humus, on the other hand, is strongly acidic and of a matted structure (attributed to the network of fine roots). There is a distinct boundary between mor humus and mineral soil, with little evidence of the activity of mixing agents. The earthworm population is smaller, although fungal activity may be greater, than in mull. Mor is characteristic of the northern coniferous zone and is especially evident under pines growing on mineral-poor soil, although some broad-leaved species, especially oaks (*Quercus*) and beeches (*Fagus*) produce a mor humus when growing on similar soils. The differences between mull and mor are generally attributed to differences in the chemical composition of the plant material from which they are derived and the nature of the mineral soil, rather than to the rate of decomposition. Occasionally in swampy conditions where acidic groundwater percolates laterally, and in conditions of heavy, regular rainfall, a thick peat, made up almost entirely of woody plant debris, may develop.

The forest structure produces less spectacular, though equally important changes in wind speed and solar radiation. Dense multistoried forests are most effective in reducing overall wind velocity in the subcanopy layers. The filtering of the tree strata also affects solar radiation. A conifer forest typically allows no more than 10 percent of visible light to reach the forest floor, and a deciduous forest about 5 percent when in leaf but between 50 and 90 percent when leafless. Forests in full canopy approach the maximum possible efficiency in their utilization of solar energy: 60–90 percent is absorbed, and the remainder is reradiated and used up by evaporation and transpiration. In qualitative terms the light values within the forest vary since the canopy vegetation absorbs more red than blue light. Blue light is, however, utilized at subcanopy and low-canopy levels and is said to reduce the tendency toward side branching, a possible explanation for the observation that understory trees are generally pyramidal but adopt a rounded crown when they achieve upper-story height. In their youth some trees, such as oak and pine (*Pinus*), require much sunlight, whereas others, spruce (*Picea*), firs (*Abies*), and beech, almost demand shade.

As indicated earlier, stratification within the forest supplies many habitats for both plants and animals and provides the opportunity for considerable interaction between organisms.

The role of decomposer organisms. The fungi of the litter and mineral soil layers play an important role in the chain of litter decomposition and the release of nutrients in the soil. Some species live in a mutually beneficial association (called mycorrhiza) with active tree roots, thus improving the efficiency of nutrient uptake by the trees. On the other hand, some species are parasitic and may cause devastation; *e.g.,* the chestnut blight, caused by the fungus *Endothia parasitica,* which has decimated the population of native chestnut (*Castanea dentata*) in North America, and the blister rust fungus, which is a serious pest of white pine (*Pinus strobus*). The fungus *Fomes annosus* attacks a number of conifer species; it is transmitted from tree to tree not only by spore infection via cut wood surfaces but also by root contact. Some higher plants are also parasitic; *e.g.,* the European mistletoe (*Viscum album*) and its North American counterpart cause stem and crown deformation. Among the animals, the insects are the most mobile in the forest. They are also the most numerous and the most highly adapted as to feeding habits. Insects are an important link in the food chain between vegetation and higher animals, particularly birds. Many tree species have flowers adapted to insect pollination; the range is wide from simple nectar-rich flowers to highly structured pollination mechanisms. Insects are often highly specialized in their feeding habitats; of 20 species known to occur on white pine, five are leaf feeders, three are bud feeders, three are twig borers, two are wood borers, and two are root borers. Species preference is also noticeable; of 43 species of British weevils whose larvae feed on Scotch pine, oaks, and birches, all except one attacks only one species of tree. Insect damage to forests reaches disaster proportions only when populations break through the natural control barriers; *e.g.,* the periodic defoliation of balsam fir (*Abies balsamea*) by the spruce budworm (*Choristoneura fumiferana*) is triggered by several summers of clear, dry weather, which favours a rapid increase in the insect population. Insects also are known carriers of the viral, bacterial, and fungal diseases of trees. As a proportion of the total insect population, however, the serious forest pests are small in number and localized in their influence.

Mycorrhiza

Forest animals and their adaptations. Larger forest animals are likewise especially adapted to their environment. Sharpness of sight and smell are apparently not as useful as a sense of hearing in the forest habitat, and most gregarious animals, including birds, are noisy in contrast to the animals of open country. The value of flying and running is also reduced, and true forest birds are poor flyers but good climbers. Movement on the ground is restricted, but deer and antelope, slender and lithe, can thread their way through the forest. Specialist adaptations include those that enable the animal to move through the dense canopy layers, sharp claws (as on woodpeckers, squirrels, martens), gripping feet or prehensile tails (as on porcupines, anteaters, marsupials), and parachute-like skin extensions (as on flying lizards and flying squirrels). Some animals, such as birds of prey, wolves, foxes, and stags, use the forest only for shelter. Although the forest is a sheltered and ameliorated environment, food is scarce and populations are relatively low.

Each forest type has its distinctive animal life. In the deciduous broad-leaved forests snails and slugs are more numerous than in the conifer forest, a reflection of the difference in the quality and quantity of the herb layer. In temperate zones forest birds distribute seeds and insects pollinate.

Animal damage rarely reaches catastrophic proportions in the forest, but local damage from grazing and browsing can result in bark stripping of young trees.

SEASONAL CHANGE

Forest rhythms

The forest community changes with the seasons. Superimposed on these changes are the daily rhythms of daylight and darkness; of these effects, the length of daylight, or photoperiod, is an important component of seasonal change, or aspection. The stimulus of these changes is environmental and as a consequence is most marked in regions where temperature, precipitation, and day length are seasonally variable. The responses of organisms to

periodic changes in the environment have evolved over a long period of time to parallel the external environmental conditions.

Aspection is also influenced by latitude, altitude, and rainfall distribution. At high latitudes the winters are prolonged, snow and frost occurring for as long as six to eight months continuously; the spring and autumn periods are short, and the growing season may be as short as three months. (Seasonal changes in the Southern Hemisphere are the reverse of those in the Northern Hemisphere.) Longer, cooler periods are associated with increasing altitude and are reflected in the forest composition. The seasonal changes in forest communities at high elevation are similar to those closely related communities at lower elevation but higher latitudes. Finally, aspection is affected by moisture availability; thus in Mediterranean climates, the resting period for many plants, presaged by leaf fall in deciduous species, results from a seasonal lack of moisture rather than from decreasing air temperature and decreasing day length.

In the middle latitudes of the temperate zone, characterized by the deciduous broad-leaved forest belt, six phases of the seasonal cycle are recognized. In this region aspection is most strongly evident, and the phases correlate climatic conditions with the outward appearance of the tree cover. In the early spring, or prevernal period, the hardy species and the lower-storied trees bud and bloom as air temperatures and day length increase. In the upper stories, bud break is delayed, but the sap rises from the roots to the crown. The late spring, or vernal period, sees soil temperatures rise with air temperatures and day length reach its maximum value. As the leaves expand, shoot growth accelerates; the transpiration rate becomes more rapid as the leaves mature and as moisture in the soil becomes more readily available. At this period the vegetation is most vulnerable to frost damage. The next two periods, estival (aestival) and serotinal periods, cover early summer and midsummer, when conditions for growth are optimal. Air and soil temperatures reach their maximum, the full-storied structure of the forest is operative, and the metabolic rate is at its highest. During this period, moisture availability becomes the critical, or limiting, factor to growth; high temperatures combined with extremely dry conditions may induce a lowering of the metabolism of the tree, characterized by premature aging of leaves and a slowing down of plant activity. The autumnal period is characterized by falling air temperatures, decreasing day length, and the advent of frosts. These changes induce a gradual reduction of metabolic activity and defoliation of deciduous vegetation. Finally, the winter, or hiemal period, is characterized by low daily maximum temperatures, heavy frosts, precipitation in the form of snow, and short day lengths. During this season the trees enter a dormant stage in which metabolic activity is diminished to the minimum level necessary for survival.

Hiemal period

Seasonal change is least marked in the higher latitudes of the temperate zone, characterized by the coniferous belt of forests. Here the species mixture is limited and simple in physiognomy, and although the conifers are evergreen, water loss through transpiration is considerably reduced as water becomes less available from the freezing soil. The evergreen habit does enable the maximum use to be made of the short growing season, and conifers are particularly well adapted to rigorous environments. The understory plants, mosses, ferns, lichens, and grasses, show no spectacular seasonal changes, and the trees themselves change little in appearance with the passage of the seasons, except for the deciduous species such as the larches (*Larix*). Even in the species-rich coniferous zones of western North America and in Asia, aspection is not well marked in the tree cover, although breaks in the canopy reveal herb and shrub layers that show the same periodic changes that are seen typically in the broad-leaved forest.

One common feature of the extratropical forest regions is the periodic activity of the tree growth. The growing tips (meristems) in the buds, which are responsible for height growth and crown development, are most active in the spring. The lateral, or cambial, meristem, which is responsible for increase in the diameter of the tree trunk, also is most active in the spring, slowing down in the autumn and ceasing activity in the winter. All temperate trees show an annual ring structure, roughly indicative of age, resulting from the episodic nature of this cambial activity.

COMMUNITY DEVELOPMENT

Forest communities represent the terminal aggregation of species in a series of associations that may have begun from primary succession—*i.e.*, colonization of bare mineral oil, exposed rock surfaces, volcanic ash, or sand dunes—or from secondary succession, in which the substrate had first been occupied by a community that, because of a catastrophic event (*e.g.*, wind devastation, fire, or human intervention), has since been destroyed. Although successional changes may vary considerably between these two major avenues of development, the time span in each case is long, and the biological interactions between the various plant and animal components are complex. Each recognizable stage in the development of a stable, or climax, community is known as a sere. Climax forest communities, because they usually represent the terminal point of seral development, are grouped together in more or less climatically controlled geographic groupings identified as forest biomes (see Figure 27).

Coniferous forests. *Taiga.* The circumpolar high-latitude forests of the Northern Hemisphere are dominated by pines, spruces, firs, and larches. The Eurasian and North American formations are similar in structure, and in both the species composition changes from east to west of the formation. Important eastern formations in Europe are the native stands of Siberian pine, Siberian fir, and Siberian and Dahurian larches; in North America, white spruce and balsam fir are the eastern dominants. In western Europe, Scotch pine and Norway spruce absolutely predominate; in western North America, lodgepole pine and Alpine fir are among the important species in the area.

Exploitation and recovery of the taiga

The taiga forests, especially in the almost pure stands, have been exploited on a large scale from the mid-19th century on. On cleared areas, on catastrophic burns, or whenever adequate precautions to ensure conifer regeneration are not taken, secondary succession begins. Dominated initially by herbaceous cover, the succession soon contains deciduous broad-leaved trees, especially birches (*Betula*) and aspens (*Populus*), prolific seed bearers, whose seeds are wind-dispersed and of a high germination capacity. Willows (*Salix*) and alders (*Alnus*) are also early arrivals in the hardwood stage of succession. Scotch pine (*Pinus sylvestris*) in Europe and jack pine (*P. banksiana*) in North America produce the advanced succession stage. Final reversion to the climax forest is achieved after a considerable time interval through the ability of the conifer species to regenerate and grow under the shade of the canopy. In addition, conifers tend to grow taller than the broad-leaved species when in direct competition and eventually shade them out. Broad-leaved species occur in the taiga in locally induced biotic climaxes and are common on the latitudinal and altitudinal margins of the biome.

Moist-temperate conifer forest. Forests of this distinctive type occur along the western seaboard of North America from Alaska to California and inland to the Rocky Mountains. Here conditions are humid and a high proportion of the precipitation occurs as mist. Species domination shifts from Sitka spruce (*Picea sitchensis*) in Alaska through the western red cedar (*Thuja plicata*)–western hemlock (*Tsuga heterophylla*) association in British Columbia to Douglas fir and finally the coast redwood (*Sequoia sempervirens*). Because of the high humidity a rich understory development, especially of mosses and other moisture-loving plants, is a feature of this forest formation. The trees are massive, and they constitute the richest commercial forest in the world. As in the taiga, secondary succession features the pioneer broad-leaved species, especially the alders and, in the south, maples (*Acer*). Where the humid conditions are maintained only under continuous conifer cover, oaks infiltrate late in the succession from the neighbouring transition to the Californian chaparral. Primary succession from the coast tends to be dominated by pioneer pine species, especially lodgepole pine (*Pinus contorta*), before the climax is rapidly formed.

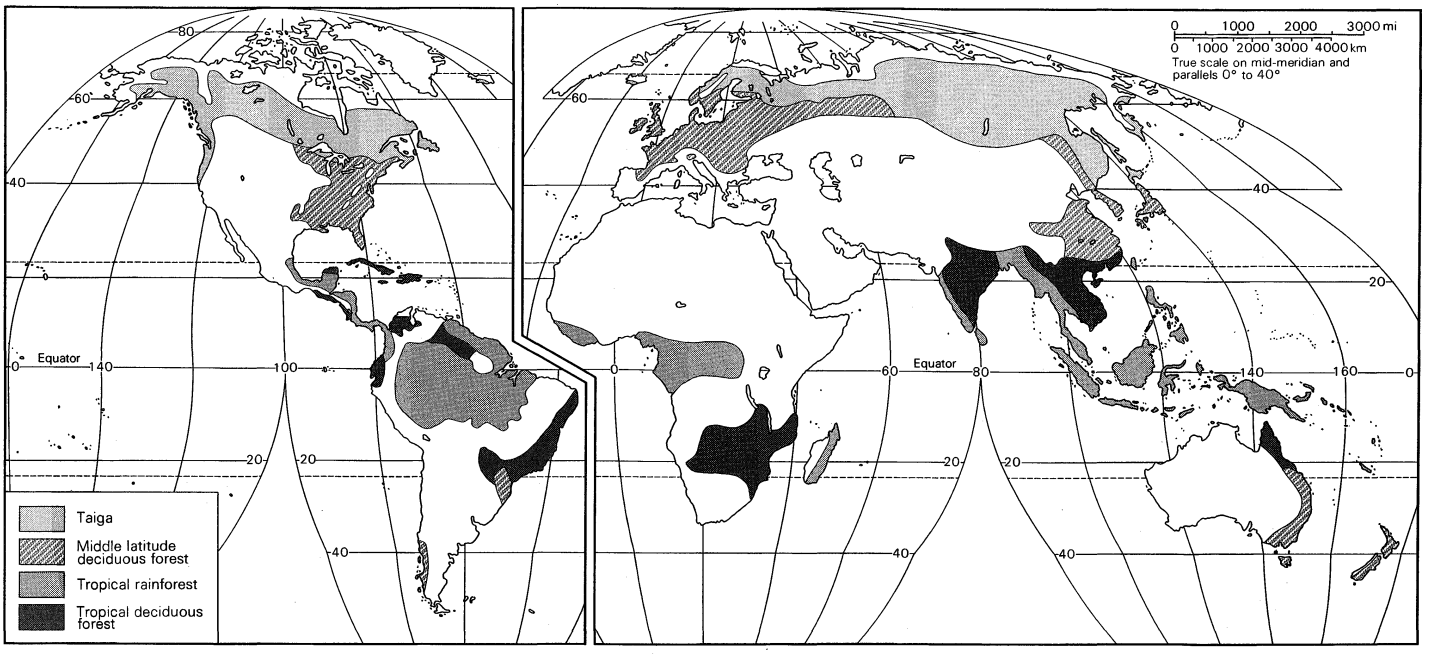


Figure 27: World distribution of major forest biomes.
Adapted from *Biological Sciences Curriculum Study Green Version High School Biology*, 2nd ed.; Chicago: Rand McNally & Co., 1968

Middle-latitude forests. *Deciduous broad-leaved forests.* This biome, which occurs in Europe and North America, is floristically the richest in the New World. Associations of beeches and maples occur in the north, maples and basswoods in the centre, oaks and hickories in the west and south, and oaks and chestnuts in the Appalachian Mountains. Native species of oaks, beeches, and birches form the associations in the northern European formation, limes, chestnuts, and elms appearing in the south. Community development in this region again arises from forest clearances; in North America, where population pressure is not so great as in Europe, abandoned agricultural land is a common starting point. The grassland community is first invaded by broad-leaved shrubby trees, and a mixed-species thicket stage develops. Later successional changes feature the invasion of jack pines for reasons analogous to those operating in the conifer biome. The richer soils and better growing conditions of the deciduous broad-leaved biome allow a more rapid growth of broad-leaved species, some of which are shade-tolerant. The pioneer pines soon become mixed with broad-leaved species, which in turn dominate the community as the climax forest develops. It is not uncommon for the succession to be deflected to a conifer climax by either regular burning or grazing. The pine barrens of southeastern North America, dominated by several pine species, are an example of such a climax.

Northern European forests

Evergreen broad-leaved forests. This formation occurs both north and south of the equator on the eastern seaboards of continental landmasses. The cooler temperate formation is found in the Southern Hemisphere; in New Zealand, southern so-called false beeches dominate. Although evergreen, the formation has much in common with the deciduous broad-leaved forest of the north. In the warm temperate forests of the Southern Hemisphere, the beeches are joined by the southern conifers: Kauri pine (*Agathis australis*) and yellowwood (*Podocarpus*) species in New Zealand, and *Araucaria* species in Chile. Particularly in the warmer regions secondary succession is complex and slow. Pure strands of southern beeches develop especially on the more exposed hillsides. The southern conifers are slow-growing and are susceptible to grazing damage; once removed they are unlikely to regain their place in the community. False beeches are very susceptible to grazing damage; recovery is difficult, and often the forest reverts to a thick grassy turf.

PRODUCTIVITY OF FOREST ECOSYSTEMS

Quantitative estimates for the productivity of forest communities are scarce and unreliable. Among the reasons for these conditions are the complexity of the forest community, the longevity of tree species, and the lack of historical records of routine weight data. Moreover, there are three levels of production that might be measured. Gross, or primary, productivity is the quantity of organic matter created by photosynthesis. From this value the energy used in respiration is subtracted to give the net production, which is measured as the dry matter, or dry weight of the total community. A third measure is the merchantable production, which is the dry weight of the usable tree stems only and is calculated by multiplying the stem volume by the basic density of the wood. Although volume figures are available for most of the species under regular forest management, this measure cannot be regarded as an estimate of total production; even for a single tree it represents less than 70 percent of the total dry weight, the unmerchantable stem (1–6 percent), stump (10–18 percent), green branches and foliage (10–20 percent) making up the remainder. For a complete evaluation of forest productivity, both the vegetation and the animal life need to be measured, a task that is overwhelming in magnitude. The difficulties of measuring even the dry weight of the total plant community present extreme sampling difficulties. Much of the data available are concerned with the plant biomass (net production), and often only for the tree component.

The three levels of production

Forests are the most efficient of all terrestrial communities in the production and storage of organic matter. A high proportion of the sunlight that falls on the forest is absorbed; the canopy provides a large area of chlorophyll-bearing tissue dispersed through a considerable volume of air for the absorption of carbon dioxide, and tree roots exploit a considerable volume of soil for water and nutrients. The three essential prerequisites for efficient photosynthesis are, therefore, effectively fulfilled, and as a consequence the forest system capitalizes on the production of organic matter. Additionally, the forest community operates efficiently through time, although forests of evergreen conifers are more efficient than deciduous broad-leaved forests in total as well as tree biomass production. Systematic studies of the changes in the rate of organic matter production through the life of the forest are lacking, but the data available from reforestation programs indicate that the tree component increases slowly as the canopy closes, maximizing at the point of greatest competition between trees, and thereafter falling as trees are removed or die. The contribution of the ground layer is most marked at the youngest and oldest ages of the crop—*i.e.*, when the canopy is open. Complementary data on

Table 10: Plant Biomass of Woodlands
(units are in dry weight 100 kg/ha)

	Pinus nigra	*Pinus sylvestris*	*Betula verrucosa*	*Quercus borealis*	*Picea abies*	*Nothofagus truncata*	*Pseudotsuga menziestli*	evergreen gallery
Location	northeast Scotland	eastern England	Moscow	Minnesota, U.S.	Sweden	New Zealand	Washington, U.S.	Thailand
Status	plantation	plantation	natural	natural	natural	natural	natural	natural
Age of tree in years	48	55	67	57	58	110	52	...
Tree height in metres	14	16	26	17	17	21	17	29
Number of trees/ha	1,112	760	...	800	924	490	1,157	16,209
Tree leaves	5.6	7.2	2.8	3.5	9.1	2.7	12.0	19.0
Tree branches	11.2	12.3	11.3	49.5	14.3	42.0	17.9	50.0
Tree trunks	95.1	96.7	156.7	111.9	85.3	224.8	174.8	225.2
Shrubs and herbs	7.0	2.6	2.0*	0.6	1.0*	0*	0.1	0.2
Roots	34.0	34.1	43.1	15.0	60.0*	39.2	12.3	88.5
Dead branches on trees	10.0	10.0	2.0*	21.9	2.6	1.1	11.2	...
Organic matter on ground	22.0	45.0	3.0	36.7	78.0	16.7	117.3	3.0
Total	184.9	207.9	220.9	238.1	250.2	326.5	345.6	385.9

*Estimates from other similar woodlands.
Source: J.D. Ovington, *Woodlands* (1965)

the dry weight of foliage confirms this pattern, and it is probable that the number of stems per unit area (stand density) is as important as geographic location for both the weight of foliage and the total biomass produced. Within a given geographic region there are considerable differences between tree species in their rate of organic matter accumulation. Various estimates of timber production indicate that conifer stands usually are more productive than deciduous tree stands. Despite the lack of comprehensive data, the biomass production from the forest community is impressive (Table 10). (G.K.E./Ed.)

Scrublands

Mediterranean vegetation

Scrublands are areas where low, evergreen, leathery-leaved, and often aromatic shrubs (or scrubs) dominate the vegetation. Mostly they are confined to a Mediterranean type of climate—hot, dry summers and cool, wet winters—and therefore are often called the Mediterranean vegetation. This type of climate is found most often near seacoasts, and it commonly occurs in a latitudinal zone between 30 and 40 degrees both north and south of the equator. Scrublands merge into monsoon forests and evergreen forests toward regions of wetter climate and into thorn-shrub vegetation and steppes toward regions of decreasing precipitation.

THE ENVIRONMENT AND BIOTA

During the dry summer months, the temperature differential between day and night in scrublands is much greater than during winter, and during the wet winter months the night temperatures are usually between 10° and 16° C (50° and 60° F), which is optimal for the growth of scrub vegetation. The night temperature is most important in the control of plant growth.

The high summer temperatures make the small amount of precipitation insufficient for growth of scrubland plants, and most of them become dormant. Scrub plants flower and produce fruit before the onset of the hot, dry summer, the scrublands bursting into colourful display during spring. In California, for example, sage, mountain lilac, tree poppy, and *Fremontia* bloom at this time. In Australia *Grevillea,* acacia bottlebrush, and *Dryandra* are the principal genera of spring flowering plants. In the Cape scrublands of South Africa various members of the protea family and heath colour the hillsides with spectacular displays of flowers. This same feature of the climate also conditions the dominance in the scrubland vegetation of winter annuals—*i.e.,* plants that complete their life cycle in a single growing season, in this case during the wet winter months. The summer drought makes it impossible for shallow-rooted herbs to survive, leaving most of the space between shrubs bare. Thus, at the beginning of the autumn rains, the seeds of annual plants are able to germinate on the bare ground.

The shrubs of the scrublands are adapted to growing in winters that combine low temperatures with adequate moisture. Typical desert plants that grow only at high temperatures, such as the creosote bush, or jarillo (*Larrea divaricata*), cannot live in the scrublands nor can shrubs native to areas with summer rains and winter drought (*e.g.,* in western Texas or the Transvaal).

The dry climate of the hot summer months predisposes scrublands to fires, and most of them can be described as being a fire-maintained vegetation. The shrubs are adapted to fires in being able to sprout from their root crown. Since most trees are unable to form new shoots from the base of their trunk, a scrubland vegetation is kept free from invading trees through frequent fires.

Many of the scrubland plants have seeds whose germination is stimulated by high temperature treatment. This is known, for example, for the chamiso (*Adenostoma fasciculatum*), whose seeds do not germinate until very old unless they are subjected to dry heat near 93° C (200° F) or else are treated by being placed near charcoal or by being leached for a long time in water.

Scrublands are not usually browsed by animals; although deer may live nearby, they normally do not eat the frequently aromatic or toxic shrubs. Another reason why the scrublands in general are not grazed or browsed by the larger mammals is their impassable, dense growth. Although not as forbidding as the thorn-shrub vegetation, scrublands are accessible to larger animals only with difficulty. In Australia, the wallabies (*Wallabia*) and even the so-called scrub wallabies (*Thylogale*) use the scrub vegetation only for resting during daytime, feeding in open grasslands during night. Scrublands also are impassable to the Australian flightless birds.

Scrubland animal life

While not inhabited by larger mammals, scrublands are frequented by small seed eaters. Typical of these seed eaters are small rodents, such as pocket mice (*Perognathus*) and kangaroo rats (*Dipodomys*), and quail. None of these animals is restricted to scrublands, however, and most prefer a more open vegetation.

Insects and insect-eating birds and reptiles are common in scrublands but not more so than in other types of vegetation. Because it is impossible to characterize scrublands by any typical animals, botanical criteria are relied upon exclusively.

SCRUBLAND DISTRIBUTION

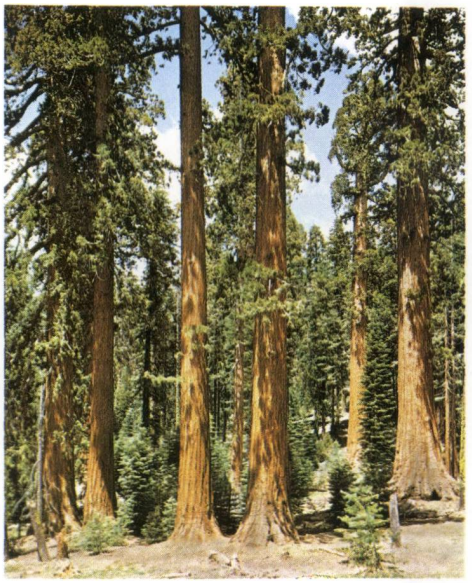
The accompanying map (Figure 28) shows the distribution of the most important scrublands over the world. A description of each of them follows.

The Californian chaparral. The scrubland vegetation of the Californian chaparral occurs on the lower slopes (generally below 1,000 metres, or 3,280 feet) of the coastal and inland mountains of southwestern North America, wherever the yearly rainfall ranges from about 250 to 500 millimetres (10 to 20 inches) per year, with a pronounced dry summer. At higher altitudes with increasing precipitation, the vegetation gradually changes into fir and pine forests, and in the interior the scrubland flora is replaced by a steppe vegetation dominated by the sagebrush (*Artemisia tridentata*).

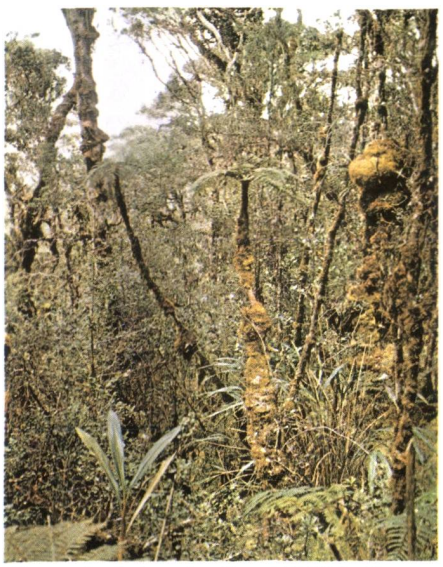
Along the coast and in the higher rainfall areas, the dominant shrubs of the chaparral are sages, mountain lilacs, rhuses, desert buckwheat, and the scrub oak. In

Terrestrial environments

Vegetation zonation (trees to stunted trees to alpine grasses) in the Colorado Rockies.

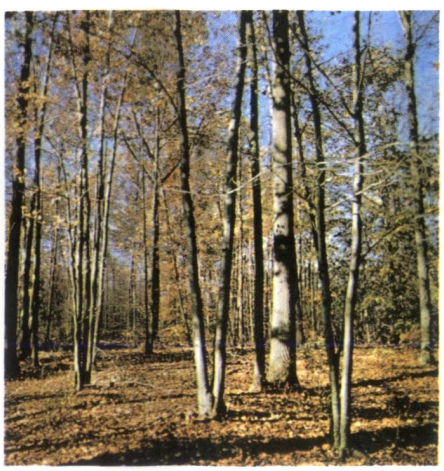


Mariposa grove of sequoias, Yosemite National Park, California.



Elfin woodland at an elevation of 10,000 feet in Celebes.

Boreal forest, or taiga, of Labrador.

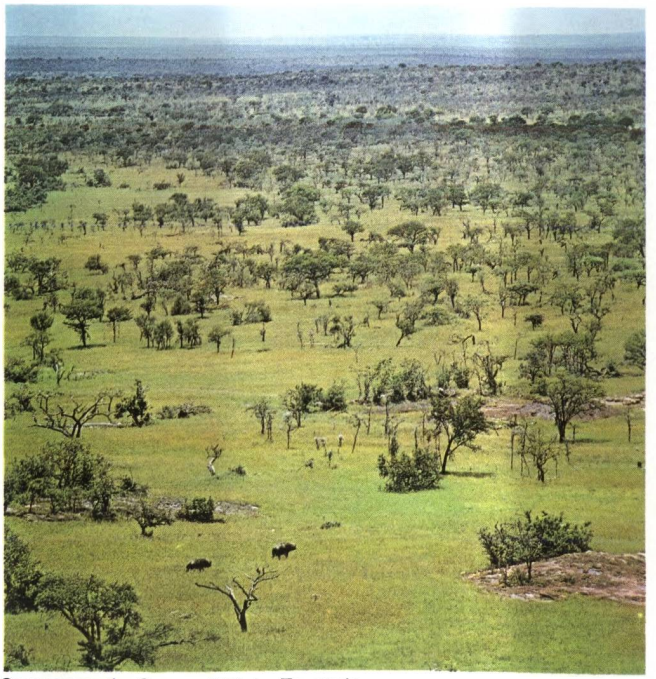


Deciduous hardwood forest of northeastern United States.

Alaskan tundra in the summer.

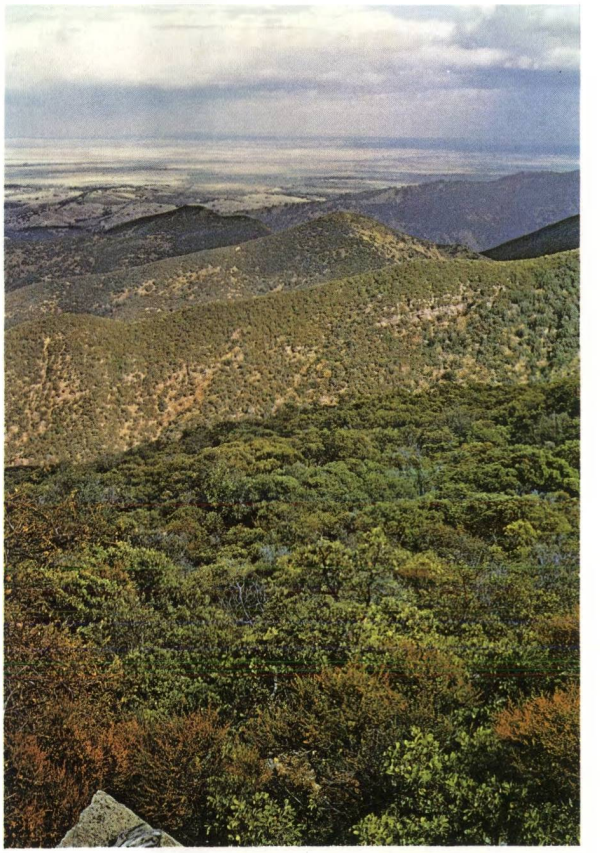
Plate 7: (Top left) Grant Heilman—EB Inc., (top right) Gene Ahrens—Bruce Coleman Inc., (centre left) A.J. Jermy, (centre right) Tom Willock—Ardea Photographics, (bottom left) Walter Chandoha, (bottom right) Willis Peterson

Terrestrial environments



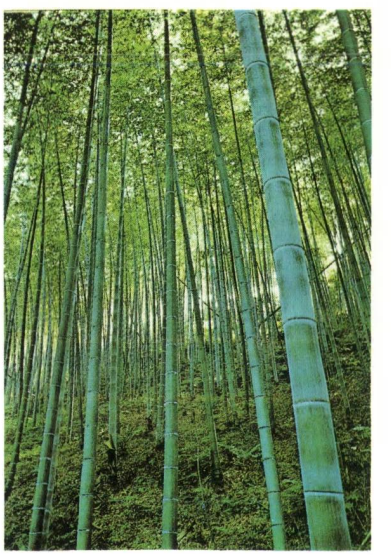
Savanna on the Serengeti Plain, Tanzania.

Chaparral in California showing (top) chaparral fire, necessary factor in the maintaining of the chaparral ecosystem and (bottom) denseness and extent of scrubland.



Tropical rainforest on western Madagascar.

Sub-alpine grassland, Yellowstone National Park.



Bamboo forest at Sagano, Kyōto, Japan.

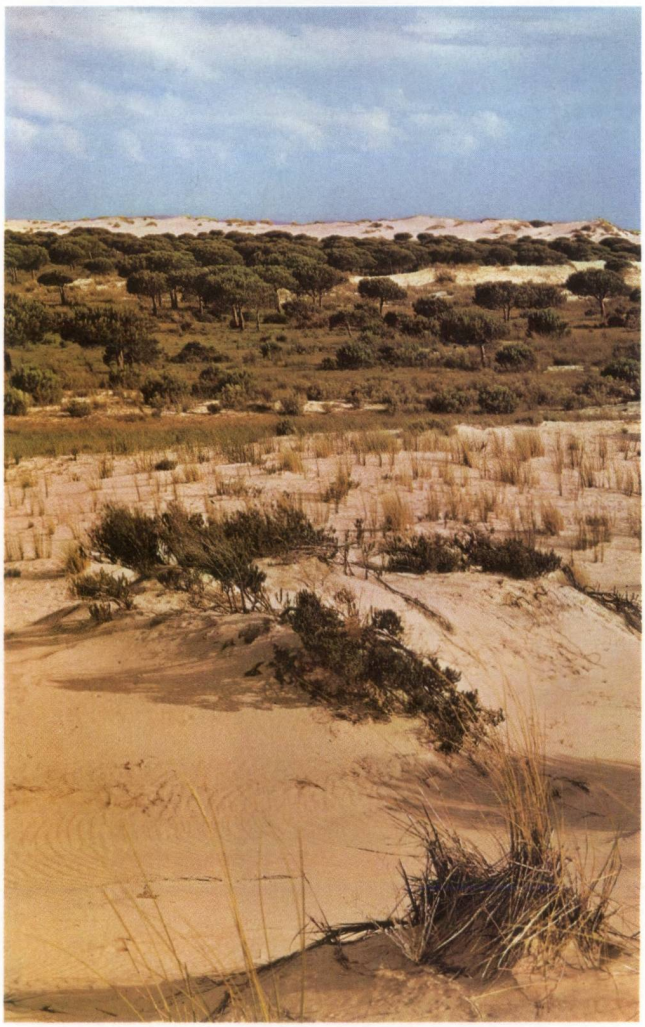
Plate 8: (Top left) Ingmar Holmasen, (top right, centre right) J.W. Wilburn, (centre left) Peter Johnson from The Natural History Photographic Agency—EB Inc., (bottom left) Charles Hannum—EB Inc., (bottom right) Maeda from Shostal—EB Inc.

Warm desert plant life on the Sonoran Desert of southern California.

Sphagnum bog in northern Wisconsin.

Cold desert in the Pine Nut Mountains of Nevada.

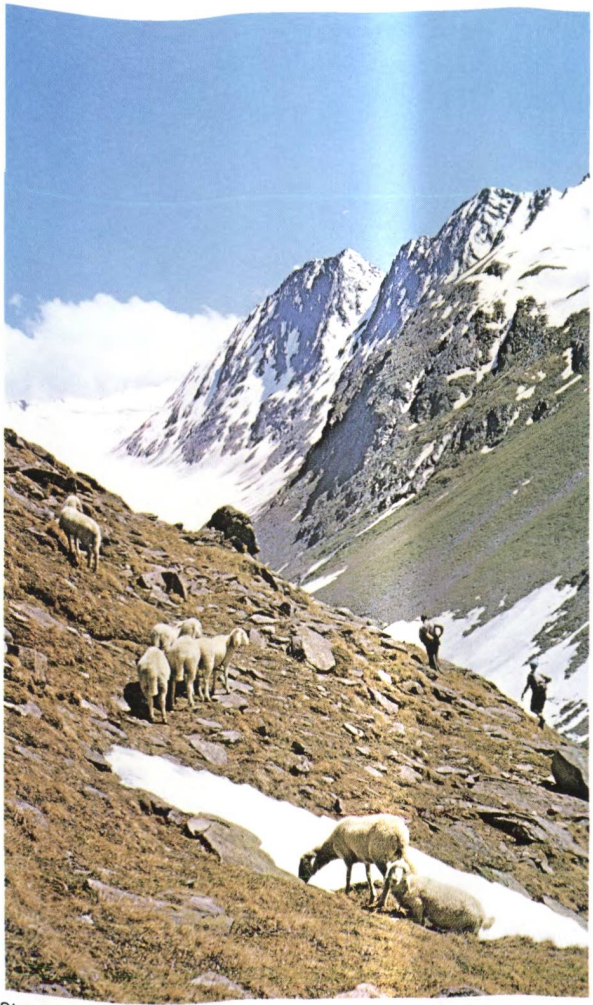
Okefenokee Swamp of Georgia.



Sand dunes on the Atlantic coast of southern Spain.

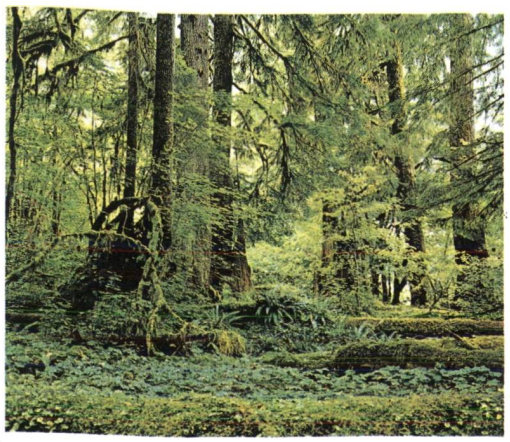
Plate 9: (Top left) Dennis Brokaw, (top right) Harold Hungerford—Photo Researchers, (centre left) J.W. Wilburn, (bottom left) C. Hardin from Shostal—EB Inc., (bottom right) Alfonso Gutierrez Escera—Ostman Agency

Terrestrial environments

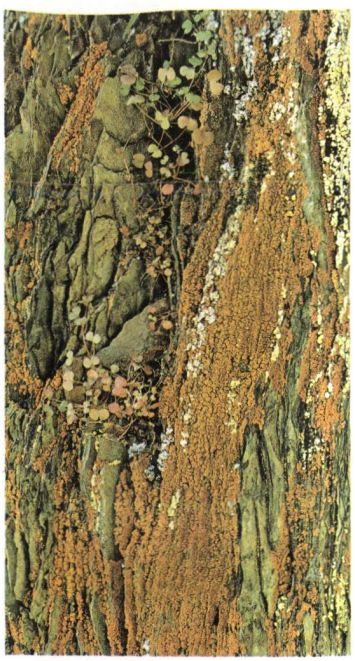


Shepherds herding their sheep on the alpine tundra of the Austrian Alps.

Oasis on the Sahara of Morocco.



Mosses, ferns, fungi, and other plant life that form the extremely dense undergrowth in the Olympic National Forest of Washington.



Lichen community thriving on a rocky cliff.

Spring wildflowers growing in the desert in Namaqualand, South Africa.

Plate 10: (Top left) A.J. Huxley—EB Inc., (top right, centre right, bottom left) Shostal—EB Inc., (top right) J. Turver, (centre right) B. Nelson, (bottom left) W. Wiley, (bottom right) Gerald Cubitt

drier localities and on poorer soil, chamiso or ribbonwood becomes dominant. If left undisturbed, live oaks and big-cone spruce invade the chaparral on the better sites of north slopes, gradually transforming them into forests. Only in exceptional localities does the chaparral remain undisturbed; normally it burns over once every 10 to 40 years. After a fire the shrubs resprout from the root crown, and many germinate from seed, together with many herbaceous plants.

The Mediterranean maquis. Although human activities have removed much of the Mediterranean maquis, in Spain, southern France, Greece, Lebanon, coastal Israel, Tunisia, and Algeria a shrub vegetation still covers the lower slopes of the mountains bordering the Mediterranean Sea. This vegetation consists of many aromatic shrubs such as savory and other members of the mint family, rockrose, laurel, myrtle, and different brooms, with olives, carob, figs, and other semitrees transforming the scrublands into forests in the absence of human interference.

Where the soils are poorer and rockier, the maquis is more open, sometimes with shrub palms (*Chamaerops*), and is called garrigue in the French Mediterranean.

Many of the plants of these Mediterranean scrublands are well known, such as the laurel and the olive. The myrtle, and innumerable spices, such as rosemary, thyme, and marjoram, are also familiar members of this vegetation.

Boundaries of maquis

Northward, toward central Europe, summer rainfall increases, and a forest climate replaces the Mediterranean climate, whereas south and eastward the climate becomes drier and changes into a steppe and desert climate. This occurs along the Mediterranean shores in Libya and Egypt, where a semidesert vegetation replaces the maquis because of an absence of moisture.

A vegetation similar to the Mediterranean scrublands extends from Turkey along the lower mountain ranges of Asia Minor. This is a reflection of the climate, with the rainfall decreasing eastward but increasing at higher altitudes. It is at the borders of the Mediterranean scrubland climate where the first successful attempts at agriculture occurred in Asia Minor, about 10,000 years ago.

With the unprecedented increase in population since Roman times, most of the original maquis vegetation has been transformed by humans into cultivated areas. River valleys are irrigated, and profitable crops now replace the maquis. On mountain slopes, olive orchards now grow in areas formerly occupied by a varied maquis vegetation.

The Australian scrublands. Australia has probably the most extensive and richest scrubland vegetation. In Western Australia rainfall ranges from about 250 millimetres (10 inches) in the interior to about 1,500 millimetres (60 inches) along the southwest coast. There are extensive eucalyptus forests in the high rainfall area, but in areas having between about 380 millimetres (15 inches) and about 1,000 millimetres (40 inches) of precipitation, varied scrub vegetation covers the land except near the roads, where extensive clearings have removed the shrubs. There are about 1,000 shrub species living in southwestern Australia, many of them with spiny branches and leaves and with brightly coloured flowers ranging from deep blue, lilac, yellow, and brown to red. Most of these shrubs belong to a few families, such as the pea family (especially the wattles [acacia]), the myrtle family, the protea family, and the rue family.

Eastward, past the desert of the Nullarbor Plain, another scrubland is found, commonly named mallee, where shrubby eucalyptus, wattles, *Banksia,* and other members of the protea family reign supreme in South Australia.

The Chilean scrublands. Also in the Southern Hemisphere are the scrublands of central Chile, where *Acacia cavenia,* groundsel bush, and tall cacti grow on the Pacific slopes of the Andes. Lower on the eastern slope of the Andes in Argentina grow desert shrubs, which are replaced farther eastward by chaco, a thorn-shrub vegetation that merges into *campos cerrados* (scattered trees in dense grasslands) in central Brazil and caatinga (stunted, sparse forests, leafless during the hot dry season) in eastern Brazil.

Caatinga

The Cape flora. The Cape flora is a restricted shrub vegetation in South Africa, consisting chiefly of heath and members of the protea family. This flora is remarkable in that it occupies an area of not more than 518 square kilometres (200 square miles), yet it is exceedingly rich in species, having more than are present on a several thousand times larger area of the central United States or of central Europe. It differs totally in its composition from the surrounding evergreen forest, karroo, veld, and desert vegetation, most of which occurs in a climate with winter drought and summer rains.

SCRUBLAND GROWTH AND DEVELOPMENT

Floristic comparisons between scrublands. All of the scrubland floras have developed locally and are not related to each other, except for a few genera they have in common. In the five main regions with scrubland vegetation, all major floral constituents are unique for each of them, yet there are certain plant genera or families that have representatives in more than one of the five areas. The scrublands of the Northern Hemisphere, for

Adapted from *Biological Sciences Curriculum Study Green Version High School Biology,* 2nd ed.; Chicago: Rand McNally & Co., 1968

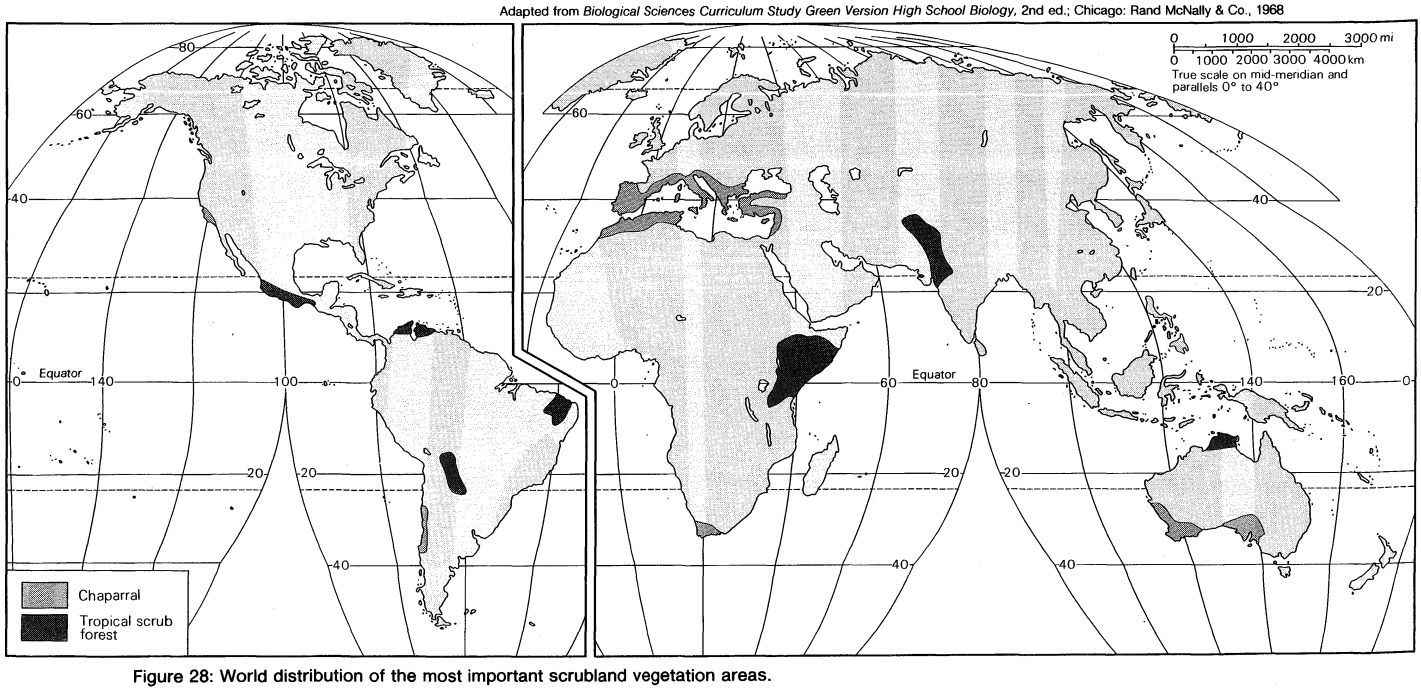


Figure 28: World distribution of the most important scrubland vegetation areas.

example, have a number of genera in common and so have those in the Southern Hemisphere. The scrublands on the same continental masses (Chile and California, Cape and Mediterranean) also have a number of related plant species. The Australian scrubland, however, has no relationship with either North American or European scrublands, nor has the South American scrubland any relationship with the European.

Seasonal aspects of scrublands. Scrublands in general do not change much in the course of a year. The shrubs are evergreen, and they change colour only slightly during the dry season, when the bright green foliage turns an olive green. After the first rains in late autumn, lush, new, green foliage appears and remains green into early summer.

The Mediterranean climate has mild winters, and most of the scrubland species are only mildly frost-resistant. In the chaparral, the laurel sumac (*Rhus laurina*) is probably the least frost-resistant, and after an exceptionally cold winter a number of shrubs of this species may be killed through freezing.

After an exceptional series of drought years, mortality among the less drought-resistant chaparral plants is greater than normal, and many sage shrubs die during such periods. The more resistant species, such as chamiso (*Adenostoma fasciculatum*) and desert buckwheat, survive even after drought periods. Thus these latter drought-resistant shrubs spread into the desert shrub vegetation. It is, therefore, not in the years with average winter temperature and average winter rainfall that selection in the chaparral occurs. It is in the exceptional years that the composition of the scrubland flora is determined.

Temperature and precipitation effects on scrubland composition

The most pronounced seasonal changes in the flora of scrublands are provided by annual plants. Their seeds germinate within a few days after the first extensive autumn or winter rain, and the plants continue to grow slowly during winter. In spring, flowers are produced in response to the lengthening of the days. Fruit is set and ripens before the long summer drought. Many of these annual plants occur both in the chaparral and in the desert. The major difference between these two habitats is the frequency of winters having rainfall sufficient for germination of the annual plants. Since annual plants require one inch of rainfall to germinate, they develop only when enough water has fallen to complete their life cycles.

Annuals occurring in both chaparral and desert are gold fields, which can produce spectacular displays in March and April, when large tracts of land become golden yellow with the flowers of this tiny composite, whispering bells, several poppy species, fiddle-necks, and different kinds of forget-me-nots.

In Western Australia the displays of annuals in the scrublands during September are just as spectacular, with the deep blue of *Leschenaultia* and *Dampiera* and the pinks of trigger plant (*Stylidium*) and sundew. Perennial herbs such as *Conostylis* and many orchids add to the flower displays in the underbrush. In the scrublands of the Mediterranean, mustards, legumes, and bulbous plants such as grape hyacinths lend special colour to the spring annual flora.

Evolution of scrublands. The scrublands vegetation provides many interesting evolutionary problems. In the southwestern United States, all scrub areas are relatively young in an evolutionary sense. Until 10,000 years ago the area now occupied by chaparral in southern California had a much denser forest vegetation in which sabretoothed tigers, giant sloths, and camellike animals lived. This oak woodland, which also occupied extensive areas in the Great Basin (now the semideserts of Nevada), gradually retreated against the onslaught of drought, which caused the disappearance of the great inland lakes—Bonneville and Lahontan—and which caused the extension of incipient deserts.

In the Mediterranean the climate probably was drier to start with and the vegetation did not change as rapidly as it did in the southwestern United States.

And in Australia the flora and fauna, like the continent itself, are very old. Consequently the mallee has not had invasions of exotic elements, and most of the shrubs growing in it have evolved in their present sites and long ago. They have evolved in directions different from other scrublands. The proportion of shrubs with spiny leaves, for example, is significant in Western Australia, as is the number of shrubs with spectacular flowers.

Occurrence of fires in scrublands. Scrublands form dense stands of vegetation that become tinder-dry toward the end of summers without precipitation and are particularly susceptible to fires. Not only the shrubs themselves but also the dry annuals and litter accumulation underneath provide plenty of fuel for extensive fires, which burn up into the forests bordering the scrublands at higher elevations. But they stop at the transition to the more arid regions where the vegetation becomes widely spaced. Not all scrublands are equally susceptible to fire. There are fewer and less extensive fires in the Mediterranean maquis than in the Californian chaparral or the Australian mallee.

All plants consist of combustible organic matter, but there are considerable differences in their flammability. There are two major factors that influence the ease with which a plant catches fire. One is obviously its water content, and thus water-filled cacti and succulents will not burn. The other is the quantity of aromatic flammable compounds contained in the plant and the ease with which they are released. The aromatic sage plants burn easily, whereas the Mediterranean rockrose (*Cistus*) is definitely fire-retardant.

Factors influencing plant flammability

A shrub fire is usually followed the next winter by destructive floods. For this reason attempts are made to reestablish as quickly as possible burned-over scrub vegetation. This is done by seeding a burned area with fast-growing annual plants. But since these plants—ryegrass and mustard—take some time before they produce a protective cover, and also because they need a good rain before they germinate, such seeding does not provide protection against the first rains after summer or autumn fire; moreover, a vegetation of annual plants needs to be renewed every year. The ideal solution, therefore, is to get as quickly as possible a renewal of the permanent perennial shrub vegetation. This is the natural sequence of events anyway. Whereas a climax forest vegetation usually does not directly regenerate itself and passes through one or several different successional stages before it becomes reestablished after destruction by humans or fire, chaparral regenerates easily. The old stumps of the burned shrubs often resprout and usually do so before the advent of the first rains. Secondly, the seeds of many chaparral plants are stimulated to germinate by the fire. Unfortunately, there usually are not enough seeds of chaparral plants present to get a good stand in the first year after the fire. Therefore a compromise is reached: as soon as possible after the fire, the slopes are seeded thinly with some annual plant that does not interfere with the reestablishment of the natural vegetation. The second year after the fire, the natural vegetation may have grown sufficiently to provide a fairly effective protection against erosion, which it does by shielding the soil against falling raindrops with foliage and litter. (F.W.We./Ed.)

Grasslands

Original extent of grassland

Evidence suggests that, before humans began the present rapid modification of environments by extensive agricultural and industrial operations, between 40 and 45 percent of the land surface of the Earth was occupied by grasslands, in which grasses and grasslike plants dominated in the absence of trees or with widely spaced trees. Although tracts of grasslands and savannas still survive in a modified state as rangelands for domesticated livestock and game animals, much has also been exposed to such intensive agriculture that the original plants and animals have been virtually eliminated.

THE ENVIRONMENTAL SETTING

Regions of natural grassland (Figure 29) occur where the environment is too arid for the development of closed forest but not so adverse as to prevent smaller, nonwoody but long-lived plants from forming a dense layer. Climate controls the biotic components of a region directly, through temperature and moisture extremes, as well as

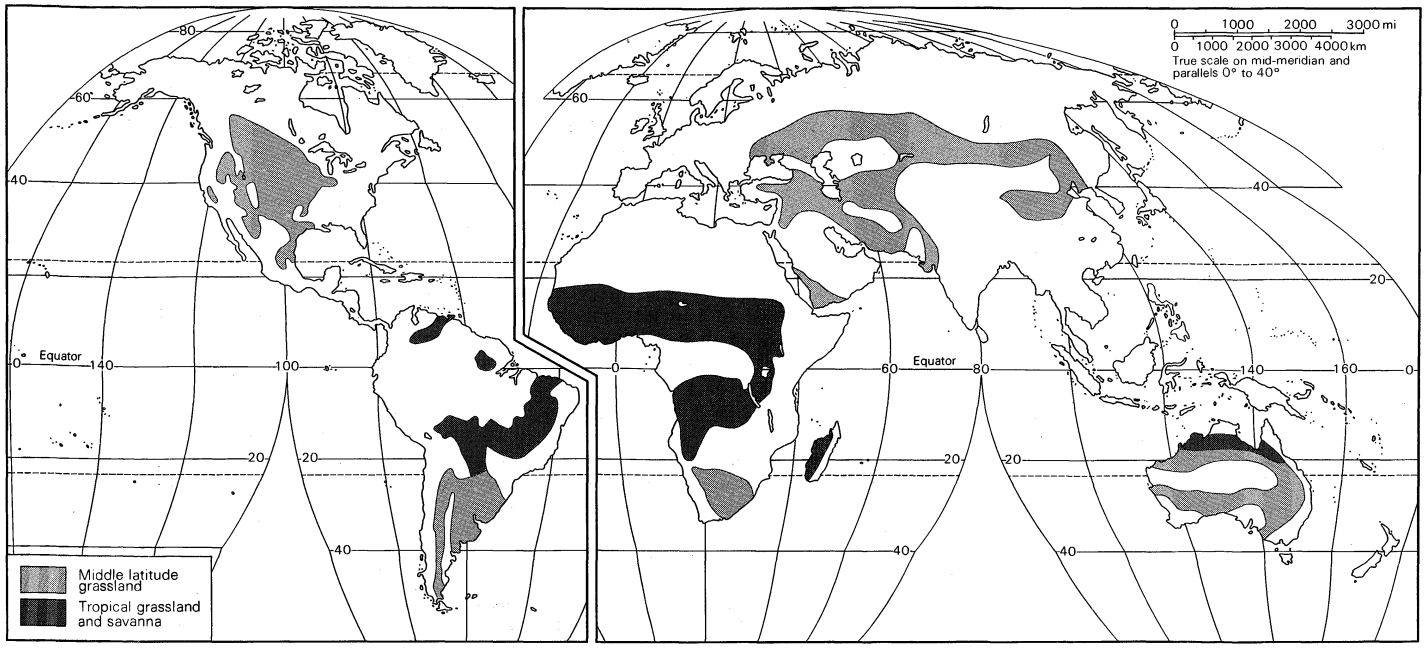


Figure 29: World distribution of grassland biomes.
Adapted from *Biological Sciences Curriculum Study Green Version High School Biology*, 2nd ed.; Chicago: Rand McNally & Co., 1968

indirectly, through its influence on soil development by such long-term physical and chemical processes as freezing and leaching. Within the grassland zone, however, local variations in topography and soils result in microclimatic conditions so different from the typical regional norms that nongrassland ecosystems dominated by plants and animals not characteristic of the remainder of the region may develop locally. In certain warm regions, for example, the woody habit is exhibited by many plant species even under arid conditions, so that specially adapted trees are scattered throughout the grassland to form a characteristic parklike vegetation type known as a savanna (see below).

Climate. The subhumid to semiarid climates of natural grassland areas are characterized by marked periodicity of precipitation, both from season to season within the year and between years. Consequently, annual drought periods of several weeks to several months are typical. The severity of drought increases with the distance from the forest margin. It is also accentuated by a cyclic climatic pattern that often results in several consecutive dry years. Droughts are particularly difficult for plants and animals in regions where the dry season is lengthy—as in the Mediterranean-type climate of some temperate grasslands and in most tropical and subtropical grasslands—or where a series of moister than average years is followed by several years of below-average moisture (see above *Scrublands*).

Annual precipitation in grasslands

Mean yearly precipitation in most temperate grasslands ranges from 250 to 750 millimetres (10 to 30 inches) and in tropical and subtropical grasslands from 600 to 1,500 millimetres (25 to 60 inches). Precipitation determines the nature and extent of natural grasslands through its effect on soil-moisture supply, since trees have difficulty competing with grasses in areas where upper soil layers are moist during part of the year but where deeper layers are continuously dry.

The adverse effect of drought on organisms is made more severe when accompanied by high temperatures. Increased duration and intensity of sunlight associated with decreasing cloudiness during periods of low rainfall tend to raise air and soil temperatures. In regions where the dry season is warm, the combined effects of drought and high temperatures make living conditions particularly adverse. Many organisms adapt to the hot, dry season in grasslands of tropical, subtropical, and Mediterranean climates and to the cold season (also a time of low precipitation) of many temperate regions by becoming dormant.

In the grasslands of tropics and subtropics, the length of the growing season is determined by the length of the rainy season, ranging from 120 to 190 days. In cold, temperate regions, however, temperature is a controlling factor. Growth begins in such regions when the mean daily temperatures reach 5° to 10° C (40° to 50° F). Although the frost-free season in cold, temperate grasslands may be as short as 100 continuous days, the vegetative season usually exceeds 165 days. Less temperature difference occurs among various grasslands during the growing season than during the dormant season. The temperatures during the warmest month in temperate grasslands are usually not lower than 15° to 20° C (60° to 70° F); in tropical and subtropical regions, the mean during the rainy season is commonly no more than 10° C (50° F) higher.

Grassland regions tend to have higher than average wind speeds, in part because of their low wind resistance.

Fire. Fire is an important feature of the grassland environment. Lightning-induced fires are common and, before their control by humans, periodically burned over extensive areas. The causes of fire are so closely related to the more or less arid nature of the grassland climate that it is not reasonable to separate fire from the more traditional climatic factors. Some geographers, however, have not been prepared to accept fires as a climatic factor and consequently have considered the natural vegetation to be that which would exist in the absence of fire. It seems more practical, in regions where aridity, continuity of combustible plant cover, and frequent thunderstorms lead to the frequent occurrence of natural fires, to consider the vegetation as it has been modified by fire to be the stable one. Fire maintains the boundaries of grasslands in climatic regions (usually the wetter portions of the grassland zone) capable of supporting forest growth. The effect of fire here, therefore, has been principally to prevent the extension of the forest into the grassland. The nature of the soil in such transition zones between grassland and forest is such, moreover, that the long-term existence of grassland is indicated.

In the temperate grasslands, a climate adverse to tree growth exists except near the forest edge, but, in the tropics and subtropics, trees and shrubs are often scattered throughout the grassland to form a savanna. The luxuriant growth of herbaceous (nonwoody) plants during the rainy season provides fuel, which, in the dry season, supports fires almost annually in many regions. Whether or not a forest would develop in the absence of fires in these tropical areas is speculative. It is most satisfactory to consider them as savannas in which the herbaceous layer predominates.

Fire's effects

Fire is more destructive to trees and shrubs than to grasses and other herbs, because the buds, from which new

growth will begin the next season, are located well up on the stems in a position vulnerable to fire. In herbaceous species, the growth zones are at or below the soil surface.

Topography. Many grassland regions are characterized by level to gently undulating topography. Often, however, the land surface is sufficiently irregular to provide variations in microclimate that result in a mosaic of vegetation types, as, for example, in the glacial moraines and loessial (loess is a soil material carried to its present location by winds) hills in parts of the Central Lowland and Great Plains of North America and in sandhills everywhere. Near the forest–grassland margin, grassland is favoured, even within the forest zone, on dry slopes that are subjected to greater radiation and wind than is level terrain. Conversely, patches of trees occur within the grassland zone on protected slopes.

Contrasts in vegetation on slopes with different characteristics are particularly noticeable in the higher latitudes, where the slopes facing the equator regularly receive more radiation than do poleward-facing ones. Various habitats also are found in sheltered basins and on exposed hilltops. Forest occurs within the grassland zone in protected basins, and patches of grassland occur within the forest zone on exposed hilltops.

Within the grassland zone, the herbaceous vegetation of depressions and poleward-facing slopes often consists of species that dominate on level ground in poleward areas. In the same regions, the species in exposed habitats are those that dominate communities that range toward the equator. Similarly, where grassland gives way to desert, desert species appear in exposed locations.

Soil. In regions where natural grasslands have developed, physical and biotic components have combined to produce soil different in its character from that of forest regions. In the temperate grassland, soil leaching (downward movement of water-soluble soil constituents) is restricted both by the scarcity of percolating water and by the low solubility of soil materials. The soil of natural grasslands, consequently, contains an abundance of organic matter at all depths, and chemical substances that are leached precipitate at the bottom of the soil profile (*i.e.,* the layering of soil as seen in a trench). Consequently, temperate grassland soils have the capacity to release nutrient elements to plants slowly and over a long period of time, thus permitting the production of annual crops, even after destruction of the natural plant cover. The abundance of organic matter (humus) in grassland soils is reflected in the colour of the upper part of the soil profile; in temperate grassland, the colour varies from black near the forest margin, where the organic content approaches 10 percent in some areas, to chestnut and to brown in areas where the organic content is below 3 percent.

Colour of grassland soils

In the tropical and subtropical grasslands, soil development is influenced by high temperatures and high precipitation in a process called laterization. By this process, soils become highly leached, and there is rapid decay of organic matter with low levels of humus accumulation. These soils are often reddish or yellowish in colour, reflecting the high content of iron left in the profile by the leaching of other minerals under conditions of high temperature. They are also much more subject to nutrient depletion under agricultural practice than are temperate grassland soils.

Differences in soil texture and soil structure throughout the grassland zones modify the vegetative cover because of influence on water regimes. Although sandy soils hold less moisture per unit volume, they permit more rapid percolation of surface moisture than do soils of finer texture, with a lower resultant loss by runoff and evaporation. Because of their greater moisture supply and the availability of moisture to greater depths, sands support stands of tall grasses or even trees in regions where grasses of smaller stature and less depth of rooting occur on soils of finer texture.

THE BIOTIC COMPONENT

Flora. The vegetation of natural grasslands is primarily composed of seed-bearing herbaceous plant species, which are classified, on the basis of gross morphology, into two groups: (1) grasses and grasslike plants (particularly sedges), which are collectively referred to as graminoids, and (2) nongrasslike herbs, mostly broad-leaved species, known as forbs. Woody components also occur, which, unlike the herbs, maintain perennial stems above the soil surface—*i.e.,* they live through several growing seasons. Frequently, there are dwarf shrubs that do not exceed the stature of the grasses; taller shrubs occur in some areas as isolated individuals or in patches. In some grassland areas, groves of trees are found in locally modified habitats such as along stream courses. Occasionally, colonies of non-seed-bearing plants, particularly lichens, club mosses, mosses, and algae, are found on the soil surface.

A considerable number of species constitute the vegetation of grasslands in areas of favourable moisture and temperature conditions, but the number decreases with increasing environmental adversity. In about 2.5 square kilometres (1 square mile) of rolling grassland in the black-soil region of eastern Nebraska, for example, more than 200 different prairie species are found; in a similar area of level brown soil 1,600 kilometres (1,000 miles) to the northwest in southern Saskatchewan, the flora of seedbearing plants is about 50 species.

The graminoids (grasses and grasslike plants) are well adapted to dominate in herbaceous communities, so that, although they typically make up less than 20 percent of the number of species present in grassland, they often furnish 90 percent or more of the plant biomass (dry weight) present. Similarly, though only a small percentage of the seed-bearing plant species belong to the grass family, their contribution to plant communities is great everywhere.

Success of grasses in occupying habitats

Grasses evolved from tropical woody ancestors, of which the bamboos are a present-day example. This origin is reflected in the relatively primitive nature of grasses composing the grasslands adjacent to forests in both tropical and temperate regions. Their adoption of the herbaceous habit, however, has permitted survival through such adversities as periodic burning and seasonal drought. Other adaptations that increase the chances of seed production in arid, windy, and cold environments include wind pollination, increased protection of flowers, and a reduced stature.

The graminoid life-form is particularly well adapted to dominate in competition with forbs. Success of grasses is associated with their ability to provide a dense plant cover in which seedlings of less aggressive species have difficulty becoming established. Graminoids are well adapted to exist in conditions of frequent fire and in areas of animal grazing by their manner of growth. New growth is from a zone of cell division located at the bases of leaves and stems rather than at the tips, as in forbs and woody plants. This permits continued growth when the upper ends of shoots or the leaves are cut, eaten, or burned. The growth zones in the stem also assist shoots that have been bent by trampling to regain a vertical position.

Grasses and the grasslands they dominate are sometimes classified as high, tall, mid, and short. High grasses, frequently reaching heights of 3 metres (10 feet) or more, are limited to tropical grasslands with high precipitation. Tall grasses are found in the most favoured parts of the temperate grasslands, moist habitats being particularly suited to them. Midgrasses, the flowering stems of which usually reach heights ranging from 30 to 90 centimetres (1 to 3 feet), dominate portions of the grassland closer to the woodlands of temperate regions. The short grasses, with leafy shoots often only 8 to 15 centimetres (3 to 6 inches) high, are dominant in the most arid grasslands. Between these extremes, the short and mixed grasses intermingle to form mixed-grass prairie, as in most of the Great Plains of North America.

The adaptation of each species of grass is also dependent on whether it is a turf former or a bunch grass. Turf-forming grasses spread by means of buds on lateral stems below the soil surface (rhizomes) or, much more rarely, above it (stolons). Bunchgrasses are able to spread horizontally only far enough to extend the size of the dense bunch, which eventually declines with age. Occasional development of seeds is necessary for the perpetuation of a bunchgrass cover but not of turf grasses. Bunchgrasses tend to dominate in arid habitats and turf formers in moist ones.

Grass species also differ physiologically in ways that adapt them to withstand adversity. Some species can continue growth during drought by extracting a greater amount of moisture from the soil. Others endure drought by achieving a dormant state before desiccation becomes excessive. Under extreme conditions, drought or overwinter, survival is achieved only by means of seeds, in which case the plants are known as annuals. Annuals are common in disturbed habitats in grassland, but they are more characteristic of the temporary herbaceous cover of deserts.

Root systems of grasses

The root systems of grasses are typically finely branched and rebranched to form a dense network extending through the soil. Depth of rooting of tall grasses and midgrasses is characteristically not greater than 183 centimetres (6 feet) and those of short grasses only 30 to 90 centimetres. These distances are also the maximum depths of moisture penetration in the soils of the habitats in which they dominate.

In some grasslands of both temperate and tropical regions, sedges (*Carex*) are associated with the grasses and have similar ecological relationships. The composites (*e.g.*, asters, sunflowers) commonly rank next to grasses and sedges in abundance. Under the favourable growing conditions near the forest boundary, forbs are abundant; the adverse conditions in the drier grassland regions, however, reduce their abundance relative to grasses in number of both species and individuals. Many forbs are adapted to arid conditions by rooting to depths much greater than the grasses, by surviving as dormant underground organs for periods as long as many years, and by storing water in their tissues. The root systems of some forbs are similar in structure to those of grasses, but a variety of other types exist, including taproots, bulbs, corms (rootlike or bulblike stem structures), and roots capable of forming shoot-producing buds. Species with the last type are particularly vigorous competitors with grasses and are especially abundant in the grasslands of the Soviet Union, where in some areas forbs are considered to dominate over grasses. Eurasian forbs with such root systems have become the most persistent introduced weeds of arable land in many parts of the world. Among them are perennial species of spurge, hoary cress, thistle, and morning glory.

The complexity of the vegetative cover of natural grassland is much greater than it appears to the casual observer. The many plants occupying grasslands may exhibit as much variety within a square metre as is present in about two hectares of forest.

The dominant and associated species are arranged horizontally into colonies and societies and vertically into layers, both above and below the ground. The number of layers discernible depends on the nature of the grassland. The uppermost one is composed of the tallest grasses and forbs. Some forbs have vegetative growth confined to lower layers but thrust their flowering parts, often on leafless stems, upward above the grass layer to ensure pollination and seed dispersal. Others grow vegetatively at a rate similar to that of the taller grasses and bear their flowers on leafy stems. A lower layer is made up of species of shorter stature, some of which are secondary species that grow early in the season, before the dominants cover them.

The lowest layer above the ground is the crust of lichens, mosses, club mosses, and algae that often occurs on the soil surface among the fallen stems and leaves (litter) of the upper layers. Various groups of plants tend to root to different depths, so that short grasses and small sedges form a shallow layer underground and taller grasses a deeper layer. The roots of some forb species often branch only below the deepest layer of grass roots; they sometimes penetrate to depths twice that of the dominants.

Seasonal aspects of grasslands

The appearance of temperate grasslands changes from season to season as a result of the successive flowering of different species. A few species thrust their flowering stems upward shortly after the disappearance of snow and before they are shaded by the spring growth of the cool-season grasses. These are followed by the spring flowers, which, in the temperate grasslands, are composed of many forbs including early legumes (members of the pea family). Early grasses and some sedges begin to flower near the end of this period. Most of the dominant grasses, of both warm and cool seasons, blossom during the summer, as do most of the legumes, all of which give a predominantly bluish and whitish tinge to the landscape at this time. The predominant colour of autumn flowers is yellow, and the floral display is made up principally of an abundance of composites. During these shifts in the seasons, the grasses also change in the early spring from the grayish colour of the preceding year's old growth through rich greens as the new leaves emerge to a straw-coloured hue of pale yellow in the late summer and early fall. Curing of shoots to a yellowish colour is characteristic of the more arid grasslands; those in more humid areas turn gray more rapidly. The growing season varies in character from year to year as differences in weather differentially stimulate or retard flowering of the various species. The grassland tends to be drab and relatively lifeless during the winter. The changes in the season have not been as well documented in the tropical grasslands, in which the main changes apparently take place with the shift of the season from dry to rainy, with the resultant prompt resumption of luxuriant green growth that ends only with the renewed onset of drought. Some tree and forb species in tropical savannas anticipate the change in the season by leafing out before the arrival of the rainy season.

Animal life. The natural animal populations in grasslands are much more diverse than is generally realized, as many surface species are small and inconspicuous, and many other species live underground for at least part of their lives. The best known are the large grazing mammals, or game animals. Other well-known groups in grasslands include grazing marsupials (pouched animals, such as the kangaroo), predators belonging to the cat and dog families, rodents of various sizes, birds, lizards and snakes, and the larger insects, particularly grasshoppers and locusts. A large proportion of surface species are either running or burrowing types. The characteristic aggregation into colonies or herds provides some measure of protection in the open type of habitat. Invertebrates of various kinds are abundant in the soil.

Vertebrates. The large grazing mammals, subject to the influence of humans, have been reduced in number, and some species survive only in protected areas. Their former ranges have become densely populated and are being grazed by cattle, sheep, horses, and goats. In some of the managed reserves in Africa, the natural and domesticated populations intermingle, but for the most part, the native grazing species are considered intruders, and their predators also are excluded by poisons, hunting, or fences.

Large animals of grasslands

The African grasslands were once abundantly supplied with a large number of species, of which perhaps the wildebeest, gazelles, and zebra were the most numerous over large areas. Other species included buffalo, giraffe, eland, antelope, hartebeest, hippopotamus, waterbuck, and impala. Some species, which are browsers as well as grazers, have had significant influence in restricting the growth of trees. These herbivores are accompanied by a variety of omnivores (animals, such as the warthog, *Phacochoerus aethiopicus,* that feed on both animal and vegetable substances) and carnivores (meat eaters), particularly the lion, leopard, hunting dog, and fox. The hyena (family Hyaenidae) is perhaps more truly a scavenger, but it also fills the role of predator while traveling in packs. The characteristic grazing species of the Asian grasslands include wild cattle, saiga antelope, wild horse, marmot, stag, and boar; predators are represented by the wolf and the fox. Wild horses and cattle, formerly abundant, have been exterminated by humans, and antelope and marmot have largely been forced out of the grassland region.

The number of species of large grazing mammals in the Western Hemisphere, while it was larger during the early geologic stages of development of grasslands, has been low in recent geologic time. In North America the principal species once were the bison and the pronghorn antelope. Their most important predators were the coyote and bobcat. In South America the large grazing mammals have been less influential; rather, the llama and its relatives and the ostrich, a native of Africa, are the most important. The isolated fauna of Australia includes several kangaroo species as the major large grazing animals, with the dingo as their prime predator.

In some grasslands the large grazing mammals are as important as climate and soil conditions in determining the nature of the vegetative community, since some plant species are more sensitive to grazing pressure than others.

Smaller herbivorous mammals, common in natural grasslands, include many species of rodents, such as mice, voles, shrews, ground squirrels, gophers, prairie dogs, hares, and rabbits. Some species have increased their populations in regions where humans have reduced the numbers of their predators, others, such as the prairie dog (*Cynomys*) in the Great Plains of North America, have suffered from a direct effort by humans to exterminate them. Rodents, now reported to be the most frequently found mammals in some grasslands, are credited with causing widespread range deterioration, because they modify the speciation of the plant cover and expose the soil surface to erosion.

Many birds, including herbivorous, carnivorous, and omnivorous species, inhabit grasslands. The herbivores are the least numerous in terms of different species, and the omnivores are most abundant. The passerines (perching birds) are particularly common and include larks, longspurs, meadowlarks, starlings, grouse, cranes, partridges, and doves. Hawks, owls, and eagles are important predators.

The amphibians and reptiles are important organisms in many grasslands. Lizards, toads, and box turtles are mainly functional as predators of insects; snakes are predators of rodents and some other small vertebrates. These organisms, probably the least conspicuous components of grasslands, are often, however, among the most abundant vertebrates.

Invertebrates. Great numbers of species of insects and other invertebrates inhabit grasslands, forming large populations of individuals. Most conspicuous are the grasshoppers (Orthoptera) and their relatives. The numbers of surface insects present in grasslands have been estimated to be as many as 1,000 per square metre (about 100 per square foot), and the number of species present in one type of grassland may exceed 200. Insect bulk, or biomass, has been reported to be greater in some areas than that of large grazing mammals. Next to the grasshoppers, the insects with the most pronounced effect on herbage are probably bugs, aphids, and leafhoppers. Ants and termites are found particularly in tropical and subtropical grasslands. The larvae of beetles (order Coleoptera) feed on roots of plants and on feces, especially those of large grazing animals. Larvae of flies (order Diptera) often affect seed production and are active in decomposing carrion. Larvae of moths and butterflies (order Lepidoptera) feed on the crowns of grasses; and thrips (order Thysanoptera) also affect seed production. The most abundant predatory invertebrates above the soil surface are spiders (class Arachnida).

Insect numbers and bulk

Small soil animals are important to the grassland ecosystem. Chief among them are roundworms, springtails, mites, and small segmented worms, as well as insects. Earthworms, although common in semipermanent and seeded grasslands, are rare in natural grassland conditions.

Microorganisms. Microflora (bacteria, actinomycetes, fungi, and algae) probably are found in greater numbers in grasslands than are microfauna (protozoans). The great majority of microorganisms present in the soil are apparently dormant or inactive at any one time, although the bacterial biomass to a depth of 15 centimetres (6 inches) in agricultural soil has been estimated to range from 330 to 720 kilograms per hectare. Grassland soils, because they have considerable root systems that provide habitats for markedly higher numbers of bacteria, thus contain an even greater bacterial biomass. Fungi, although present in appreciably lower numbers than bacteria, have much larger cell sizes and are estimated to possess a biomass twice that of the bacteria. Algae are commonly estimated to contribute less than 10 percent of the total microfloral biomass in mature soils.

COMMUNITY DEVELOPMENT

Stages. Grassland ecosystems have developed through an orderly series of changes in plant cover and soil, as one has affected the other through climatic influences. These progressive changes in plant cover are known as primary plant succession. As plant succession progresses toward a stabilized plant community, a soil gradually develops a profile that reflects the nature of the plants occupying it. The end of this process, a climax situation in which the organisms and soil are in equilibrium with the climate, is a stable system with many buffers to protect it from changes in climate and animal populations.

Primary succession of grasslands begins on bare, dry land surfaces or in water. In each situation, development is toward a more moderate water regime. This is accomplished in the dry habitat by the development of an insulating plant cover that protects the surface from the desiccating effects of the sun and wind and by the incorporation of organic matter into the soil, which increases its water-holding capacity. In wet habitats, the process involves gradual elevation of the substrate (bottom of the water body in question) by deposition of both organic and mineral particles.

The series of changes in plant cover resulting in grassland begins with a layer of crustose lichens, followed by foliose (leafy growth form) lichens. Eventually, annual herbs and a series of perennial species—first forbs and short-lived grasses, finally a stable community of long-lived grasses and forbs—becomes established.

Gradual succession of water bodies to grassland

The succession beginning in a body of water starts with submerged aquatic species and progresses through floating-leaf, reed-swamp, and sedge-meadow stages as the water level decreases. This moist habitat may survive for thousands of years and permit the development of various vegetative stages (even of shrubs and trees) before the depression containing the water gradually fills from the edges and the vegetation becomes dependent, for the first time, on climatic factors (chiefly precipitation) that permit the establishment of a climax grassland cover.

The uniformity of the climax vegetative cover depends on the degree of irregularity of the landscape, for the final stages of the dry-habitat succession will not be reached while exposed locations survive on sunny and windy slopes and on hilltops, and protected situations are preserved on sheltered slopes and in depressions. The regional characteristics of the vegetative cover, however, are expressed in topographic locations between these extremes. It is on this portion of the landscape, which in most grassland areas constitutes the major area, that the following brief survey of the grasslands of the Earth is based. While this survey considers the temperate regions separately from the tropical and subtropical grasslands, the distinction is arbitrary. Subtropical grasslands give way imperceptibly to temperate communities through the arid areas where contact usually takes place.

Temperate grasslands. The most extensive grassland areas of the temperate zone are the steppes of Eurasia, the prairies and plains of central and western North America, and the pampa and adjacent areas of Argentina. Less extensive areas are found in the velds of South Africa, in the mountain grasslands of South America, and in Australia and New Zealand.

North America. The North American grassland is composed of seven regional types. The largest, the mixed prairie, extending from southern Alberta and Saskatchewan southward to western Texas, lies between the 100th meridian (west longitude) and the foothills of the Rocky Mountains to the west. It is dominated in the north by bunch-forming spear grasses and turf-forming wheatgrasses, which are midgrasses that gradually give way to short grasses southward, particularly blue grama (*Bouteloua*) and buffalo grass (*Buchloë*). The next largest area of grassland east of the Rocky Mountains occurs in a higher precipitation belt extending from the mixed prairie to the eastern deciduous forest. Although it is now the "corn belt," this area was formerly dominated by spear grass and dropseed in the drier habitats and by bluestems in more favourable locations. Both types occur in areas east of the Rocky Mountains, where there is an early-summer peak in precipitation, followed by late-summer drought. West of the Rockies the most northerly grassland area occurs in eastern Washington and adjacent parts of Idaho and Oregon and in the valleys of British Columbia. The dominant grass species is a large, bunch-forming wheat-

grass. The upper slopes in the northern part of this region are occupied successively by remnants of mixed prairie to the east and then by fescue prairie. The latter also occurs as a belt around the mixed prairie's northern edge in the eastern foothills of the Rocky Mountains and in parklands of Canada's Prairie Provinces. Southward, in the Mediterranean climate of the valleys of California, the native, perennial grass cover has been replaced—through overgrazing—by annual weedy grasses, but spear grasses formerly dominated. The southwestern United States and north-central Mexico contain the most arid prairies, the desert plains grasslands, which surround the warm desert at elevations between 300 and 1,500 metres (1,000 and 5,000 feet). Short grama and wire grasses abound here in an environment similar to that of the dry subtropics. The region of Texas adjacent to the Gulf of Mexico also has subtropical vegetation characteristics.

South America. The best-known temperate grassland in South America is the pampa of east central Argentina, which occupies a region of level, black soil. The annual precipitation of 1,000 to 1,250 millimetres (40 to 50 inches) is the highest of the temperate grasslands, and frequent drought periods do not occur. Similar grassland extends on more rolling topography to the north through Uruguay into the semiarid campos region of southern Brazil and northwestward in Argentina. Westward, the pampa becomes drier and the soils gradually change to chestnut and brown and then to gray in colour near the Andes Mountains. In the driest areas, shrubs and small trees (mesquite, senna) form a dwarf savanna. Southward, the pampa gives way to the semiarid regions of Patagonia. Only part of these areas have grassland character. The extremely dry regions contain shrubby, cold desert vegetation. The pampa and the adjacent less humid grasslands have been exposed to grazing by domesticated livestock for about 400 years. Heavy grazing, particularly in the last century, has modified the vegetation to the extent that the identity of the natural dominant species is not known, although spear grasses are thought to have been included among them. The vegetation is now composed principally of introduced species. Cool mountain grasslands are extensive at various altitudes in the Andes, occurring at high elevations even in the tropics. These areas are dominated by species of spear grass, fescue, bluegrass, and reed grass.

The pampa

Eurasia. The body of natural grasslands in Eurasia lies within the Soviet Union, extending from north of the Black Sea in the southern Ukraine eastward through northern Kazakhstan and southern Siberia. These steppes are bordered on the north by forest steppe where the elements of the boreal forest intermingle with the grassland, as is the case in the similar climate at the northern edge of the Great Plains grassland of North America. Southward, the climate becomes warmer and drier, particularly in the Asian portion, and the soils gradually lighten in colour from black to dark brown, brown, and gray at the edge of the cold-shrub desert region. The most important dominants are species of spear grass, feather grass, and fescue. Here, the species of fescue that dominates over large areas is more drought-resistant than the dominant spear grasses and feather grasses, a situation the reverse of that in the northern American grasslands.

Africa and Australia. The South African veld, the tussock grasslands of New Zealand, and the grasslands of Australia are much less extensive. They are similar mainly in that oat grasses are dominants in each. In Australia, however, the oat grass-dominated grasslands are restricted to the southeast, with more extensive grasslands dominated by Mitchell grasses, which occur in arid situations blending into subtropics. Grassland of temperate character also occupies the highlands of East Africa.

Tropical and subtropical grasslands. Tropical and subtropical grasslands are located in central Africa and central South America.

Africa. In Africa, semidesert and desert grasslands occur in areas where the annual rainfall averages less than 500 millimetres (20 inches). Important species in such lands include spear grasses and wire grasses, and the woody element is limited to low, thorny shrubs.

South America. The most extensive tropical grassland area in South America is the Llanos of Venezuela, which lies between the Andes Mountains and the Orinoco River. This region is composed of a complex of plant communities including grassland and savanna, with high grasses dominating in some parts and tall grasses in others. Over large areas there are no trees, apparently because of waterlogging of soil. The best-known dominants in many of these areas are Para grass and guinea grass. The semiarid region of northeastern Brazil is similar to those on the equator side of the deserts in Africa, with wire grasses among the dominants and also with scattered shrubs.

The Llanos

Lands surrounding the Mediterranean Sea and those in Asia Minor and parts of southwestern Asia (Iran, Pakistan, Afghanistan, Tadzhikistan [in the Soviet Union], and India) have been exposed so long to the effects of human occupation that it is no longer possible to determine their original character. It is probable that many areas that are presently desertlike were previously grassland. Elsewhere, the change from grassland to desert because of human occupancy has been observed during the last century.

FUNCTIONING AND PRODUCTIVITY OF GRASSLAND ECOSYSTEMS

The organic component of a grassland ecosystem is a complex of producers, consumers, and decomposers that are highly organized into a food web. The system is driven by energy fixed from sunlight during the process of photosynthesis in the green tissues of plants (the producers). This energy is passed on in the form of organic matter—used as food—from one to the other of various groups of organisms, each of which liberates some of the energy in its own body functions. Finally, the photosynthetically produced organic matter is transformed back into mineral elements, which may again be used by plants in photosynthesis.

The biological productivity of such a system is expressed in terms of the rate of fixation of solar energy by the producers and is known as primary productivity. The rate at which the consumer biomass increases, by transformation of plant materials into animal tissue, is referred to as secondary productivity. Primary productivity for natural, temperate grasslands ranges mostly between 200 and 1,000 grams per square metre of dry matter per year; that in the most favourable tropical areas is much greater. Secondary productivity is considerably less than primary productivity in all regions. In the central Great Plains of North America, for example, yearling cattle consume only a portion of the vegetation present in their grazing area, and of this it is estimated that only 9 percent is used in meat production, 48 percent being lost as heat to the atmosphere and 43 percent to other food chains in the form of feces and urine.

Plant species in natural grasslands are sensitive to the removal of more than a moderate proportion of their shoots. Unconsumed shoots die and remain standing, in some communities for periods as long as two or three years, during which time they protect the soil against erosion and contribute toward increasing the rate of its water intake. In some grasslands, there is more of this "old growth" present than the current season's crop of green shoots. The old shoots gradually decay or fall to the ground to produce a litter layer. The depth of litter depends on amount of growth, amount of consumption, and rate of transformation (by microorganisms and soil animals) to soil humus.

Old growth

The plant biomass (including underground parts) of natural temperate grasslands is usually in the range of 1,000 to 3,000 grams per square metre of dry matter. As much as 85 percent of this biomass occurs underground. Dead material constitutes only a fraction of the aboveground parts (both standing and as litter) but up to 75 percent of the underground parts. About half of the energy fixed in a given year is deposited in underground tissues. The rate of decomposition of dead, unconsumed plant parts is such that, on the average, the amount of biomass present disappears in about two years above ground and about four years under ground. The amount of plant litter present usually reaches an equilibrium state (*i.e.*, the condition that exists when the rate of disappearance equals the rate

of new litter added), which varies from less than 100 grams per square metre in dry, temperate grasslands to about 1,000 grams per square metre in black-soil areas. Lower values than these prevail in areas where a higher than usual proportion of the shoots enter the consumer food web or where recurrent fires occur.

In grassland stands in the subhumid tropics and subtropics, biomass of herbage sometimes exceeds 5,000 grams per square metre; but the rate of decomposition of organic matter is high in such areas, leading to low quantities of litter and soil organic matter.

It has been estimated that about two-thirds of the energy of the grassland ecosystem is released through the activity of reducers and decomposer organisms that feed on detritus and animal wastes. This activity is almost all in the soil or in the litter at the soil surface. Only general conclusions can be given concerning the proportion of this activity caused by invertebrates and other decomposing reducers, since it has not yet been quantitatively evaluated. It is generally agreed, however, that the soil microflora (microscopic plants, including fungi and bacteria) is the single most important group of organisms affecting the turnover of energy. The biomass of soil microflora has been estimated to be 400 to 600 grams per square metre of soil surface on a dry-weight basis; however, probably less than 1 percent of this is active at any one time.

It is not yet possible to identify to what extent various groups of soil microflora contribute to the total of decomposer activity in any given system. The rate of decomposition in standing dead vegetation is usually slower, however, than in litter and in vegetation in contact with the ground. The bulk of the microbial decomposition does not occur until the litter has either made contact with the soil or has become densely compressed just above the soil surface. The smaller macrofauna (animals bigger than microbes), such as litter-feeding insects and worms, are active in bringing litter into more intimate contact with the soil.

UTILIZATION OF GRASSLANDS

Many early human civilizations developed in grassland regions, so humans should be familiar with the ecology of grasslands. Greater interest, however, has been shown in converting grasslands to the growth of annual crops than has been devoted to considering whether they would have been a more valuable resource in an untilled state.

Domesticated grazing animals

Domesticated grazing animals occupy the most important role in the human conversion of natural-grassland plant growth. While conservationists in the past may have seen the domesticated grazing animal as an intruder in natural grassland, this does not necessarily mean that the grassland environment deteriorates through the replacement of natural by introduced animals. The philosophy of range management that has developed in North America is based on the concept of obtaining the highest sustained level of animal production on natural grassland that is compatible with maintenance of the resource. Range ecologists have been much more conscious of the need to conserve land resources than have agriculturalists. In the management of arable lands, for example, the guiding principle has been almost exclusively determined by the need to produce the maximum harvestable yield, a practice hardly compatible with conservation. The advantage that ecologists see in the use of domesticated livestock in the rational management of rangelands is that the distribution and density of these animals is under their control to a far greater extent than would be possible with native animals. The domestication of the range is thus seen as a stabilizing situation.

The success achieved in increasing the harvestable yield of intensively managed arable land and improving semipermanent grasslands of woodland climate has led to the suggestion that the plant cover of nonarable grasslands should be changed as much as possible by the introduction of domesticated forage crops that have been selected and bred for high yield and for optimum response to fertilization and management. A high degree of environmental control is needed, however, to utilize the higher potential of these species, and the indication is that the native grass cover cannot be excelled by introduced species for range production on nonarable land.

Effects of herbicides and fertilizers

Attempts have been made to increase the productivity of natural grassland by the use of herbicides and fertilizers. Weed control of rangeland is economically practical only in cases where the weedy situation has been induced by mismanagement and when this situation is corrected. The effect of fertilizers is variable, yielding better returns where moisture conditions are most favourable. Some native species do not respond to increased levels of nutrient supply and may be replaced after fertilization by species that are more productive but dependent for survival on a continued supply of fertilizer nutrients.

After the initiation of tillage, highly fertile, temperate grasslands gradually (over a period of 50 to 100 years) attain a new level of equilibrium, which is associated with a lower content of soil organic matter. The full impact on organic-matter content will not occur until the original organic material is replaced by that formed from the annual agricultural plant cover. The concept that corrective measures can be taken by future human generations by addition of chemical fertilizer does not account for changes in soil structure that may be associated with declining organic content.

The maintenance of both arable and nonarable ecosystems in grassland zones is vital to the continued provision of food for the world. The temperate grassland zones include an important portion of the cropland of the Earth (for example, 90 percent of the grain for commerce originates here); the tropical and subtropical grasslands provide a possible means for expanding agriculture when technology is developed to manage these lands on a long-term basis. (R.T.C./Ed.)

Deserts

Deserts are arid areas of sparse to absent vegetation and low population density that constitute more than one-third of the Earth's land surface, if semiarid regions are included. Approximately 5 percent of the Earth's land area can be categorized as extremely arid; the regions involved are the central Sahara and the Namib Desert areas of Africa, the coastal areas of Ethiopia and Yemen (Ṣan'ā') near the southern end of the Red Sea, the Rub' al-Khali in Saudi Arabia, the Takla Makan Desert in central Asia, the Atacama Desert of Peru and Chile, and parts of the southwestern United States and northern Mexico. Although the desert environment exists on parts of each of the continents, the nature and diversity of deserts are not widely understood.

The fallacy of the popular view of deserts

The popular view of deserts as predominantly sandy areas that have existed in their present locations throughout geologic time is an unfortunate misconception. Sands, for example, may cover about 10 percent of the surface area in the Sahara and, where dunes or sand sheets occur, are certainly prominent features of the Namib, Arabian, and some other deserts. In general, however, desert sands must be regarded as a minor portion of the deserts as a whole and are essentially absent over vast reaches of arid terrain. And with regard to the permanence of deserts, it is fair to say that the land areas subjected to arid conditions have varied widely throughout Earth history. This is a consequence of both continental drift—the shifting about of landmasses on the Earth's surface, thus bringing different regions into conjunction with arid climatic zones—and climatic changes per se.

During the Quaternary period (the last 1.6 million years) some profound climatic changes of uncertain synchroneity with the ice ages affected some desert areas. Artifacts indicate that early humans hunted large mammals in areas that could not possibly sustain such life today; forests of oak and cedar grew in such highland areas as Tibesti, in the central Sahara; lakes and lake systems were at one time extensive in such regions as the Kalahari, the Iranian desert, and the western United States, where there is evidence of a former body of water 180–210 metres (600–700 feet) deep in Death Valley, for example. Indeed, relict surface features of many kinds, as well as some relict animals (specialized fishes and some crocodiles in oases), point toward the existence of former moist conditions in many of the world's desert areas in the relatively recent

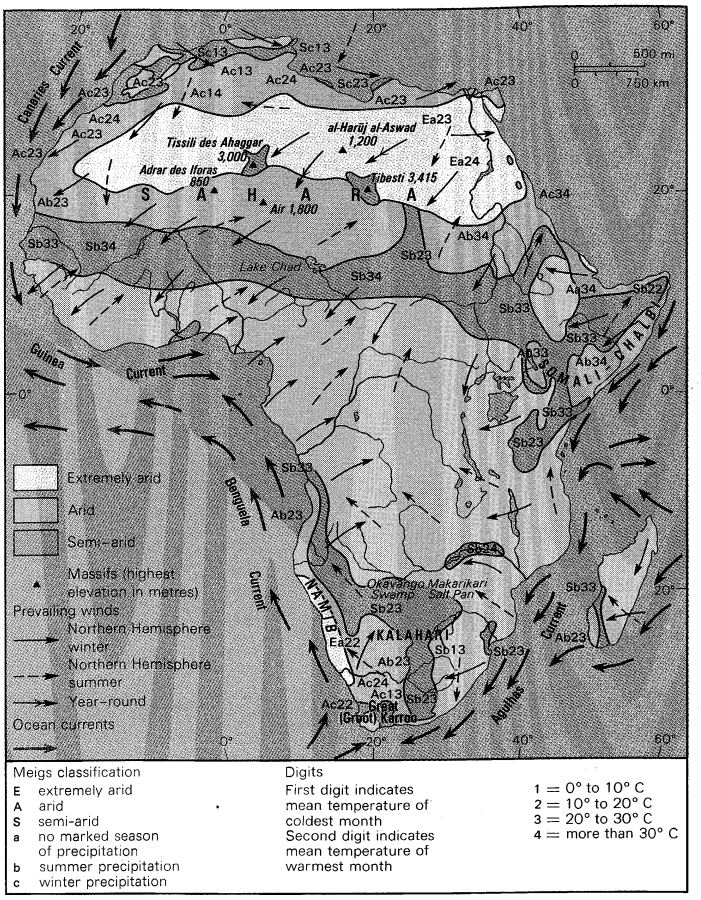


Figure 30: Deserts of Africa.

Adapted from UNESCO, Reviews of Research on Arid Zone Hydrology, "World Distribution of Arid and Semi-arid Homoclimates" (1953)

past. The same may be said of habitation in Roman times in North Africa. The theatre at Leptis Magna, near present-day al-Khums, Libya, for example, was designed to seat many thousands, but the site today is reminiscent of a ghost town.

ENVIRONMENTAL FACTORS

Topography. A considerable number of landforms occur in desert areas, and though some are considered characteristic of arid regions, nearly all have counterparts in the more humid environments. Mountain ranges and plateaus, volcanic peaks and isolated bedrock knobs, peripheral landforms that are alluvial fans if composed of sediment and pediments if consisting of a sloping rock surface, sand sheets and sand dunes, playas and other dry lakes, and desert plains and depressions—these are the principal elements of the desert landscape, which is often described as stark, harsh, and angular in profile. This is to some extent more closely related to the relative absence of vegetation and soil cover in arid regions than to some fundamental physiographic distinction that is climatically based (see further CONTINENTAL LANDFORMS).

Principal landforms

Aridity and its causes. The general cause of the arid zones of the Earth and, hence, of the present location of the world's deserts resides largely in the dynamics of the atmospheric circulation. In the simplest sense, the engine that drives the planetary atmosphere is incoming solar radiation combined with the effects of the Earth's rotation. Much more incident heat is received in the equatorial region than at the poles, and this tends to lead to a poleward transfer of heat and to a meridional atmospheric circulation pattern. The rotation of the Earth produces latitudinal wind zonation at the Earth's surface—*e.g.*, the trade winds, or easterly winds, in equatorial regions and the westerlies in middle latitudes. Although the physical causes of these two circulatory patterns are complex, the reason that they lead to the existence of arid zones is easily understood.

Trade winds and westerlies

Basically, air that is heated at the equator by incident solar radiation rises and cools; its moisture condenses and is released in the tropical zone; the air then subsides toward the Earth's surface near latitudes 30° N and 30° S, thus producing two great belts of subtropical high pressure. This descending air is also the source of the easterly trade winds that blow toward the equator. Some of the world's greatest deserts are located beneath or near these high-pressure belts—the hot, dry trade winds blow across the Sahara (Figure 30), the deserts of the Middle East and South Asia (Figure 31), and part of the North American Desert (Figure 32) in the Northern Hemisphere, and this effect of the general planetary circulation is also responsible for the occurrence of the Atacama-Peruvian desert of South America (Figure 33), the Namib Desert and the Kalahari in southern Africa, and the Australian desert (Figure 34) in the Southern Hemisphere. It should be noted, however, that where the trade winds blow onshore, as along the east coasts of Africa, South America, and Australia, the moisture they bear precludes the existence of deserts.

These generalizations concerning the locations of the world's deserts and wind systems are borne out by several of the maps shown in Figures 30 to 34. Aside from the disruption in continuity of the arid zone along the east coasts of the continents mentioned above, it also may be noted that several of the world's deserts can be characterized as high-latitude types—*i.e.*, they lie north or south of the principal arid belts. Included in this group is much of the North American Desert and the Patagonian Desert of Argentina. These arid areas result principally from physiographic causes: a rain-shadow effect is involved. The latter term is applied when warm, moisture-laden winds must cross a mountain system or similar topographic barrier. When this occurs, as on the west sides of the Andes and the Sierra Nevada, the air masses cool as they rise, and condensation and precipitation occur over the mountains proper. When the air descends on the leeward sides of such barriers, it is thus devoid of moisture and deserts result. The deserts of central Asia, principally the Takla Makan and Gobi, are also related to physiographic causes in the sense that their locations in the continental interior are remote from any sources of moisture. The distribution of land and sea can therefore be cited as a second general cause of aridity on the Earth's surface.

Principal hydrologic factors. The factors that govern the hydrology of arid regions are climatic, topographic, and geologic. Considering climatic factors first, precipitation and evaporation are perhaps most important because these two variables tend to control the water balance of an area. Precipitation is highly variable in arid areas, and as a general rule the variability is inversely related to the mean annual precipitation—*i.e.*, as annual precipitation decreases, the variability increases. This variability, in turn, is directly related to the nature of surface runoff in desert regions; it is spasmodic, and streamflow on a large scale will occur only after intense rainfall.

Variability of precipitation

Another moisture source in deserts is dew, which is worthy of mention because its contribution to the water input and, hence, to the general hydrology of an area has been overemphasized upon occasion. Dew can derive from the condensation of atmospheric water vapour or from water vapour that is emitted from moisture present in the soil. In the latter instance, there clearly can be no net water gain, so the dew derived from soil moisture can be disregarded. The maximum total dewfalls reported do not exceed 0.2 millimetre (0.008 inch) per day in any case, whereas total daily evaporation of about 5 to 6 millimetres is not uncommon. On these grounds it can be argued that the contribution of dew to the overall water balance in arid regions is of negligible import. Its greatest significance is (1) in west-coast deserts such as the Namib in Namibia, where the influx of humid air from adjacent oceans permits optimal dewfall that benefits vegetation and (2) as a necessary moisture source for the chemical weathering of rocks.

Evaporation in all desert regions is a factor of great significance. It is here designated a climatic factor, because it is related to incident solar radiation and air temperature and to atmospheric humidity. The total annual evaporation

Evaporation in desert regions

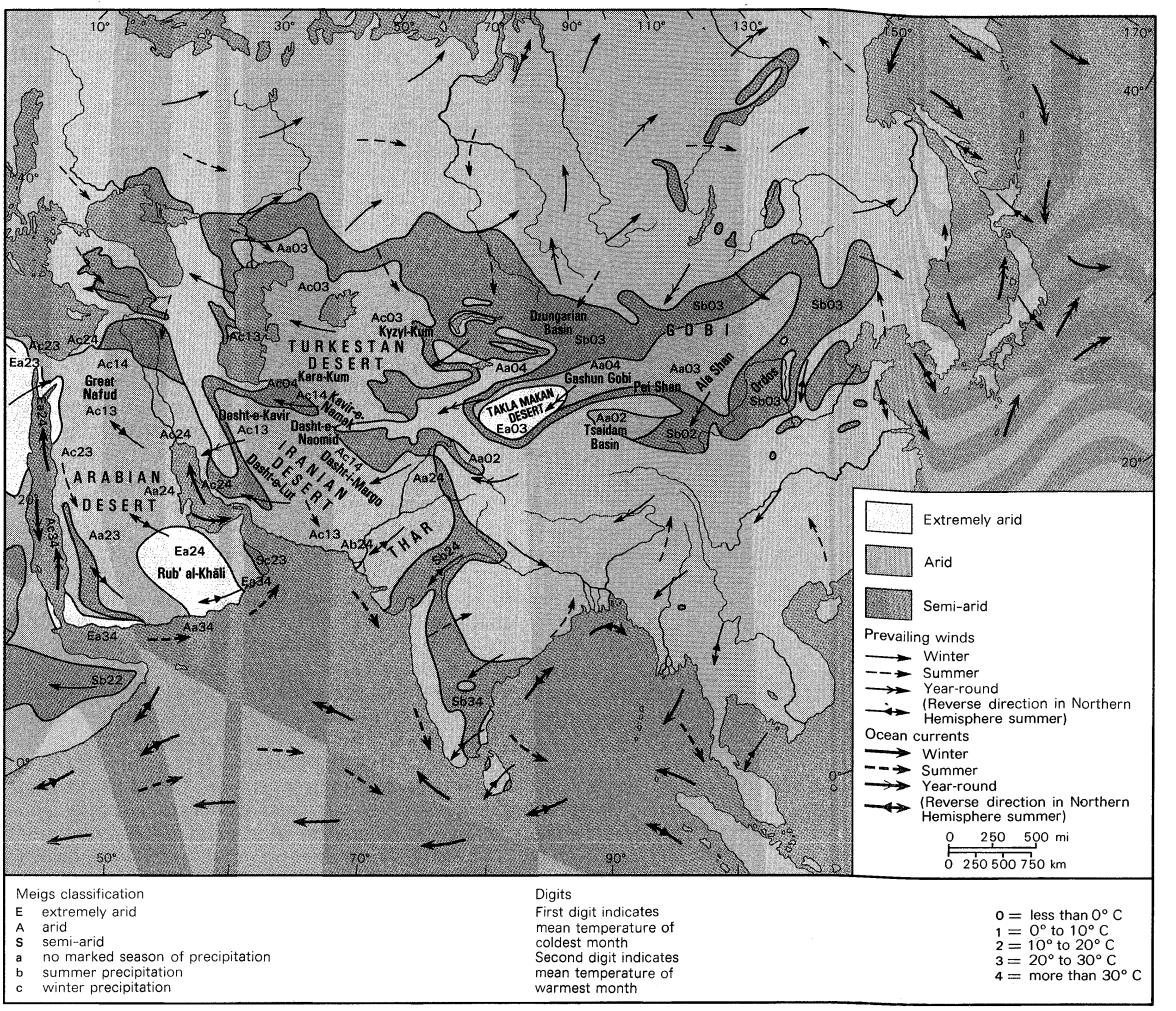


Figure 31: Deserts of Asia.

Adapted from UNESCO, Reviews of Research on Arid Zone Hydrology, "World Distribution of Arid and Semi-arid Homoclimates" (1953)

everywhere exceeds total precipitation significantly, and, indeed, this relationship can be considered one of the climatic hallmarks of the arid regions. Wherever evaporation exceeds precipitation, the presence of free-standing natural water bodies such as lakes is generally precluded, unless some permanent source of water is involved. The Nile, for example, can cross the eastern part of the Sahara despite high evaporation rates (approximately 3,000 millimetres [120 inches] per year at ʿAṭbarah, The Sudan) because its source is in the equatorial highlands and because groundwater contributes to its flow over part of its course. This is flowing water, however; an equal volume of standing water whose only input was derived from precipitation would eventually dry up.

Evaporation is greater over water surfaces than over bare soil surfaces, but it occurs over the latter as well and has much to do with depleting soil moisture and inhibiting plant growth. The occurrence of hot dry winds will tend to increase evaporation rates, and for these and other reasons—including difficulty of accurate measurement—existing evaporation data are in some instances debatable.

With regard to topographic and geologic factors that influence desert hydrology, these can best be considered together. Topography may on occasion promote increased aridity by the rain-shadow effect previously mentioned, but it also may promote rainfall where highlands occur within arid regions. The Tibesti massif (Figure 30), within the central Sahara, for example, attains a maximum elevation of 3,415 metres (11,204 feet); this is sufficient to render it a semiarid "island" in an extremely arid area by reason of the increase of precipitation with increasing elevation.

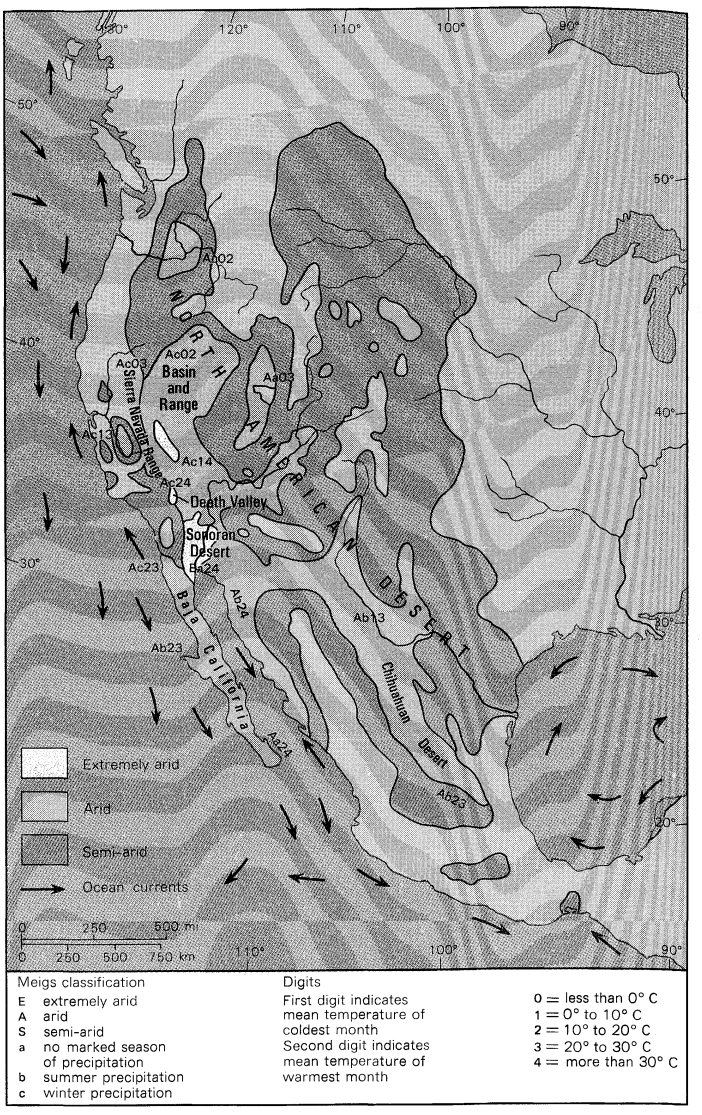
Geologic factors such as rock type and structure, the character of the soil, and the porosity and permeability of surface materials in general also can affect the hydrology in two ways. If water will readily seep to the subsurface because of the presence of rock fissures and fractures or of highly permeable materials, then the runoff from a given amount of precipitation will be less than normal because of water losses. On the other hand, this same set of conditions may promote recharge (the addition of water) of the groundwater reservoir in the area. Where geologic factors inhibit the downward percolation of water, runoff following a storm may be enhanced, but, ultimately, greater evaporation may result.

Geologic factors

Surface water. Although hydraulic and hydrologic principles are unvarying in all environments, some substantial differences exist between the characteristics of streamflow in arid regions and in more humid areas. The drainage networks that exist in the latter, for example, generally reflect the nature of runoff conditions today, whereas in desert areas there are many drainage systems that reflect establishment during Pleistocene time (1,600,000 to 10,000 years ago), when moister conditions prevailed and streamflow was more abundant.

Streamflow in humid and semihumid areas has been studied in detail, and it has been established that the channels of all rivers are in adjustment to a number of hydraulic parameters—*i.e.*, the width, depth, and velocity of flow and the sediment discharge are related to the bank-full water discharge, and this water discharge governs the nature of the channel. The frequency of occurrence of the bank-full discharge is generally about once every 1.5 to 2.5 years, and overbank floods of great magnitude tend to occur every 50 to 100 years. Rivers in these regions exhibit far more days of low flow than of high flow per

Streamflow in humid areas



Meigs classification		Digits	
E	extremely arid	First digit indicates mean temperature of coldest month	0 = less than 0° C 1 = 0° to 10° C 2 = 10° to 20° C 3 = 20° to 30° C 4 = more than 30° C
A	arid		
S	semi-arid		
a	no marked season of precipitation	Second digit indicates mean temperature of warmest month	
b	summer precipitation		
c	winter precipitation		

Figure 32: Deserts of North America.

Adapted from UNESCO, Reviews of Research on Arid Zone Hydrology, "World Distribution of Arid and Semi-arid Homoclimates" (1953)

year; they usually transport the bulk of their suspended sediment load during high-flow stages, but the many low flows during a year serve to transport much of the total load (the suspended load, dissolved load, and bed load). Finally, flow is perennial, with rivers linking up to ultimately reach the sea.

In desert areas, streamflow is extremely sporadic. An effective precipitation sufficient to overcome both evaporation and seepage losses into the channel bed and the surrounding dry terrain of the drainage basin involved is required for streamflow to occur. This usually involves a storm yielding relatively intense precipitation (*e.g.*, 36 millimetres [1.4 inches] of rain were recorded in 40 minutes during a storm at Tamanrasset, Alg., in the central Sahara). Low flows tend to be few by reason of the seepage and evaporation losses, and, when flow does occur in arid areas, it will likely be of high stage, violent, and highly charged with sediment that is derived from the abundance of loose sand and gravel that blankets most areas. Discharge of any magnitude, however, and particularly bankfull discharge, may have a frequency of occurrence that is as small as once in 100 or more years, and the recurrence interval of major overbank floods may be measured in thousands of years in extremely arid areas. Streamflow is thus characterized as ephemeral rather than perennial, and because drainage generally fails to reach the sea, save for rivers of the first rank such as the Nile or those with relatively short distances to traverse such as the Kuiseb in the Namib Desert, it is termed interior drainage (Figure 35).

Where surface flows are ponded or terminate in low depressions, marshy areas or lakes may result. If there is sufficient precipitation over the site of termination or if there is sustenance by groundwater flow, then a relatively freshwater lake may result. Most desert lakes are without outlet, however, and thus become highly saline, because each increment of inflow transports additional salts in solution and the constant evaporation of water from the lake causes salt concentrations. Salinity values as high as 200 to 300 parts per thousand are not uncommon.

Subsurface water. Subsurface water is present in each of the desert areas, but lack of exploratory drilling precludes definitive statements on the total amount of water involved in most instances. Groundwater is easily the principal water resource of the arid regions. Water seeps into the ground and flows downward under the influence of gravity—always assuming, of course, that the materials through which it migrates possess the requisite porosity and permeability. Some of this water migration occurs in the deserts proper; surface water losses caused by downward percolation into the sands and gravels of stream beds have been previously mentioned. Indeed, there is always some close relation between groundwater and surface water wherever stream channels occur, but much of the groundwater also is related to the vast areas over which rain may fall and ultimately reach the saturated zone in the subsurface.

Soil moisture

In the near-surface zone, the water present, principally derived from rainfall, is termed soil moisture. Some of the moisture is lost because capillary action brings it back to the surface, where it may form dew and is in general subject to evaporation. There is a second zone extending to perhaps 100 metres (330 feet) beneath the zone of soil moisture in many areas. This is a relatively dry zone, but it may contain lenses of water-bearing sediments surrounded by more impermeable beds or layers. Such groundwater is termed perched in reference to its mode of occurrence. Beneath this zone the groundwater proper occurs; the surface of this saturated zone is called the water table.

In many instances the general subsurface structure, or configuration of strata, in desert areas is that of a basin. When the principal water-bearing strata, or aquifers, are upturned and are exposed at the surface in regions of high rainfall, much of the water input to the groundwater system will be so derived—rather than from seepage due to rainfall over the entire basin region. In the Great Artesian Basin, in Australia, for example, the total area is about 1,760,000 square kilometres (680,000 square miles), but the principal water intake is on the eastern edge of the basin, over an area of about 100,000 square kilometres (38,610 square miles), in which precipitation is 625 millimetres (25 inches) per year. The water flows downward through upturned sandstone aquifers and thence into the centre of the basin, which attains depths of as much as 2,100 metres (6,900 feet). This circumstance of distant intake for groundwater, wherever it occurs, is responsible for the age of the water that has been determined in North America, in the Libyan Desert of the eastern Sahara, and elsewhere. All absolute age measurements on groundwaters in desert areas indicate a range in age of between 20,000 and 35,000 years. This means that waters tapped by wells in Egypt, Libya, and adjacent regions of the Sahara entered the groundwater system from Pleistocene rainfall in highland areas to the south. The groundwater is often called fossil water for this reason.

Age of groundwaters

If the principal water intake is at the upturned edge of a basin structure and if the water-bearing strata crop out at the surface (or occur near the ground surface) farther out in the basin, then there will be a pressure difference caused by the difference in elevation of the water within the aquifer. When this condition exists, the flow of water is termed artesian, and artesian water is the source for most springs, oases, and near-surface water wells in desert regions. The other principal mode of occurrence of these lifesaving caravan stops is also related to groundwater flow. Instead of artesian conditions, the termination of a subsurface aquifer within permeable sediments or the existence of fractures that penetrate the aquifer may be responsible. (L.K.L./Ed.)

Climatic patterns and the biota. Biological activity is inevitably limited when water is scarce or lacking. Above all, it is the green plant that is so limited. Few plants can continue photosynthesis for long in the absence of water; and in the absence of plant growth other living forms have no opportunity to develop. Animals and microorganisms can exist only if energy sources from green plants are available. They are further limited, though, in the absence of water, for most animals and all microorganisms can be active only while moisture is adequate.

In the cold deserts in the interior of continental masses, drought is severe for at least a part of the year, and the daily and annual range of temperature may be great. Moreover, low winter temperatures have a more severe impact on the biota than they might in a region with greater precipitation, because the protection by snow cover is less.

In the tropical deserts cold is no problem, but the clear skies result in a high radiation load on organisms and in high daytime temperatures, often with large fluctuations from night to day. These factors, too, create more intense stresses on organisms than occur in most environments.

Other effects on the biota of the desert environment arise from the abrasive effects of windblown sand and from the high surface salinity in areas where temporary waters accumulate and evaporate or where evaporation leads to upward movement of salts through the soil profile.

Limited leaching and the small accumulation of organic matter in deserts result in desert soils remaining much closer to the types found in pioneer habitats than do soils in more moist environments. Desert soils seem particularly subject to the formation of impervious surface layers—either by physical processes or through the surface development of algae and lichens—so that the effective drought is accentuated by excessive runoff.

Adapted from UNESCO, Reviews of Research on Arid Zone Hydrology, "World Distribution of Arid and Semi-arid Homoclimates" (1953)

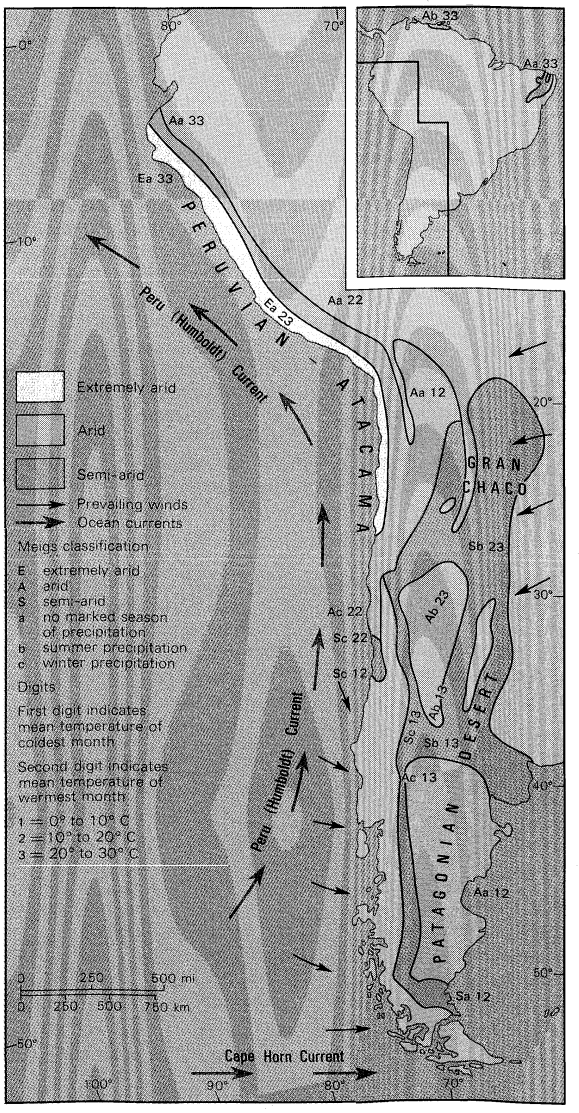


Figure 33: Deserts of South America.

The pioneering conditions that organisms encounter in the desert are permanent, unlike those of other bare and inhospitable habitats, such as coastal sand dunes, rocks, and denuded areas left by human activity. In these climatically more favoured situations, the pioneering organisms initiate changes in the parent soil material that render it more hospitable to newcomers; and a succession of different living communities, each more complex than the last, can take place, leading to a climax community. In the deserts, harsh conditions are not ameliorated by biological activity, and the pioneers themselves constitute the climax community.

Pioneering organisms form the climax community

There are limited desert areas—notably along the coast of South America but also in Baja California, Mauritania, and Namibia—where the extremely low rainfall is supplemented by dew and the aridity is ameliorated by frequent mist and cloud. The relatively low evaporation rates make conditions in such areas much more moist than a simple examination of their precipitation records would suggest.

Spatial dispersion of biota. All of the organisms in a desert do not live under the same environmental conditions. Although the variety of habitats is less than in many wetter environments, it is still great enough to provide effective living conditions for a wide diversity of organisms. Moreover, many organisms show structural and behavioral responses to the environment that tend to neutralize at least some of its extreme features.

It is well known, for instance, that perennial plants in the desert tend to be widely spaced. But, although there are large unoccupied areas between the aerial parts of the plants, excavation may reveal that the root systems are in contact and they exploit the soil so effectively that establishment of new individuals becomes difficult. When rain penetrates the soil, it is rapidly absorbed by the continuous net of roots, and little penetrates to the subsoil. If situations differing in rainfall but otherwise similar are compared, it is found that the plants are more widely spaced as the rainfall decreases. This implies that the amount of plant material remains roughly proportional to the rainfall and that the individual plant is receiving about the same amount of rainfall, thanks to its inhibitory effects on the establishment of competitors.

Perennial-plant spacing

A similar response leads to the establishment of denser vegetation around the watercourses, even when they run infrequently, for there the longer periods of availability of subsoil water permit more plants to have similar productivity without more water per individual.

Different plant species in the desert under the same climatic conditions may differ markedly in their local environment, thanks to differences in soil conditions and topography. Even in the tropics there may be a marked difference in growth conditions between the steeper slopes that face north and those that face south. Desert soils, though with little organic matter and less well-developed profiles than are usual in more humid conditions, differ sufficiently in their mechanical and chemical properties from place to place to support different types of vegetation. And, where stones and boulders are scattered over the soil surface, the microclimates they generate—and the local increase in rainfall penetrating the soil around their edges—permit even some moisture-loving plants such as mosses to develop. Oases similarly are places where a locally dense vegetation corresponds with a locally available water supply.

For animals, the variety of habitats is still greater, for they are generated by the vegetation itself, and since animals are not rooted and dependent on light as green plants are—facts that oblige plants to be more or less exposed to the prevailing atmospheric conditions, however harsh they may be—animals are free to move to more comfortable locations if they desire, even underground. Each species of plant modifies the conditions for the animals around it and may generate living conditions that permit particular animal species to exist, whether providing food,

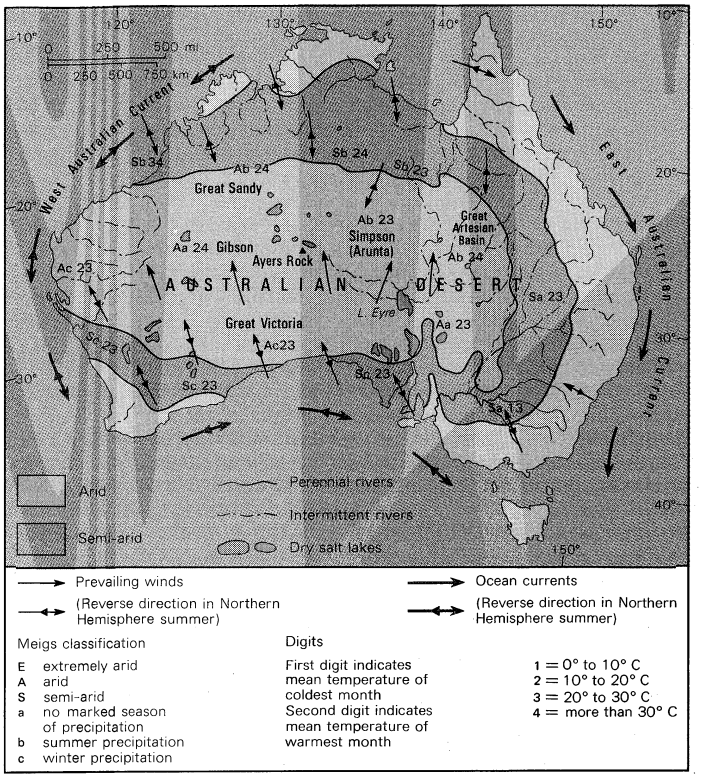


Figure 34: Deserts of Australia.

Adapted from UNESCO, Reviews of Research on Arid Zone Hydrology, "World Distribution of Arid and Semi-arid Homoclimates" (1953)

shelter (shade or protection from predators), or layers of litter in which they can live. Movement between different habitats further varies the living range of animals, as when bats (order Chiroptera) roost in caves during the day and emerge in the evening to feed or when the subterranean larvae of darkling beetles (family Tenebrionidae), feeding on living and dead roots, emerge as adults into the harsh world of the desert soil surfaces to begin their reproductive activities.

Movement between habitats

Like plants, some animals may show a wider dispersion in deserts than elsewhere, thus giving each individual a broader area of ground as "its own." Kangaroo rats (*Dipodomys*), for instance, are reported to manifest more aggressive behaviour in the deserts than in other habitats, resulting in larger and more effectively defended territories for each individual animal.

TYPES OF DESERT ECOSYSTEMS

Cold deserts. These occur mainly in the interior of Asia and in the intermountain zone of North America. Winter temperatures are well below freezing, and the ground may be snow-covered for considerable periods. Summer temperatures are moderately high (up to 40° C [104° F]), though the temperature extremes of regions farther south near the tropics are avoided since the sun is not so high in the sky. Precipitation often is between 200 and 300 millimetres (8 and 12 inches), and a good part of it commonly falls as snow.

The dominant vegetation is of shrubs, with a discontinuous canopy, but not as widely spaced as in the drier deserts nearer the equator. In the spring, as the snow melts, there is a flush of growth, and many short-lived annual plants occupy the soil surface. Among the shrubs, sagebrushes are prominent, as are a number of species of the goosefoot family (Chenopodiaceae), particularly in lower-lying areas, where there is some accumulation of salt.

Shrub deserts and thorn scrub. These ecosystem types occupy a large part of the area of the tropical deserts. The peripheral zone of the Sahara (the Sahel) is dominated by the acacias, many of them thorny. The Mojave and Chihuahuan deserts in North America are dominated by a range of shrubs, of which the creosote bush is probably the commonest. The same is true in similar deserts of South America. Much of the Australian dry country is dominated by many sparse, small acacia trees, while other acacias, *Eremophila,* sennas, and members of the goosefoot family form a denser understory. In other peripheral parts of the Australian arid zone, the low-growing mallees (eucalyptus) take the place of the treelike acacias.

Under the shrubs there is often a partial ground cover of tufted perennial grasses, while the interspaces are filled with short-lived, small flowering plants following rain.

Larger plant-eating animals are not uncommon in this type of country, entering from the surrounding areas of somewhat higher rainfall and more grass during periods when forage is more abundant and retreating when times are harder. Some, such as the desert white-tailed deer (*Odocoileus virginianus*) and collared peccary (*Dicotyles tajacu*) in America and the kangaroo (family Macropodidae) in Australia, make their permanent home in such conditions.

The deserts of cactus and cactuslike plants

Stem-succulent deserts. These striking formations are of limited extent and consist of only the Sonoran Desert (Gran Desierto) of Arizona and northwestern Mexico, the area between Peru and Argentina, and the Namib Desert in Namibia. The special appearance of stem-succulent deserts rests on the geographic availability of certain plant groups absent from many of the deserts—the cactus family (Cactaceae) of the Americas and the arborescent members of the spurge family (Euphorbiaceae) mainly found in Africa. These plants have adopted a special growth habit that provides one alternative solution to the problem of survival in drought. Elsewhere, their role is taken over by shrubs.

Herbaceous deserts. There are desert areas where shrubs or stem succulents are few or absent and where the bulk of vegetative matter is in the form of perennial herbs. These are nonwoody plants whose tops die back each year but whose roots survive to produce new tops year after year. The Nullarbor Plain in southern Australia is of this type; the soil is thin over a continuous limestone substrate, permitting water to penetrate freely, and these harsh conditions are doubtless responsible for the lack of shrubs. Even the perennial species of the goosefoot family form only a sparse ground cover. In southern Africa, there are areas where the dominant vegetation is formed of low-growing succulents.

Salt deserts. Within the deserts, there are many areas where the soil contains a high concentration of salts; most commonly sodium chloride is predominant, but in some places sulfates or carbonates may be more important, and other substances may also, in part, take the place of sodium. These areas are usually in basins of internal drainage, and they may surround highly saline lakes, such as the Dead Sea in Jordan and Israel and the Great Salt Lake in the western United States, or large salt pans flooded only after exceptional rains, such as Lake Eyre in South Australia. Few organisms can grow in a concentrated salt solution, and in consequence the area of a salt pan has a simple community and may appear bare. As

L.K. Lustig—EB Inc

Figure 35: Red sand dunes of the southern Namib Desert separated from the northern gravel plain (foreground) by the Kuiseb River. The channel is marked by a row of vegetation, visible at the base of the dune area.

the salinity decreases, either around margins of such an area or through seasons of high rainfall, a limited range of higher plants can establish themselves—mainly shrubs or subshrubs of the goosefoot family. It is noteworthy that these are closely related to—sometimes, indeed, the same as—the species found in the coastal salt marshes. There, too, a high and variable salinity prevails, and near the spring high-tide mark, in some seasons, high salinity and low moisture content persist for a week or more. In the salt deserts, on the other hand, these conditions may persist for months. Whereas in salt marshes the increase in soil level with buildup of organic matter may eventually take the area above tide level, so that leaching can take place and succession proceed, the vegetation in salt deserts is a soil-dependent climax and will not develop further.

Vegetation on dunes and along ephemeral streams

Sand dunes. Where sand dunes occur in deserts, their consolidation or stabilization by vegetation is much more difficult than along seacoasts. In both situations consolidation depends on plant species with wide-spreading root or rhizome (underground rootlike stems) systems. At the coast, this normally permits invasion and establishment of other species, and thus begins a succession. In the desert, the pioneers are the climax, and the barely consolidated dunes are likely to become mobile again if any disturbance breaks the fragile cover and permits wind blowouts. The free drainage of the sands implies that moisture conditions, as well as substrate mobility, are particularly harmful to plant life; and sand dunes in an arid landscape are likely to be among the most persistently bare areas on the Earth's surface.

Wadis and watercourses. Along the ephemeral wadis and watercourses of the desert, the occasional violence of the waters discourages perennial plants; but along the sides, where flow is less frequent and less violent, a special vegetation of deep-rooted moisture-requiring shrubs and trees develops. These benefit not only from water that actually flows down the channel but also from its much deeper penetration into the soil—often quite permeable in these situations—and consequent longer availability. Some of these plants, called phreatophytes, may be able to tap water many metres below the surface and thus survive a drought during which most of the vegetation around dies. A special population of birds and other animals is associated with the vegetation around the watercourses.

Aquatic ecosystems in the deserts. A torrent that flows for two or three days and is then dry clearly does not permit the development of an aquatic ecosystem. But some waters in the deserts are longer-lived, and they have their own special biota. Even in the temporary watercourses, there may be pools that remain filled for weeks, and in these an ecosystem with algae, insect larvae, and amphibians will develop. The same applies to the temporary lakes known as playa lakes lying in hollows and filled for a month or so following rain. If these temporary lakes are saline, the ecosystem may be much simpler, for the salt concentration may considerably exceed that of seawater and the species that can endure it are few.

Virtually the only permanent waters of low salinity originating in the deserts are springs, renewed by flow from underground. These may develop complete ecosystems of similar structure to aqueous systems elsewhere, though generally simpler—due, doubtless, to isolation. These include rooted aquatic and riparian higher plants, algae, crustaceans, insects, amphibians, and fish.

THE DESERT BIOTA AND THEIR ADAPTATIONS

Protists. Soil microorganisms occur in desert soils in about the same degree of diversity as in other soils and can survive frequent desiccation and remoistening. In some deserts, there is a noteworthy development on the soil surface of blue-green algae, which are thought to be active in nitrogen fixation (converting atmospheric nitrogen to forms that plants can utilize) and may also modify the infiltration properties of the soil.

Algae, fungi, lichens, and mosses

Plants. A great deal has been written of desert plants and their adaptations to the rigours of desert life. Apart from the blue-green algae mentioned above, other algae are practically confined to the limited water areas of the desert. Parasitic fungi are as abundant as in other environments, though mainly confined to the higher groups not requiring free water for swimming reproductive cells. Saprophytic fungi (*i.e.*, fungi obtaining food from dead organic matter) are not abundant, the limited organic matter in and on the soil providing little in the way of nourishment, and active growth being possible only during the short periods when the soil is moist.

Table 11: Desert Climatic Data for Representative Stations of the World

station	number of years on record	mean annual precipitation (mm)	maximum recorded precipitation for a 24-hour period (mm)	relative humidity (percent)	mean annual temperature (°C)	maximum recorded temperature (°C)	minimum recorded temperature (°C)	average number of days of sunshine per year	average wind velocity (km/hr)	maximum recorded wind velocity (km/hr)
North America										
Phoenix, Ariz. (Sonoran Desert)	40	178	108	33	20	47.7	−8.3	209	8	96
Las Vegas, Nev. (Mojave Desert)	22	102	66	18		47	−13.6	220	14.4	86*
Monterrey, Mex. (Chihuahuan Desert)	10	711	...	57	22	42	−9	...	...	...
South America										
Los Canchones, Chile (Atacama Desert)	40	0	...	...	17	36	0.6	264	...	...
Lima, Peru (Peruvian desert)	10	25.4	...	72	19	34	9	...	...	...
Africa										
Algiers, Alg. (Sahara)	10	762	...	61	18	42	0	...	...	...
Cairo, Egypt (Sahara)	10	25.4	...	29	20	47	1.1	...	...	...
Tripoli, Libya (Libyan Desert, Sahara)	25	381	...	62	19	45	0.6	...	...	...
Asia										
Baghdad, Iraq (Syrian Desert)	25	152	...	29	...	49	−7.7	...	...	...
Lu-k'o-ch'in (Takla Makan Desert)	2	...	...	...	13	48	−21	...	...	...
Ulaanbaatar, Mong.	1	208	...	54†	−3	16	−26	...	...	...
Australia										
Alice Springs, Australia (Australian desert)	90	252	147	...	21	47	−7.2	...	6.6	...
Marble Bar, Australia (Australian desert)	70	340	305	...	27.7	49.2	1.1	...	...	...
Antarctica										
McMurdo	1	101	...	...	−17	−3	−29	...	...	...

*Six-year record. †Fifty-four year record.

Lichens are numerous and often prominent—particularly the crustose types. They often form an almost continuous covering to the soil surface and are then able to take advantage of any wetting of the soil for short-lived periods of photosynthesis.

Mosses and liverworts are few and are often confined to special habitats more moist than the deserts as a whole. Ferns are not infrequently found in rock crevices and similar habitats. Club mosses also occur, some showing the remarkable adaptation of being able to dry out and withstand desiccation and lack of active growth in the form of a ball, which may be blown about freely.

Some gymnosperms (nonflowering seed plants) are noteworthy inhabitants of the deserts. Conifers and cycads are few in desert areas, but several ephedras are important desert shrubs—a role to which their near leaflessness suits them; on the other hand, the remarkable Tumboa (*Welwitschia*), with its large, split leaves, is also a desert plant, confined to the Namib in Namibia.

Among the angiosperms (flowering seed plants), there are some families that are especially characteristic of the deserts. The goosefoot family is such a characteristic inhabitant of salt deserts (and other saline habitats) throughout the world. Some members of the caltrop family (Zygophyllaceae) are also normally desert plants. In the Americas, species of the cactus family are characteristic of the deserts, and the small candlewood family (Fouquieriaceae) are found there only. Another group playing an important part in many of the world's deserts, though not confined to them, is the family Mimosaceae, and particularly the genus *Acacia*. Species of this genus characterize the peripheral zones of the Sahara, are dominant in many parts of the Australian desert, and are far from negligible in the American and Asian deserts.

Plants in the deserts have to surmount a difficulty that the photosynthetic mode of nutrition imposes on them. In order to produce food by means of photosynthesis, carbon dioxide from the atmosphere must be taken into the plant. This unavoidably leads to the loss of water from the plant at the same time, since the same small openings in the plant surface (called stomates) serve for both purposes. Various solutions to this problem are to be seen in desert plants.

One way to avoid the adverse effects of limited water availability is to live through dry periods in a completely inactive form—as a seed. This is the solution to the problem adopted by the ephemerals, or short-lived annual plants. These plants shorten the duration of their life cycle so that seed germination follows promptly after an adequate storm and seed formation is already in progress by the time soil moisture is reduced to the wilting point. This means, of course, that the total period of photosynthesis is limited to a few weeks every year—perhaps in some years to none at all. Nevertheless, this group of plants is a successful element in desert vegetation, as is evident in the flowery display often seen over the desert floor a month or so after rain.

Annuals' solution to limited water

Another solution, rather similar, is for the plant to restrict itself during the dry season to inactive tissues that have limited gas exchange and to form tissues capable of free gaseous diffusion only during periods when soil moisture is available. This mode of growth is called the deciduous habit, and it is expressed by plants everywhere that drop their leaves during winter or, in the case of deserts, during dry periods. A good many species of desert plants have short-lived leaves and are bare for much of the year. The spectacular ocotillo (*Fouquieria splendens*) of the Sonoran Desert of southwestern North America is one example.

Another solution is to separate temporally (*i.e.*, in time) the gaseous-diffusion process from that of energy utilization. Normally, the photosynthetic surfaces must be open to carbon dioxide diffusion during daylight periods, in order that radiant energy may be converted to chemical energy in the photosynthetic process, and this incurs the penalty of water-vapour loss at a time when the diffusion gradient for it is particularly high because of high daytime temperatures and dry air and that for carbon dioxide is no more favourable than at any other time. If carbon dioxide can be taken up at night instead, the concomitant loss of water can be much reduced. Many desert plants have this ability, carbon dioxide being stored in the form of carboxylic acids at night, which are converted to carbohydrates when radiant energy becomes available in the day. The stomates in these plants open by night to allow entry of carbon dioxide into the intercellular spaces and close by day to restrict loss of water vapour while photosynthetic reduction of the "stored" carbon dioxide proceeds. This so-called crassulacean metabolism is not confined to the stonecrop family (Crassulaceae) but is found in a good many other plants of the deserts.

Another mechanism, like the deciduous habit, restricts gas exchange when drought conditions prevail not by loss of leaves (a wasteful process) but by direct restriction of gaseous diffusion. The stomates, through which most gas exchange occurs, can generally open or close according to prevailing environmental conditions; adaptations may increase their responsiveness to water stress and at the same time reduce the amount of gaseous diffusion in other ways. The means used include increased thickness of the cuticle (outer layer of stem and leaves), development of wax layers, rolling of leaves in drought, and so forth—a series of adaptations often described under the heads of xeromorphy or xerophily.

Another mechanism, often combined with restriction of gaseous diffusion by stomates during periods of drought and with crassulacean metabolism, is water storage in succulent tissues. This has developed in rather few taxonomic groups—the families of carpetweed (Aizoaceae), stonecrop, cactus, some milkweeds (Asclepiadaceae—especially the subfamily Stapeliae), some goosefoots, some lilies (family Liliaceae), the arborescent spurges, and a few others.

In some cases, plants can grow in deserts without suffering water shortage through having an extremely deep root system penetrating to subsoil water. A high ratio of root to shoot tissue is indeed a common feature of desert plants, whether the root system is deep or spreads widely.

A hazard of the deserts, when water loss from aerial parts is restricted, is the high temperature in these organs that may result from intense sunlight when cooling by evaporation is limited. Resistance to high tissue temperatures may be part of the adaptational equipment of these plants.

Adaptations of plants to high temperatures

Temperatures on the soil surface in deserts may be very high (70° C, or 158° F), and plants growing there must be prepared to endure them. Lichens and blue-green algae, which compose the bulk of the soil-surface flora, are usually dry at times when the soil-surface temperature is highest, in which state they are highly resistant to extreme heat. Perennial plants usually provide shade to the soil immediately around their bases, so that high temperatures are eliminated there. For small annuals, on the other hand, high soil temperatures around their stem bases when soil moisture has been depleted may cause some death of tissues and accelerate flower and seed maturation.

Another factor to which adaptation is required in some desert situations is the high salinity of the soil solution. Usually plants growing in such situations have difficulty taking water into their roots from such solutions. Many plants have adapted to these conditions with a highly concentrated sap, enabling the root to take up water even from saline solutions.

Invertebrate animals. Many groups of invertebrates require a marine environment and are consequently absent from the deserts; other groups occur in desert waters but not on land. The flatworms (phylum Platyhelminthes) occur as parasites inside the bodies of animals, as elsewhere. The various groups of nematodes (phylum Aschelminthes)—animal parasites, plant parasites, and those living free in the soil—are found in the deserts as in other habitats, though activity in the soil is restricted by their need for free water. The segmented worms (phylum Annelida) are represented by a few earthworms, occurring only in favoured situations. There are a fair number of snails (class Gastropoda) found in deserts, some species of which are remarkably resistant to arid conditions.

It is, however, of the insects and spiders and their relatives (phylum Arthropoda) that one thinks when desert invertebrates are mentioned. The crustaceans of that phylum are represented by the brine shrimps that live in

concentrated salt solutions and by the terrestrial isopods (wood lice, or sow bugs) that occur fairly widely. Myriapods are reasonably numerous in the desert, both the vegetarian millipedes and the carnivorous centipedes. All the groups of arachnids (spiders, scorpions, etc.) except the primitive marine horseshoe crab (*Limulus*) occur in the deserts, and some are more at home than in many other environments. Scorpions are numerous, as are ground-dwelling spiders (web-building spiders are not particularly common), and there are plenty of mites, including those inhabiting the soil.

Insects are abundant, though forms with aquatic larvae such as the dragonflies (order Odonata) and mayflies (order Ephemeroptera) are found only around the waters; termites, beetles, butterflies, moths, flies, ants, bees, and wasps are all universally present in deserts, as are the grasshoppers and true bugs.

Adaptations to desert conditions by insects and their relatives

For many insects and arachnids, the problems of life in the deserts are easier than for most vertebrates. Like many desert plants, they have a waterproof cuticle, or skin, to retain water; and the frequently short life cycle, often combined with metamorphosis (changes during the life cycle from egg to larva, etc.), enables vulnerable stages to be timed for the less adverse seasons. By persisting through unfavourable periods in an inactive stage (egg or pupa), the animal can avoid many of the difficulties of desert life, much as does the ephemeral plant that persists as a seed or the deciduous plant when reduced to inactive woody tissue during dry periods. Another means of avoiding environmental difficulties is to burrow beneath the soil. Many arthropods burrow for at least a part of their life cycle; and there is an abundant fauna that spends its whole life beneath the soil. This is true of nematodes and the few desert earthworms, of course, as well as large numbers of mites and almost all the springtails (order Collembola). These soil arthropods form part of a trophic cycle (food chain) within the soil mass, in which material originally derived from plant roots is eaten by herbivores and they, in turn, by several carnivore levels. This material eventually returns to detritus, or the decaying organic state, to be acted upon by microorganisms, which are again taken into the animal sequence.

Some of the subterranean arthropods are larval stages of adults that spend their life on the soil surface. There are important groups, however, whose adult life is divided between the upper and lower worlds. The most important are the ants (order Hymenoptera) and the termites (order Isoptera). Not all ants have subterranean nests, but, in deserts, by far the majority do so; and all termites avoid the light, building covered runways when they come above ground or developing channels through woody stems. Thus the ants, and even more the termites, avoid the risk of exposing the more vulnerable immature forms to the high radiation, temperature, and evaporation rate of the open desert; and they limit the time of exposure of the more resistant adults to what is required for food gathering (in the case of ants) and defense.

Vertebrate animals. The desert waters, when permanent and not too saline, contain fish, and in some cases the species are endemic (peculiar to the locality), because of the isolation of some of these ecosystems. They are generally closely related to species in nearby nondesert areas. Little is known of their special adaptations, though they must often tolerate a higher and more variable salinity than do their relatives elsewhere.

Frogs and toads are far from uncommon in the deserts, despite their need of an aqueous environment for reproduction. Consequently, desert amphibians need to be opportunistic, responding quickly to rain that leaves standing water and then passing through the stages of egg laying and larval life before this water dries up. The selection pressure for rapid larval development must be intense, for those larvae still dependent on gill breathing when the pools dry will die. The adults, with a moist and unprotected skin, are also vulnerable to high evaporation rate and spend most of the time between rains inactive in a protected spot, often burrowing deep into the soil.

Reptiles are perhaps the most characteristic group of the desert animals. Lizards and snakes are numerous, tortoises occur, and crocodiles are common in some tropical desert waters. The majority of the lizards are insectivorous, though some are herbivores. The snakes mostly feed on other vertebrates, chiefly small mammals and birds' eggs and nestlings. Some of them have refuges in holes in the soil—perhaps more from predators than from the climate. But the majority spend most of their life exposed to high temperatures and evaporation on the surface of the soil. Their scaly skin is indeed resistant to water loss, but the radiation load they suffer is often considerable. They have no means of temperature regulation except seeking shade, and their tissues must be resistant to temperatures well above the normal limits for warm-blooded animals. Though some show bright warning coloration, many are camouflaged in the colours of the soil surfaces on which they live.

Desert birds

Birds are as numerous in deserts as food availability permits but often spend only part of their lives in this environment. Apart from seasonal migration, they may move between roosting or nesting sites and feeding areas. The lack of trees or tall shrubs to provide nesting positions excludes many species (apart from foraging excursions), and many desert birds nest on the ground. The eggs are in some cases covered with mounds of vegetable material, which give partial protection and provide a more uniform temperature for incubation. Running birds are numerous in the deserts and include some of the largest—the ostrich (*Struthio camelus*) in Namibia (and previously in North Africa and the eastern Mediterranean countries); and the emu (*Dromaius*) in Australia, both of which are herbivores. Some birds are present that are prominent predators and carrion eaters, such as the roadrunner (*Geococcyx*) and magpie (family Corvidae) of the North American deserts. Raptors (hawks, eagles, etc.) are important and have in the deserts excellent visibility for attacking prey from the air. They may nest in rocky parts of the deserts themselves, but their range of movement is such that they can often operate from areas in adjacent but different vegetation.

Some bird species show notable adaptation to the desert environment in their reproductive habits. Gambel quail (*Lophortyx gambelii*), for instance, do not reproduce in years with below-normal rainfall, when food for the young would be scanty; ducks (family Anatidae) in central Australia have lost the annual reproductive rhythm and will breed at any time when rain falls.

There are many mammals in the deserts, though the modifications required in behaviour or physiology (or in both of these) are considerably greater than for the reptiles. This is largely because they are warm-blooded and because they excrete urea rather than uric acid, which imposes extra demands on body water.

Most mammalian orders, other than those that are purely aquatic, are found in the deserts. Primates (apes, monkeys, etc.), mainly tree dwellers, are naturally not prominent, although several species of baboon are found in the African and Arabian deserts. Edentates (anteaters) and insectivores are few, but bats are often numerous, roosting in the caves and rock crevices that are commonly found in desert regions. The rodents (rats, mice, etc.) and lagomorphs (rabbits, hares, etc.) form the largest group of desert-inhabiting mammals. Some of them are remarkably successful there—the European rabbit, for instance, introduced in the 19th century. The native rodents of the deserts often show marked physiological adaptation to the desert, water requirements being reduced to a minimum. Some, in fact, can subsist on the metabolic water produced by respiration of their foodstuffs. Many desert mammals avoid high radiation loads and decrease their water losses by spending the midday hours in the shade of bushes and feeding in the early morning, late evening, or at night. Many seek the shelter of burrows in the earth, with their much more moist and consistent conditions, at least during the daytime hours, and their life above ground may be limited to nocturnal foraging.

Herbivorous mammals

Herbivores in the desert can benefit from the ability to move rapidly, from one feeding place to another or between feeding and drinking places, so it is not surprising that a fair number of ungulates (hoofed animals) make the desert their home. Many are nomadic and range over large

areas, thus being able to benefit from vegetation growth following local storms, without the risk of starvation imposed by eating out the forage within a limited territory. There are numerous antelopes, goats, and sheep in the deserts of all continents but Australia. Some of the wild asses are at home in the deserts of western Asia and North Africa, and escaped domestic equines (burros in North America, brumbies in Australia) have taken successfully to the deserts and desert margins. Outstandingly successful among ungulates in the deserts, however, are the camels (*Camelus*), whose adaptation to life there is well known. Their broad feet suit them to movement over sandy areas (though such areas are extensive only in limited parts of the desert regions); they have large reserves of fat, to enable them to live through long periods of scanty forage; and they can live for many days without drinking, making up their water deficit by enormous drafts when opportunity presents itself.

Among the carnivores, lions inhabit the desert margins of Africa and Asia, and some smaller cats also hunt in the deserts, in both the Old World and the New. Several of the hyenas are desert animals, and some of the dogs are also prominent desert carnivores, notably the coyote (*Canis latrans*) and kit fox (*Vulpes macrotis*) in North America and the dingo (*Canis dingo*) in Australia.

The marsupials (pouched animals) include a number of species well adapted to life in the Australian desert—many of the kangaroos and wallabies and a variety of smaller herbivorous and omnivorous species.

Man. A wide variety of human races have adapted successfully to life in the deserts. Caucasoid peoples inhabit the deserts of North Africa and western Asia, and the Australian Aborigines, thought to be more closely related to them than to other major racial groups, live in the deserts of Australia. Negroid peoples live on the southern margins of the North African desert and in Namibia. Mongoloid peoples occupy the central Asian deserts and those of North and South America.

Human uses of deserts

As would be expected in an animal of such behavioral diversity as man, his adaptations have been behavioral and cultural rather than physiological. Irrigation civilizations have been developed in many arid regions of the world—in fact, the Nile valley and the valley of the Tigris and Euphrates, which were two of the cradles of civilization, run through deserts. Where irrigation was not possible, low vegetative productivity imposed a nomadic culture on desert-dwelling societies.

Apart from nomadism and irrigation works, man's behavioral adaptations include his use of thick-walled dwellings with few windows; water storage in deep, covered cisterns; clothing, often white, protecting against the sun; and domestication of desert animals, notably the camel.

PRODUCTIVITY OF DESERT ECOSYSTEMS

The natural productivity of the deserts is low compared with that of most other ecosystems. The dependence of all plants on water and of other organisms on plants means that the arid conditions limit productivity where radiation, temperature, and other growth factors would otherwise permit the highest levels of productivity on the Earth. The proportion of solar energy incorporated in organic matter, rarely exceeding 2 percent in natural communities, is in the deserts a small fraction of 1 percent. The bulk, or biomass, of vegetation in the Syrian Desert has been reported as ranging from 470 to 4,800 kilograms per hectare (kg/h), and in the cold deserts of central Asia from 55 to 7,000 kg/h.

In cold deserts in North America dominated by sagebrush, figures between 1,000 and 3,000 kg/h of dry matter per year have been recorded, as compared with 50,000 to 100,000 kg/h for mountain forests in the same latitude. For the Sonoran Desert a figure of 1,400 kg/h per year has been quoted.

While the standing plant biomass may be well over 1,000 kg/h, the animal biomass is far less. Figures on which to base even a rough estimate are scanty, but 100 kg/h would probably be exceptionally high, and 10 kg/h the more usual order of magnitude.

The deserts have been used by food-gathering human societies for many millennia. Provided the population is sparse and nomadic, a reasonably reliable living may be wrung even from these unpromising habitats. Good examples are the Australian Aborigines, who made use of a wide range of fruits and roots and insects, reptiles, and mammals. Other peoples use the deserts by grazing on them flocks of domesticated livestock and constantly moving the flocks to fresh pastures as the limited growth is grazed off. (D.W.G./Ed.)

Jungles and rain forests

In modern usage the term jungle has come to mean tropical forest and connotes luxuriant, tangled, impenetrable vegetation in a hot, steamy environment, teeming with wildlife. The rain forest exhibits many forms but seldom approaches its popular image; in fact, it is darkly shaded, with little ground cover, and is fairly easy to travel in. Where light can penetrate to the ground, however, as along tracks and riverbanks, a dense and tangled mass of vegetation appears, and because most people see the forest from a track or from a boat it is easy to believe in the impenetrable jungle. Tangled and dense bush often invades as a secondary growth after a rain forest has been cleared, and the word jungle is sometimes used in a more technical sense for such regrowth.

Characteristic amount of rainfall

Rain forests, as the name suggests, occur in regions of high rainfall, commonly over 1,800 millimetres (70 inches) per year. It is overall wetness that forests require, however, and no single figure really expresses the most desirable conditions. This is because the effectiveness of precipitation depends on losses from runoff and evapotranspiration (the amount of water that is lost by direct evaporation and by transpiration through plants) that in turn depends on topography, soil, and temperature. Rain forests, therefore, span a wide range of environments, with widely varying geology, water regime, soils, landscapes, and potential use.

Commonly, most rain forest areas are places of relatively low human-population density. The rapid clearing of rain forests in the late 20th century for commercial uses, however, promised an increase in human habitation as well as far-reaching negative consequences from the unchecked destruction of these ecosystems, which contain the vast majority of the world's species and account for a significant portion of oxygen production in the biosphere.

THE ENVIRONMENTAL SETTING

Varieties and characteristics of rain forests. Some authorities divide the tropical rain forest into "equatorial" and "tropical" or "tropical" and "subtropical" forest, the former in each case being the more permanently wet. Here, "equatorial" and "subtropical" are used for these divisions, and "tropical" is a general term for both. The term selva is synonymous with tropical rain forest. The world distribution of jungles and rain forests is shown in Figure 27.

Equatorial rain forests. Equatorial rain forests grow in areas of little varying climate with rain throughout the year, no frost, high average temperatures, and no marked seasonal variations, though occasional dry periods of a month or two can occur. Average daily temperatures in equatorial rain forest regions usually range between a daytime maximum of about 30° C (86° F) and a nighttime minimum of about 20° C (68° F), with a monthly and yearly average of about 27° C (81° F). Wide variations from these temperatures are exceptional, and the average annual range is usually only one or two degrees. The average daily range, normally between five and eight degrees, is greater than the annual range, so it may truly be said that night is the winter of the tropics. Any seasons that can be recognized are dependent on variations in the amount of rainfall rather than in the amount of temperature.

Rainfall in equatorial forests

Annual rainfall generally amounts to between 1,500 and 3,500 millimetres (60 and 140 inches), generally well distributed throughout the year (see Figure 36). There may be a drier period, and some places have two periods of maximum rainfall that roughly coincide with the two periods of vertical noonday sun. Some equatorial rain originates from thermal convection, but most comes from organized,

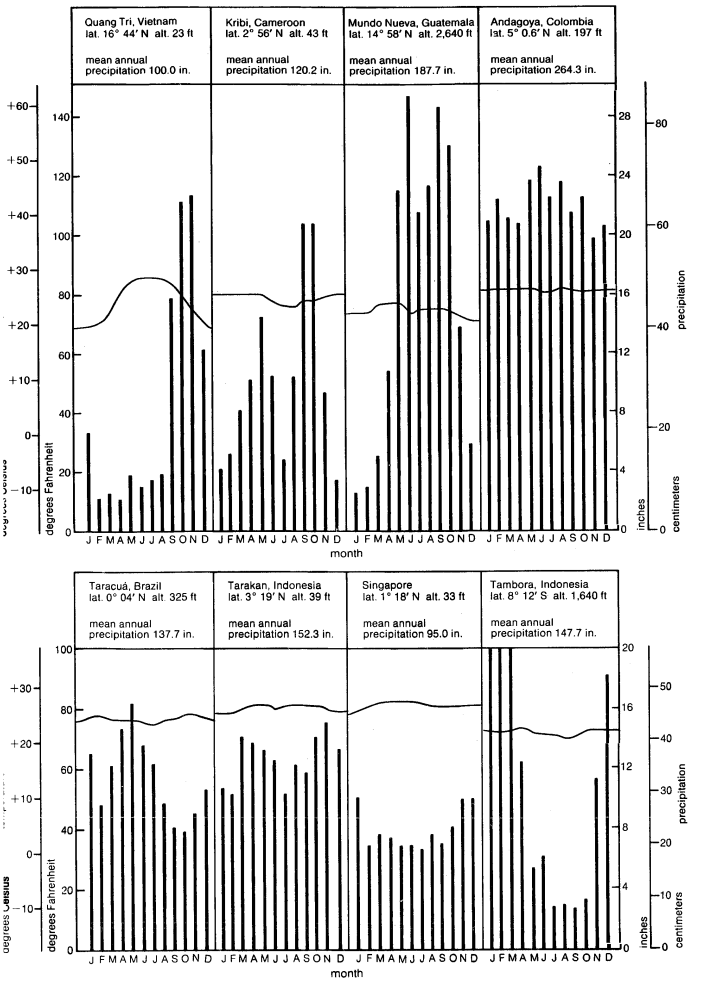


Figure 36: Representative temperatures (lines) and rainfall regimes (bars) in tropical areas.

Reprinted with permission of the Macmillan Company from *Climatology and the World's Climates* by George R. Rumney. Copyright © by George R. Rumney, 1968

migrating atmospheric disturbances. Relative humidity in an equatorial rain forest is high at all times.

There is a large water surplus and thus a large runoff; soils are permanently moist, leaching is intense, and erosion is potentially severe. Chemical reactions affecting soil and vegetation take place rapidly because of the high temperatures and abundant rainfall. Plant growth is continuous, though individual species have their own seasons of leaf shedding. The light intensity on the forest floor is usually less than 1 percent of that just above the forest canopy. The ground is littered with fallen leaves and wood, and low-growing plants are few except for tree seedlings.

Main areas of equatorial forest

The main areas of equatorial forest are the Amazon lowlands, the Congo lowlands (together with a coastal zone extending from Nigeria to Guinea), parts of Indonesia (especially Sumatra), and several Pacific islands.

Rain forest can sometimes extend into areas where rainfall appears to be inadequate if it is augmented by groundwater—forests often extend along valleys surrounded by other vegetation types—or if generally high humidity effectively reduces water losses, as in parts of southern Nigeria.

Equatorial forest grades into subtropical rain forest on windward coasts, into monsoon forests in parts of Southeast Asia and Australia, and into montane forest with increasing altitude.

Subtropical rain forests. Subtropical rain forest is located on the windward coasts, roughly from 10° N and 10° S to the tropics, but occasionally it extends farther. It differs from the equatorial rain forest by having a season of reduced rainfall or even drought, although the wet season is dominant. With distance from the equator, there is greater variation in length of day and greater seasonal variation in temperature; as a result, this type of rain forest is slightly more open and lower and has a smaller number of species and fewer lianas.

Such forest is found in Central America and in the Caribbean—especially on the windward side of islands; in the Western Ghāts of India and coastal areas of Myanmar (Burma), where there is a short dry season; and in Vietnam, the Philippines, parts of the Brazilian coast, and Madagascar. In the Everglades of Florida there are small patches of such jungle, known as hammocks, which contain mahogany, strangler fig, and epiphytes (air plants).

Monsoon forests. Monsoon forest, also known as tropical deciduous forest, occurs in regions with a large total rainfall but a marked dry season. Monsoon climates are characterized by seasonal pressure and wind reversal resulting in heavy rainfall during the high-sun period of onshore winds and little or no rain during the low-sun period of offshore winds. Average temperatures are high, often over 25° C (77° F), but the annual range is greater than in areas of equatorial rain forest. This forest is lower than equatorial forest, trees have more spreading branches, and there is more light penetration. Because of the dry season, a high proportion of the trees are deciduous, and there is a general period of leaf shedding. Teak is a common species, lianas and epiphytes are locally abundant, and thickets of bamboo are common. Undergrowth is likely to be denser in monsoon forest than in equatorial rain forest because of the greater openness and light penetration, and there is more "jungle" formed after clearing than in tropical rain forest.

This sort of forest is found in monsoon Asia, Myanmar (behind the coastal tropical forest already mentioned), Thailand, Kampuchea (Cambodia), Indonesia (especially Java and Celebes), northeastern Australia, and in parts of West Africa and South America bordering the tropical rain forest.

Mangrove forests. Mangrove forest grows along many tropical coasts, especially on low-lying muddy shores. Trees are commonly between 2 and 10 metres (7 and 33 feet) high, but some reach 30 metres (100 feet). The tangle of stilt roots, air roots, and general debris unites to trap mud and thus build up the land along a low coast. Mangrove can tolerate a wide range of salinity and often lines tropical estuaries and deltas, sometimes extending for tens of kilometres inland. Several species may grow in belts parallel to the coast. Mangrove wood is used for fuel in some areas.

Land development

Montane forests. Montane forest occurs in the high altitude regions within or bordering the tropical rain forest and includes the Ruwenzori Range of central Africa, the New Guinea highlands, the Amhara Plateau of Ethiopia, and the Gotel Mountains of Cameroon. With its lower temperatures, this forest has many affinities with the temperate rain forest (see above); but, although altitude brings lower average temperatures, other features, such as length

Eric Lindgren—Ardea Photographics

Figure 37: Tropical rain forest in New Guinea, showing the discontinuous character of the upper tree stratum.

of day, cloudiness, and seasonal variation, are unlike those of the temperate regions, and the two forest types are therefore dissimilar. Tropical forest gives way to submontane forest at an average altitude of 1,000 metres (3,280 feet) and to true montane forest, or "mossy forest," at about 1,500 metres (4,900 feet).

Rock type and landscape. Rain forest is found on many rock types and in a wide range of topographic situations, but a few major varieties of landscape may be distinguished.

High plateau and gentle uplands are found in large areas of Africa and in the highlands of the Guianas, Brazil, and other parts of South America. Broad, shallow valleys are separated by broad, gently convex or flat interstream areas, on which erosion is slight. Weathering to a depth of 100 metres (330 feet) is common, and detailed investigations usually show that the lower limit of weathering is irregular with respect to depth; rapid lateral changes in the depth of weathering occur. Occasional hills of fresh rock rise abruptly from a surrounding plain, to a height of tens or hundreds of metres. Such hills are called inselbergs, and some authorities regard them as typical of the humid tropical environment.

Inselbergs

Where erosion has followed widespread laterite formation (see below *Weathering and soils: Lateritic soils*), flattop hills capped by lateritic ironstone are common. Lateritic ironstone requires seasonal drying out for its formation, however, and so is more frequent in savanna regions and around the edges of the tropical forest than in the pure forest.

A completely different kind of landscape is present in some rain forest of highland areas and even in areas of low hills, where V-shaped valleys with sharply angular interstream areas produce a landscape consisting entirely of slopes that are often steep and straight. This landscape is dominant in the New Guinea highlands, in the island arcs and mountain belts of Indonesia and Central America, and around the edges of high plateaus. Similar topography may also be found in mountainous areas of temperate forest, as in New Zealand.

Volcanoes in various stages of dissection by erosion are found in several parts of the humid tropics, including Central America and Indonesia. These often produce fertile soils. Limestone areas in tropical forest often give rise to a distinctive kind of landscape with "haystack hills," which rise abruptly from the surrounding plains. Broad alluvial plains, such as the lowlands of the Amazon, the Congo, and the Fly in New Guinea, make up large areas of rain forest. There is much variation in alluvial landforms, and these lowlands are by no means uniform with regard to sediments, landforms, or soils.

From A.N. Strahler, *Introduction to Physical Geography* (1965); John Wiley & Sons, Inc.

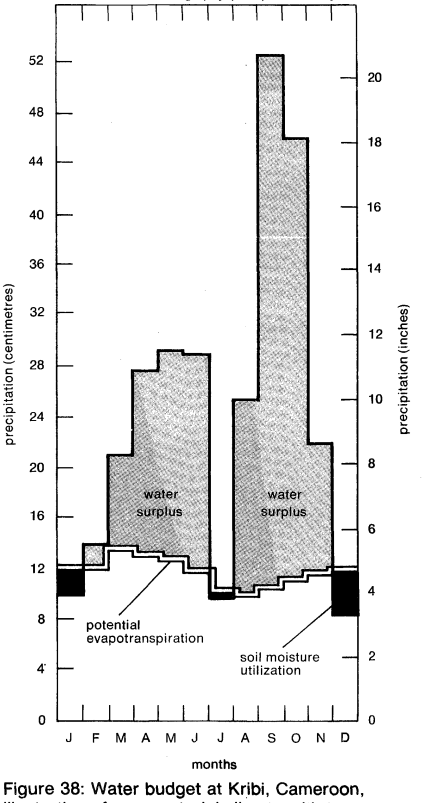


Figure 38: Water budget at Kribi, Cameroon, illustrative of an equatorial climate with two rainfall maxima.

Precipitation, evapotranspiration, and runoff. In rain forest areas there is an excess of precipitation over evapotranspiration for most, if not all, of the year (Figure 38). Only a small proportion of the raindrops fall directly to the ground except at canopy openings. Most are intercepted by the leaves in the canopy and either re-form as waterdrops that fall from the drip tips of the leaves or trickle down the tree trunks. Tropical forest trees generally have broad leaves with drip tips to help shed the water, but, even so, the rain forests probably intercept more rain than any other cover. Estimates of precipitation reaching the ground under various types of cover are: dense forest, 70–80 percent; dense, high grass, 80 percent; cereals and other crops, 80–85 percent; and bare soil, 100 percent.

These data are based on an assumed average of 4 millimetres held (intercepted) per shower. This means that rains of less than 4 millimetres (0.2 inch) do not contribute much to the supply of soil moisture or runoff. The amount retained depends on the total rainfall per day; with heavier showers a greater proportion gets through.

The approximate quantity of rainfall retained by vegetation in a typical site in Indonesia was estimated to be:

Average rainfall	2,400 millimetres (95 inches)
Number of rainy days	160
Rainfall per rainy day	15 millimetres (0.6 inch)
Retained by vegetation (160 × 4)	640 millimetres (25 inches)
Amount reaching soil	1,760 millimetres (69 inches)

Similar results were obtained in Brazil, where it was estimated that about two-thirds of the total rainfall reaches the ground, one-third as raindrops and waterdrops and one-third by streaming down tree trunks.

Evaporation rates can be determined, but results are difficult to correlate with evaporation from a ground surface. Available data indicate an annual evaporation of 542 millimetres (21 inches) at Jakarta, Indon.; 1,139 millimetres (45 inches) at Surabaya, Indon.; 2,300 millimetres (91 inches) from a pond in Madras, India; and 1,930 millimetres (76 inches) from a reservoir in Bombay.

Evaporation from soil is hard to determine, but, in Indonesia where temperature is fairly uniform throughout the year, it was found empirically that evaporation (in millimetres for periods of a month) could be expressed by the formula $E = 60 + 0.125P$, in which E is evaporation and P is precipitation. This means that with less rainfall there will be less evaporation, and with no rain at all there will still be 60 millimetres (2 inches) of evaporation per month.

Extent of transpiration

Transpiration usually accounts for a major share of the water lost by an area, and different species of plants transpire differently under similar conditions. Some types of vegetation, including many trees, may utilize moisture equivalent to nearly all of the rain falling even during the wettest months, whereas grasses or mixed forest may transpire only a fraction of the available moisture. This may account for a rise in the water table after land clearance. The average annual transpiration of plants in Java (in millimetres) has been estimated to be:

Mature forest (trees 740) (undergrowth 130)	870 (34 inches)
Bamboo forest	1,500 (60 inches)
Jungle below 1,000 metres	1,200 (45 inches)
Jungle at 1,000 metres	1,200–500 (45–20 inches)
Jungle at 2,500 metres	600–500 (25–20 inches)

The annual evapotranspiration in the forest zone of the Congo basin is estimated to be 1,650 ± 100 millimetres (65 ± 4 inches).

Forest soils tend to be permanently moist and spongy, but their structure allows a great deal of infiltration. Nevertheless, tropical storms are usually of such great intensity that runoff is common during rains, especially on steeper slopes. There are significant differences between the hydrologic regimes in areas of different topography, especially between the plateau landscapes and the ridge and valley landscapes.

Sediment transport by streams

In ridge and valley areas large runoff produces high rates of erosion, and rivers carry a large sediment load. In flatter areas, minor streams are frequently blocked by fallen trees and associated plant growth to form swamps and lakes. This reduces the ability of the streams to transport debris and leads to sedimentation in temporary pools. Major streams, however, are open and often fast-flowing, their sediment load depending on their catchments (drainage areas).

In the Amazon basin there are "black water" streams, that is, streams draining swamp-forest peats that carry black, acid water that is stained with organic matter but has little sediment. This black water contrasts markedly with the muddy water of other streams and has the incidental advantage to explorers of being free from insects. In New Guinea the Fly River has a comparatively small sediment load, although it is nevertheless murky and opaque; however, its tributary, the Strickland, coming from a more mountainous interior, carries much greater quantities of sediment in its turbulent waters. The difference in the two streams is sufficient to give rise to different alluvial landforms and different vegetation patterns along the respective rivers.

Flooding is important in lowland rain forest. It is often of considerable range, such as the 8-metre (26-foot) variation of the Fly River, which is a great obstacle to development or permanent settlement of such areas. Such floods cover the ground and trees over many square miles. Over-bank floods spread out over the land, depositing new sediments, building levees, and adding new layers of silt to the soils of the lowland forest.

Change occurs in the water table after forest clearance. In Nigeria a rise of 60 metres (about 200 feet) has been noted in 50 years. Larger streams continued to flow in the dry season longer than they formerly did, new springs appeared, and one valley was drowned by a new lake.

Climatic changes also alter hydrologic regimes, but not always in easily predictable ways. In main valleys the reduction in flow due to greater desiccation might be more than offset by increased flow from runoff and from accentuation of groundwater ridges.

Weathering and soils. *Tropical weathering.* In hot, wet areas, chemical weathering is most intense, and, in the intertropical zone, weathering of rock produces clay-rich weathering mantles up to 100 metres (328 feet) deep or even more. The distinction between soil and weathering mantle is hard to draw.

The weathered rock that is still in place retains features of the unweathered hard rock and is separated from an overlying loose mantle of more or less transported material by stone lines of varied composition and controversial origin. All of the rain forests have abundant rainfall, so on well-drained sites there is a washing out, or leaching, of dissolved chemicals. In all of the rain forests, bases such as calcium and magnesium tend to be leached out, but in the tropics there is also removal of silica, which leads to the formation of different soils under tropical and temperate rain forest.

On well-drained sites in the tropics the leaching of bases and silica leads to an accumulation of aluminum and iron oxides and hydroxides and the formation of the clay mineral kaolinite, which is relatively poor in silica (see further MINERALS AND ROCKS: *The major rock-forming minerals: Clay minerals*). Many upland soils in the tropics consist of little more than kaolinite and iron hydroxides, which give the soil a red or yellow colour. There are two main divisions among the tropical red soils: tropical red loams and lateritic soils.

Tropical red loams

Tropical red loams (known also by other names, such as krasnozem and tropical red earth) are deep, friable, and easily tilled soils. They are porous and freely drained at all depths, and there is little change in colour or texture of the soil material in the vertical soil profile. Clay is common, but it is so combined with the iron oxides that it behaves more like a loam and retains a porous, freely drained structure even when wet. It does not break into larger clumps, such as commonly happens in clay-rich soils of temperate regions. Soil profiles of this kind are not affected by water tables within the soil profile. In the tropical rain forest these soils are much more common than the lateritic soils.

Lateritic soils. In contrast to the uniformity of the tropical red-earth profiles, lateritic soils have profiles with well-marked horizons (layers). The uppermost horizon is variable but is commonly a red loam, lighter in texture than the underlying horizons and sometimes containing ironstone gravel. Beneath this is a layer rich in iron. So long as it is kept moist under a forest cover this material will remain soft, but if it is dried out it will irreversibly harden and may even be used for brickmaking. It was to such material that the term laterite was first applied, but the word has been used loosely since then. There was in the 1970s a tendency among soil scientists to return to this early usage, and the whole sequence of associated horizons may be referred to as a laterite, or lateritic, profile. In some profiles the laterite horizon is hard, even within the intact soil profile. Beneath this layer there may be a mottled zone and beneath that a bleached-white, kaolin-rich zone, known as the pallid zone. These zones may be many metres thick.

The formation and potential hardening of the iron-rich zone of lateritic soils are their most distinctive and important features. Hardening can come about in several ways: forest clearing by humans, drying out of the soil due to climatic change, and drying of the soil due to valley downcutting and drainage change.

The iron-enriched layer occurs closer to the ground surface with increasing distance from the equator, and north of 14° N only fossil crusts are found. In Côte d'Ivoire, for instance, there is a latitudinal sequence: in the south, under rain forest that receives more than 1.5 metres (5 feet) of rain annually, are soils that do not form crusts so long as the forest cover is intact; whereas to the north, under subtropical forest that receives less than 1.5 metres of rainfall, the soils are less deep, and ironstone crusts exist on dry sites.

With increasing distance from the equator and longer dry seasons, less silica is removed, and the soils of the humid tropics give way to the ferruginous (iron-rich) soils of the savanna areas. In extremely wet and hot areas, on the other hand, even more silica is leached away, leading to the formation of aluminum-rich bauxite rather than laterite. In the humid tropics the influence of parent rock is less than in temperate areas, and lateritic soils can form on practically any parent rock.

Nevertheless, there is considerable variation. In South America, for example, dark lateritic soils occur in humid areas or on rocks that are rich in calcium and magnesium; red lateritic soils occur on acid rocks (rich in soda and potash); and brown lateritic soils occur on volcanic ash or basalt. Volcanic soils are generally more fertile than others, and in regions of active volcanoes, such as parts of Indonesia, occasional eruptions produce layers of ash that act like a rich topsoil in restoring fertility. Terra rossa soils (clay-rich, red, granular soils) are common on limestone, including that of coral islands, which often have rain forest cover.

Volcanic soils

On base-poor rocks, such as sandstones, podzols (leached soils with distinct sandy topsoils) may form. Tropical podzols are widespread in the catchment areas of the Negro River in the Amazon basin and in Guyana, Malaya, Borneo, and Thailand. These podzols are not under normal forest but instead are under "heath forest," the padang of Southeast Asia.

On tropical mountains there is a general tendency for tropical red loam or lateritic soils to be produced on lower slopes and podzols to be formed higher up, but there

are many complications due to parent material, erosion, aspect, drainage, and other factors.

Thus far, discussion has centred on the soils of leached sites, the well-drained areas on higher ground. A well-marked division is found in many tropical areas, however, with red, leached soils on the uplands and black clays, gray soils, and fresh alluvium in the lowlands. Furthermore, swamp forest with acid soil consisting largely of peat is found in Java, Sumatra, Borneo, Celebes, Malaya, and Guyana, although it is apparently absent in Africa. This peat occurs in large areas, is often more than 6 metres (20 feet) thick, and is composed almost entirely of the remains of trees and other woody plants of the swamp forest.

The common idea that all of the tropical soils are laterites is a gross oversimplification, and detailed investigations in New Guinea and elsewhere reveal many soil types.

Despite the luxuriant plant growth, tropical forest soils are not particularly fertile. Under intense weathering, leaf litter quickly decomposes and thus provides the nutrients essential for the trees, which absorb some of the released ions before they are leached from the soil profile. Most plant nutrients are thus recycled by the trees, and there is little reserve of nutrients in the soil. The humus layer is normally only a few centimetres thick, and decomposition of vegetation is rapid.

Removal of tropical rain forest to allow cropping swiftly eliminates leaf litter, and the water-retaining, spongy soil structure is destroyed. The mineral cycle is broken and the soil productivity declines; laterite dries out because of increased evaporation and so accelerates the development of undesirable physical features. Clearing of lateritic soils also accelerates erosion of the meagre upper horizon, exposing the ironstone layer and facilitating its hardening. Ironstone crusts have been known to form in as little as 30 years in West Africa from such causes.

Cleared forest is therefore of use for only a short time before it is exhausted, so indigenous inhabitants of tropical forest practice shifting cultivation. If only small patches of forest are cleared, regeneration is possible, but on any large scale the changes of soil become progressively worse and irreversible, and the forest gives way to other vegetation types. In some areas there is tangled secondary regrowth, or jungle, while in others, especially in areas of climatic stress, savanna expands at the expense of forest; in New Guinea, for example, large areas of kunai grass (*Imperata*) appear to be of anthropogenic origin, replacing original forest. (C.D.O./Ed.)

Cleared-forest succession

BIOTIC FACTORS

The tropical rain forest is the most complex ecosystem on Earth, and its plant and animal life is richer and more diverse than that of any other type of forest. A two-hectare (five-acre) sample often has more than 100 different species of trees 30 centimetres (12 inches) or more in diameter (in the Malay Peninsula more than 200 species have been recorded), compared with about 25 tree species in the richest temperate broad-leaved forests (North America). Similar data for animals are difficult to obtain, but comparisons have been made of the number of bird species: more than 600 in a 780-kilometre-square area of rain forest in Panama, as against about 100 in a similar area of temperate forest of the eastern United States. The number of bird species, however, is small compared with the number of insects.

Climate and the biota. One obvious reason for the enormous variety of plant and animal life in the rain forest is the climate, which is hot and moist throughout the year. The soil never dries out, and the humidity of the air is always high. This constantly warm and moist environment allows plants and animals to grow and reproduce all year long, though they may not in fact do so. In the rain forest the struggle for survival is, thus, less against a hostile physical environment, as in climates with a cold winter or summer drought, than against the competition of other organisms. This has led, among both plants and animals, to the evolution of many species filling many different niches, or ecological roles, some of which are highly specialized.

Another and less obvious reason for the great diversity of species in the rain forest is probably its age. In the tropics the climate has changed little since the Cretaceous period, and therefore evolution has proceeded uninterrupted for some 66 million years.

Forest structure and microclimates. The trees and other plants that grow together in a primary, or undisturbed, tropical rain forest may seem a chaotic mass of vegetation, but the forest, in fact, has an irregular pattern of structure not easily discernible at ground level. Tree crowns are arranged in several stories, or strata (Figure 39), called A, B, C, D, and E from above downward. These layers are generally poorly defined, so that only an arbitrary boundary can be drawn between one stratum and the next. Most of the trees in the lower strata are immature individuals of species that may later reach a higher stratum. The A stratum, of trees 30 to 50 metres (100 to 165 feet) high, is never continuous; their crowns rarely touch and may be widely separated. For this reason, the roof of the forest, as seen from the air, has an uneven surface, the crowns of the largest trees standing out like gigantic, dark green cauliflowers (Figure 37). The B-stratum trees, whose crowns are smaller and more closely spaced, likewise do not generally form a continuous layer. The trees of the A and B strata are often called the emergents; together they form the forest canopy.

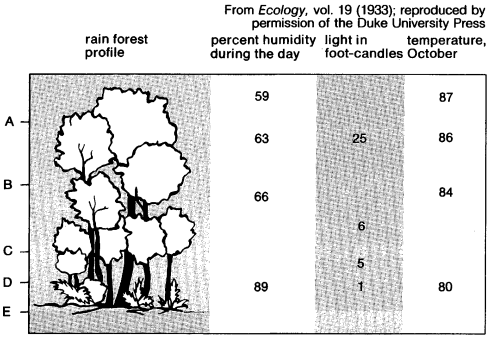


Figure 39: Stratification of a tropical rain forest (see text).

The C and D strata constitute the densest layers of the forest, where little space is unoccupied by trunks, branches, twigs, or foliage. In contrast, the lowest, or E, stratum is often thinly stocked, and much of the ground surface itself is bare except for a thin covering of dead leaves and litter. Contrary to popular belief, the undergrowth of the rain forest is seldom impenetrable: it is usually possible to walk about and to see another person from at least a distance of 50 metres. Only where sunlight reaches the ground in sufficient intensity, in clearings and on the edges of roads and rivers, does the undergrowth become the dense jungle commonly described. Similarly, the undergrowth of young secondary forests, which spring up in clearings left by timber operations or by agriculture, is much denser and more tangled than that of undisturbed primary forest.

The scant undergrowth of jungle

The leaves and branches act as a filter to light and heat radiation and as a barrier to air movements, resulting in wide differences in temperature and humidity of the air in the lower stories, compared with the relatively open canopy. In the undergrowth there is little air movement, and, apart from occasional flecks of sunlight, most of the light has passed through or been reflected from leaf surfaces. Within the forest, conditions are different from those in a clearing or in the open, and there is a range of "local climates" (microclimates).

At midday in sunny weather the temperature in the A and B strata might be 32° C (90° F), while in the undergrowth it could be 26° C (79° F). During cloudy or rainy weather the difference would be smaller, but the upper strata are always warmer than the undergrowth during the day. At night the position is reversed, and, especially on clear nights, the air around the treetops is cooler than that near the ground. Although at night the air in all strata is nearly saturated with moisture, during the day there are considerable differences in humidity between the upper strata, where relative humidity often falls to 60 percent in the afternoon of a sunny day, and the undergrowth,

where the humidity rarely drops below 80 to 90 percent. The undergrowth of the tropical rain forest is in fact one of the most constant biological environments on Earth, and it is not surprising that, as in caves, another unusually constant environment, some strange forms of animal life are found—some of which, such as the ricinuleids (class Arachnida) and the wormlike walking worm (*Peripatus*), can be regarded as "living fossils."

Character of rain forest communities. The rain forests of tropical America, Africa, and Southeast Asia to Oceania (northern Australia and the western Pacific islands) are similar in many features, but each has a distinctive plant and animal life, with almost no species common to all three and not many common to any two of the tropical land sectors. Whole groups of organisms are confined to a single sector. For example, the dipterocarps (family Dipterocarpaceae), are the most important family of trees in Indo-Malayan rain forests but are absent in the Americas and in Africa are represented by only a few small trees of the savannas. The pineapple family (Bromeliaceae), so common as air plants (epiphytes) in the American tropics, have only one species in the Old World tropics. Much the same is true of the animals: tropical America alone has monkeys with prehensile tails, sloths, toucans, and leaf-cutting ants, but it has no apes, hornbills, or large mammals, such as elephants and rhinoceroses, all of which are confined to tropical Africa and Asia. Often a group in one sector replaces one of another; thus, the flower-visiting sunbirds (family Nectariniidae) of the Old World occupy a niche similar to that of the hummingbirds (family Trochilidae) of tropical America.

Plants. The tropical rain forest is overwhelmingly dominated by trees and woody-stemmed vines (lianas). Woody plants are the chief producers of the food on which all the animals depend and of the shelter in which they live. Most of the nonwoody plants are epiphytes living on the trunks and branches of the trees, but a few, including grasses, grow on the ground.

Rain forest trees

Rain forest trees range from giants 60 metres (almost 200 feet) high down to treelets of not more than 1 metre (just over 3 feet). The great majority belong to the large group of flowering plants known as dicotyledons, but palms and other monocotyledons occur in the lower stories and in open areas. Conifers are found only in certain parts of the rain forests of Borneo, New Guinea, and neighbouring areas. Although belonging to a large number of different families, rain forest trees are remarkably uniform in appearance, the majority having thick, leathery leaves, like those of temperate laurels in size and shape. Most are evergreen, but some are deciduous, remaining leafless for a few days or weeks. Because the behaviour of the different plants is not closely synchronized, the forest as a whole always appears in full leaf. The young leaves of the trees are often bright red and hang limply, as if wilted.

Flangelike buttresses, extending up the trunk for 3 to 4 metres (10 to 13 feet) and outward for about the same distance, are characteristic of many of the larger trees, and stiltlike aerial roots are not uncommon among the smaller trees. Root systems in general are shallow—even those of the largest trees seldom penetrate deeper than a metre—and most of the fine nutritive roots are concentrated in the humus-rich layer within a few centimetres of the soil surface.

Large, brilliantly coloured flowers are not common in the rain forest; those of many trees are inconspicuous and white or greenish. The production of flowers on the main trunk, rather than on the twigs or branches, is characteristic of many small- to medium-sized trees; this feature may be related to pollination by bats or other animals that cannot easily reach flowers hidden within a mass of twigs and leaves.

Lianas are plentiful, especially where clearings are formed by the death of old trees. These woody vines reach the tops of all but the tallest trees, linking them together and competing with them for space and light. In the eastern tropics, climbing palms (rattans) are common and grow to a length of 100 metres (328 feet) or more. Strangling plants, which include many figs (*Ficus*) and, in the American tropics, species of autograph trees (*Clusia*), also compete with the trees. These plants start life as epiphytes on tree branches to which their seeds are carried by birds; later they send aerial roots to the ground and enclose the host trunk in a network of stout, woody roots that may kill it, leaving the strangler as an independent tree.

Epiphytes are commonest on branches 30 metres (about 100 feet) or more above ground in positions fully exposed to the sun. Common rain forest epiphytes include orchids, bromeliads, and specialized ferns as well as lichens, mosses, and liverworts that grow on the surface of leaves in the humid, shaded undergrowth. Most epiphytes do not harm the host tree but simply depend on it for physical support. They obtain mineral nutrients from rainwater and from the organic debris that collects among their roots, often brought there by ants. Many epiphytes have special adaptations, such as bracket- or pitcher-shaped leaves that collect water and humus. The most remarkable of such adaptations are the "tanks" of bromeliads, formed by closely overlapping leaf bases. These may hold several litres, or quarts, of water and provide a home for many kinds of aquatic organisms, including mosquito and other insect larvae, tadpoles, and even crabs.

Many rain forest plants are not green and so cannot make their own food; they live either on decaying matter as saprophytes or on other living plants as parasites. Although most of these nongreen plants are fungi, a few are specialized flowering plants. One of the most spectacular of all rain forest plants is the Malaysian monster flower (*Rafflesia*), which is a parasite of certain woody vines; its flowers may be a metre (more than three feet) in diameter.

Animals. The variety of animals even in a small area of tropical rain forest is so vast that a large proportion of species of all of the groups, except the vertebrates, are still unnamed and unclassified, a formidable obstacle to research in rain forest ecology. All groups of vertebrates except fishes are represented and almost all groups of terrestrial invertebrates. Even groups that are normally aquatic, such as flatworms (planarians) and polychaete worms, are able to live in the continually moist undergrowth of the rain forest. Land leeches (class Hirudinea) that attach themselves to passing animals (including man) are one of the most troublesome pests of the rain forests of the Philippines and other parts of the eastern tropics.

Wide representation of animal groups

Rain forest mammals range from such large creatures as the elephant (of Africa and Asia) and forest rhinoceros (of Malaysia) down to tiny mice and shrews. In tropical America the range of size is less, because there are no mammals larger than jaguars and tapirs. A tendency to gigantism is shown among rain forest invertebrates, which include the giant snail of Africa, "bird-eating" spiders, and the goliath beetle. The splendid *Ornithoptera* butterflies of Malaysian forests, the blue morpho butterflies of tropical American forests, and the huge Atlas moth of Asian forests are among the largest Lepidoptera.

Although the animals of the tropical rain forest are abundant and many striking and colourful in appearance, often few can be seen except for some butterflies and bees, an occasional hummingbird, ground squirrel, or armadillo, and the omnipresent ants. One reason for this is that a large proportion of the animals, including most of the birds, live in the trees and rarely come down to ground level. Others are extremely well camouflaged—*e.g.*, the leaf butterflies, which are virtually invisible when they rest on dead leaves; green tree snakes, which hide among the foliage; and the numerous shy, brownish birds and mammals, which are well concealed until they move.

Another reason for the apparent scarcity of animal life is that many of the rain forest animals are nocturnal, including such mammals as lemurs and tarsiers, such birds as nightjars and owls, most toads and frogs, and most moths, as well as a host of other insects of many different families. All of these animals become active only after dark, when the temperature is lower and the dangers from water loss and predators less than in the day. Many nocturnal animals are highly adapted to their way of life; the large saucer eyes of tarsiers, bush babies, and owls and the luminous organs of fireflies are a few examples of the ways in which nocturnal animals are adapted to their way of life.

Often more evident than the animals themselves are their habitations. Birds' nests are seldom easy to find, but the beautifully constructed nests of wasps, bees, ants, and termites are everywhere to be seen. Termites, which are among the most abundant of all of the forms of life in the rain forest, play a vital part in removing deadwood and plant debris. During the day they remain within their nests (termitaria) and in covered passages of cemented wood and soil particles. Termitaria are found both on trees and rising from the ground as fantastic, spired or conical structures sometimes a metre or more in height. Both types of nests often have highly developed rain-shedding devices: in the tree nests, chevronlike arrangements of ridges direct the water away, and in some freestanding termitaria superposed "hats," somewhat like the roof of an African grass hut, drain water away.

Animal stratification

The animals of the tropical rain forest, like the plants, are stratified, but, because of the difficulty of exploring the forest above ground level, zoologists usually divide the forest merely into canopy and undergrowth, or into top, middle, and lower layers, rather than into the five strata mentioned earlier.

In the canopy not only is there more open space but, because of the higher light intensity and greater photosynthetic activity of the foliage, more young leaves, fruits, and flowers are available as food. In the undergrowth there is less variety of plant food, though leaves and wood that continually fall from above provide a plentiful source of food for some animals and plants.

For the larger animals the canopy provides opportunities for flying, gliding, and leaping, as well as for climbing and running along branches; both sight and hearing are important senses in seeking food and escaping predators. The life of an undergrowth animal is different. Running, fluttering, hopping, and climbing are the chief modes of locomotion, and, because the animals tend to be camouflaged, communication between them is more by sound than by sight. Many undergrowth birds—for example, the bellbirds, which include the cotingas, honeyeaters, and some of the thickheads and shrikes—have dull, concealing colours but surprisingly loud, clear calls.

The American naturalist William Beebe was one of the first to study the stratification of animals in a tropical forest. In Guyana he recognized five height "zones," each mainly inhabited by different mammals and birds. More recent studies in Malaya, where the forest animals are similarly stratified, have shown that the animals of each level tend to differ in both their mode of locomotion and their feeding habits. For example, the mammals living on the ground include large species, such as the elephant and rhinoceros, with no climbing ability, and smaller species that climb only to a limited extent. This group includes herbivores, which browse on leaves and eat fallen fruits (elephant, deer), mixed feeders (sun bear), and carnivores (tiger, leopard). The inhabitants of the middle levels and canopy have little contact with the ground; they climb, leap, or swing from branch to branch (gibbons, monkeys) or glide for long distances (flying squirrels, flying lemur). Mostly they are fruit and insect eaters, though some are carnivorous (clouded leopard, marten). The birds of different strata show similar differences in their movements and feeding habits.

The animals of the rain forest show many striking structural adaptations to their mode of life, those of the treetops, in particular, having evolved special characteristics not found in those living nearer the ground—*e.g.*, the winglike skin flaps of the flying squirrels and lemurs, the prehensile tails of the New World monkeys, the peculiar limbs of sloths (which allow them to hang from branches upside down), the stiff tails of woodpeckers (used as an aid to climbing), and the beaks of parrots (well adapted to cracking nuts and to assist in climbing).

Adaptations to the requirements of climbing in different strata are particularly well shown among rain forest primates. The only forest primates that live mainly on the ground are the chimpanzees and gorillas of Africa; they have limbs adapted to walking and climbing but not to venturing high in the trees. The tarsiers, galagos, and lemurs, found mostly in the small trees of the B and C strata, have short legs and arms; they climb well, run, and make short jumps. The primates of the treetops include the orangutan and the athletic gibbons of Malaysia and monkeys of many kinds. The orangutan and the gibbons have long arms with which they swing from branch to branch, while the monkeys are better adapted to running along branches and leaping from tree to tree. In feeding habits the different groups of primates are also well adapted to the strata in which they mainly live.

Specialized tree-bound animals

Among the animals that rarely come down to earth are certain frogs and toads, which must obtain water for their tadpole stage. Some lay their eggs in water held in holes in trees or by epiphytic bromeliads; in others the tadpole stage is abbreviated, and the developing eggs are carried attached to a moist membrane on the mother's back. One Central American species lays its eggs in packets fixed to branches of trees overhanging streams, so that the tadpoles fall into the water when they hatch.

Rot holes and bromeliad "tanks" also provide treetop habitats for mosquitoes and other insects with aquatic larvae. Mosquitoes that breed in the treetops are of much practical importance because some are carriers of malaria, yellow fever, and other diseases of humans and animals. Malaria-carrying anopheles mosquitoes breeding in forested regions of tropical America sometimes come down to ground level and become a source of infection for people living in the area. Similarly, aëdes mosquitoes, which carry yellow fever, normally bite monkeys living in the treetops in the forests of Africa and the Amazon, but when trees are felled these mosquitoes may bite lumbermen and initiate epidemics among susceptible human populations.

There are many rain forest inhabitants normally found below the soil surface to a depth of a metre or more (three or four feet). This underground community includes burrowing vertebrates, such as armadillos, which also spend much time above ground, and others, such as caecilians (wormlike amphibians), worm snakes, and many more, which are wholly subterranean. Much more important than these, however, is the host of small soil invertebrates (worms, mites, insects, etc.) that, together with fungi and microorganisms, play a large part in decomposing dead vegetation and freeing the elements contained therein for recirculation in the forest ecosystem.

The year-round warm and moist conditions in the rain forest are of great significance for animals as well as for plants. In general, rain forest animals are unrestricted in their activities by seasonal changes of climate. The environment demands no prolonged rest periods—neither hibernation nor estivation—and resting stages of insects are either absent or short, with little relation to the time of year.

An important aspect of the environment for animals is that food is always available. Foods such as leaves, flowers, and fruit may vary in abundance but are never unobtainable as in a temperate-zone winter. Many forest birds depend exclusively on fruits, and bats as well as hummingbirds and sunbirds have evolved flower-visiting habits.

Some animals actually reproduce at all times of year. Some African forest birds, such as fruit pigeons (family Treroninae) and flycatchers (families Muscicapidae and Tyrannidae), breed at all of the seasons, and some tropical butterflies, such as the swallowtails (*Papilio*), have a succession of broods throughout the year rather than the one or two as in temperate Europe and America. Although breeding is always possible, the majority of animals have fairly definite breeding periods; these periods, however, are not synchronized for different species, so there is no time at which some birds, mammals, and insects are not reproducing. The breeding season of many rain forest birds and mammals seems to be related to variations in abundance of their chief foods.

Reproduction tends to take place more frequently or over a larger part of the year than in temperate animals, and there is a tendency for fewer young to be born in each brood. This is especially true of tropical birds, which lay fewer eggs at a time than do temperate birds.

Relationships between organisms. The large and diverse populations of organisms in a tropical rain forest interact

Complex association forming a stable ecosystem

with each other and with their environment to form an extremely complex but stable ecosystem. As in any other natural community, the basic relationships between the constituent organisms are the food and energy links. Radiation from the Sun provides energy for the photosynthesis of trees and other chlorophyll-containing plants. These are eaten by herbivorous animals, the most important of which in the rain forest are probably the insects. These primary consumers are the food of secondary consumers, such as insectivorous birds, which are in turn eaten by others, so that there are food chains with up to four or five "links" (trophic levels). The food chains end with such predators as jaguars, eagles, and owls, all of which are relatively scarce and, because of their size and large food requirements, unselective in their diet. Most rain forest plants and animals have parasites, which may in turn be attacked by parasites (hyperparasites), establishing another kind of food chain.

An important difference between a tropical forest and other ecosystems is that the food chains (or food webs, because they are interconnected) are so many and so complex. A further characteristic of the rain forest is that within the main ecosystem with its food webs, there are small subsidiary systems, which are partially isolated. One subsidiary system is that of the bromeliad tanks in tropical American rain forests. In these the chief sources of food are plant debris falling into the tanks or brought there by ants and production of food by populations of certain green algae. The consumers are worms, insect larvae, tadpoles, etc., some of which remain in the tanks, whereas others, such as frogs and mosquitoes, leave them when mature to find food elsewhere in the forest.

Another subsystem consists of the large number of organisms dependent in one way or another on ants. Many kinds of insects live in ants' nests in various relationships with their hosts, the nature of which is often not clear. Some large predators, such as anteaters, feed largely on ants and termites, while antbirds (family Formicariidae; a characteristic group of thrushes and other species) follow the migratory army ants (*Eciton*) of tropical America and the similar driver ants (subfamily Dorylinae) of Africa as they advance through the forest, eating not the ants themselves but the insects and small vertebrates disturbed by them.

There are also many kinds of ant–plant relationships, the most remarkable perhaps being that between green plants and the tropical American leaf-cutting ants (*Atta*). These ants bite out small semicircular pieces from leaves and carry them in procession into the recesses of their moundlike nests. Special fungi, which grow on the leaf bits, produce knoblike outgrowths on which the ants feed.

Another kind of ant–plant relationship found in all of the tropical rain forests is that between certain types of ants and certain species of plants in which they nest, known as myrmecophytes ("ant plants").

Stratification of pollinators and seed dispersers

Among rain forest trees, in contrast to temperate trees, pollination by wind is uncommon, and insects, birds, or bats are generally the pollinators. Specialized coadaptations between flower and pollinator have often evolved, the most extraordinary being perhaps the fantastically elaborate pollination mechanisms in some tropical orchids. Several unusual features common among rain forest trees, such as flowers borne on the main trunk or hanging from the branches on long cordlike stalks, seem to be adaptations to bird or bat pollination.

Animals of many kinds play a part in dispersing the seeds of many rain forest trees and vines, though some, especially species colonizing gaps and clearings, have windborne fruits. Trees of the A and B strata often have heavy fruits that crash through the foliage beneath and are subsequently distributed by rodents and other small animals. Other trees have edible nuts or berries, which are dispersed by parrots, monkeys, and other treetop inhabitants.

Many forest plants have evolved protective features that make them distasteful or toxic. Members of the pea family, to which many rain forest trees and vines belong, often have bitter seeds avoided by insect larvae. Similarly, the leaves of the madder family (Rubiaceae), the largest family of tropical plants, often contain poisonous or bitter compounds; it can hardly be a coincidence that few caterpillars are known to feed on them.

BIOLOGICAL PRODUCTIVITY

The luxuriance of the tropical rain forest and the speed with which clearings are invaded by secondary growth suggest that the rates of photosynthesis, growth, and other physiological processes in tropical plants must be high compared with those in temperate climates. To test such impressions, estimates of biomass (total quantity of organic material per unit of land area) and of the amount of new material added yearly (biological productivity) are required.

A few estimates for some primary rain forests indicate that from 300 to 400 metric tons of plant material occupy a hectare, which is greater than the biomass of most temperate forests but less than that of the redwood forests of California. It is probable that the biomass of the forests of the Malay Peninsula, however, reaches as much as 500 metric tons per hectare.

In every case, more than 80 percent of the rain forest biomass consists of woody stems and branches, leaves forming only a small proportion of the total. The organic material that is available as food for herbivorous insects and other leaf feeders is thus a small fraction of the total amount present. If the biomass can be regarded as the forest's capital, then the annual increment of organic material is its income, and this latter depends on several factors, of which the most important is the amount of available energy. More solar radiation falls on the tropical zone than on any other part of the Earth. In the humid tropics the radiation actually reaching the Earth's surface is about double the amount received at 60° N, the latitude of Leningrad, and the northern tip of Newfoundland. It might, therefore, be expected that the tropical rain forest would produce more organic matter per unit of land surface than any other kind of vegetation, but this does not in fact seem to be the case.

Surprisingly low organic productivity

A reliable estimate of organic production in the tropical rain forest of Côte d'Ivoire (West Africa) gave the figure of nine metric tons per hectare per year, not much higher than the productivity of a beech wood in Denmark. For several reasons this forest was perhaps untypical, but estimates for other tropical areas do not suggest much higher figures.

If it is indeed true that tropical forests are not much more productive than those in temperate climates, the explanation may lie mainly in the high respiration rates of tropical plants. The total amount of organic material produced by a tropical tree in a year may be prodigious (about two or three times that produced by a temperate tree), but much of it is unavailable for plant growth because it is lost in respiration, the process by which oxygen is taken in and carbon dioxide given out, with release of heat and chemical energy. It seems that, generally, tropical plants lose through respiration a larger proportion of the carbohydrate material they build up during the day than do temperate plants.

The tall trees of the forest, whose crowns are more or less fully exposed to the sun, photosynthesize more actively than those in the shaded undergrowth. In Kampuchea (Cambodia) the A stratum contributes more than half the productivity of the forest, though its foliage forms only about a fifth of the total leaf area.

The rate of growth in girth and height, like the productivity, of primary rain forest trees does not seem to be unusually high. The evidence suggests that primary rain forest trees do not, in general, grow as rapidly as the fastest-growing temperate trees, such as poplars, eucalyptus, and some pines. Very high growth rates, however, are found in trees of young secondary rain forests, such as the trumpet tree (*Cecropia*) of tropical America and the African parasol tree (*Musanga cecropioides*). The latter can reach a height of more than 20 metres (65 feet) in 20 years. Such trees do not, in fact, grow as fast as such temperate plants as the sunflower under comparable conditions. Temperate plants, however, grow fast only during a short summer season, whereas the growth rates of trees such as the parasol tree are maintained through the year.

At tropical temperatures, decomposition is normally much faster than under temperate conditions. As a result, plant and animal remains do not accumulate as litter or as humus incorporated in the soil. Studies in the Thailand rain forest showed that more than 80 percent of the carbon present is in the form of living wood and leaves and less than 20 percent as dead material on and in the soil. By contrast, in a northern coniferous forest more than half the carbon may be present as organic matter in the soil or as litter on its surface.

Elements released from the dead organic material by decomposition are almost immediately reabsorbed by tree roots, often with help from fungi living in close association with them. The small amounts of nutrients washed down into the soil by the rain or carried away by streams are replaced by additional nutrients set free by the weathering of soil minerals. There is thus an almost closed cycle of mineral nutrients, and a rain forest soil is able to support a luxuriant vegetation and abundant animal life in a permanent equilibrium with its environment. Because the loss of nutrients is so small, the concentration of mineral elements in the water of rivers, such as the Amazon, that drain large rain forest areas is scarcely higher than in rainwater.

Effect of forest clearing

In the undisturbed rain forest the greater part of the mineral resources is locked up in the biomass itself, but there is a drastic change if the forest is felled. When clearings are made by the slash-and-burn system—the commonest method of growing food crops in rain forest areas—good harvests are seldom obtained after the first year. When the vegetation is burned, the minerals contained in it are suddenly released in soluble form, and a large proportion is washed away in surface erosion or carried down into the soil and drained away later to the rivers. The conversion of vast tracts to pastureland results, in addition to the destruction of species, in the rapid depletion of nutrients and an inferior utilization of solar energy. (P.W.R./Ed.)

Other tropical-subtropical woodland complexes

At the edge of forests are transitional woodland formations called forest ecotones. Two types are briefly described below. In tropical latitudes the savanna and thorn forest ecotones are intermediate formations between the rain forest and semiarid desert.

SAVANNAS

Savannas lie between the tropical rain forest and semidesert vegetation in a broad latitudinal zone between 5° and 20° N and S of the equator (Figure 29). The dominant vegetation is an open parklike tree cover over dense, tufted and often tall grass. The climate is tropical and semihumid, with summer rain and a long dry season. Precipitation varies between 600 millimetres (about 23 inches) and 1,500 millimetres (about 60 inches); at the lower range of rainfall the savanna passes into thorn scrub, whereas at the upper limit the savanna grades into semievergreen or mixed deciduous forest. The savanna occurs extensively in Africa, both north and south of the equator, covering most of the East African highlands. In South America the savannas of Venezuela and Brazil are separated by the Amazon forests. In Australia savannas are found between the forests of the wet coastal regions and the deserts of the interior. These "savanna woodlands" of tropical Australia include many areas in which the tree cover consists mostly of eucalyptus. In all localities the savanna offers a unique environment for large animals, many of which, such as the zebras and giraffes of Africa, are particularly adapted to the open woodland and some, like the carnivores, to sparse woodland conditions. The environment is especially prone to fire, and it is probable that large areas of savanna are stabilized by fire. Annual burning is a common practice in Africa, and the savanna vegetation is subject to severe selection; only species resistant to periodic fires can maintain themselves. If burning is abandoned, increase in woody plant development takes place.

Savanna trees

The dominant tree species of the savanna are typically deciduous, shedding their leaves during the dry season; they have a gnarled, Y-shaped form, branching a few feet above the ground and spreading out to an umbrella-shaped crown. In Africa, outside the influence of the savanna-closed forest margins, the miombo (*Brachystegia-Isoberlinia*) woodland grades into an acacia formation as drought conditions increase. This formation includes also elephant grass mixed with other tussock grasses of the genera *Imperata, Andropogan,* and *Hyparrhenia.* At their most luxuriant they are often more than 3.6 metres (12 feet) tall at the end of the growing season; their stature diminishes as conditions become drier, until, under the acacias at the border of the semiarid desert and thorn forest, they give way to triple-awned grasses.

THORN FORESTS

The thorn forest is stabilized by the climate over most of its area, although occasionally it is maintained by grazing or burning. It is found on the equatorial margins of subtropical and tropical semideserts, where rainfall in a two-to four-month summer season rarely exceeds 700 millimetres (about 27 inches). The forest resembles the savanna woodland, but it is somewhat shorter in stature; the ground layer is confined to sparse arid-adapted grasses but is richer than its savanna neighbour in understory woody species.

The vegetation of thorn forests is markedly adapted to dry conditions, the trees being deciduous or, if evergreen, with leaves protected against drying; the branches often bear thorns as protection against grazing animals. Cacti and other spiny succulents appear in the ground layers and understory. The common dominant trees are native acacias, which appear in the South American, African, and Australasian formations, in the last case displacing the eucalyptus of the savanna. In the Asian formation khair (*Acacia catechu*) is associated with teak (*Tectona hamiltoniana*), a deciduous nonthorny species. Like the tree species of the savanna, the thorn forest dominants are deep-rooted, competing most successfully on soils where water penetrates deeply and quickly. (G.K.E./Ed.)

BIBLIOGRAPHY

General works: EUGENE P. ODUM, *Fundamentals of Ecology,* 3rd ed. (1971), is a comprehensive college textbook and reference, designed also for the citizen, educator, and political leader. *Scientific American,* vol. 223, no. 3 (1970), is an entire issue devoted to the ecology of the biosphere, with excellent general articles. W.C. ALLEE *et al., Principles of Animal Ecology* (1949), is a classic reference work that is valuable for most topics. N.M. JESSOP, *Biosphere: A Study of Life* (1970), includes discussion of energy flow and material cycling. Three works that emphasize the ecosystem approach to ecology are EUGENE P. ODUM, *Ecology* (1963); EDWARD J. KORMONDY, *Concepts of Ecology,* 3rd ed. (1984); and ROBERT H. WHITTAKER, *Communities and Ecosystems,* 2nd ed. (1975). HOWARD T. ODUM, *Environment, Power, and Society* (1971), is a semipopular introduction to systems ecology, with emphasis on technological, political, and economic solutions to environmental problems. JOHN WIENS (ed.), *Ecosystem Structure and Function* (1972), contains the proceedings of the 31st Biology Colloquium, in which five authors deal with ecosystem concepts. ALDO LEOPOLD, *A Sand County Almanac, and Sketches Here and There* (1949, reissued 1987), is an environmental classic, with near-poetic essays on "The Land Ethic," "Wilderness," and "Conservation Esthetic," all imbued with ideas about the place of humans in ecosystems. EDWARD J. KORMONDY and J. FRANK MCCORMICK (eds.), *Handbook of Contemporary Developments in World Ecology* (1981), provides good summaries of developments in countries in five continents. (D.M.G./E.P.O./Ed.)

Distribution of organisms: Local distribution is considered in detail in H.G. ANDREWARTHA and L.C. BIRCH, *The Distribution and Abundance of Animals* (1954, reissued 1970); and VERA COPNER WYNNE-EDWARDS, *Animal Dispersion in Relation to Social Behaviour* (1962, reissued 1972). Large-scale distribution is dealt with in LÉON CROIZAT, *Panbiogeography,* 2 vol. in 3 (1958); PIERRE DANSEREAU, *Biogeography: An Ecological Perspective* (1957); PHILIP J. DARLINGTON, JR., *Zoogeography: The Geographical Distribution of Animals* (1957, reprinted 1982); L.F. DE BEAUFORT, *Zoogeography of the Land and Inland Waters* (1951); DAVID J. DE LAUBENFELS, *A Geography of Plants and Animals* (1970); M.J. DUNBAR (ed.), *Marine Distributions* (1963); SVEN EKMAN, *Zoogeography of the Sea* (1953, reissued 1967; originally published in German, 1935); PETER A. FURLEY, *Geography of the Biosphere: An Introduction to the Nature, Distribution, and Evolution of the World's Life Zones* (1983); WILMA B. GEORGE, *Animal Geography* (1962); RONALD GOOD,

The Geography of the Flowering Plants, 4th ed. (1974); ROBERT H. MACARTHUR, *Geographical Ecology: Patterns in the Distribution of Species* (1972, reprinted 1984); and NICHOLAS POLUNIN, *Introduction to Plant Geography and Some Related Sciences* (1960). The factors behind dispersal are examined in SHERWIN CARLQUIST, *Island Life: A Natural History of the Islands of the World* (1965); CHARLES S. ELTON, *The Ecology of Invasions by Animals and Plants* (1958, reissued 1977); CARL H. LINDROTH, *The Faunal Connections Between Europe and North America* (1957); ROBERT H. MACARTHUR and EDWARD O. WILSON, *The Theory of Island Biogeography* (1967); MIKLOS D.F. UDVARDY, *Dynamic Zoogeography: With Special Reference to Land Animals* (1969); and L. VAN DER PIJL, *Principles of Dispersal in Higher Plants,* 3rd rev. and expanded ed. (1982).

(P.R.G./Ed.)

Biotic interactions: PAUL BUCHNER, *Endosymbiosis of Animals with Plant Microorganisms,* rev. ed. (1965; originally published in German, 1953), provides a detailed, authoritative treatment of symbiosis between insects and microorganisms. THOMAS C. CHENG, *Symbiosis: Organisms Living Together* (1970), a semipopular book, encompasses types and origins of symbiotic associations, including human parasites and their life cycles. PAUL DEBACH (ed.), *Biological Control of Insect Pests and Weeds* (1964, reissued 1973), is a technical treatise, with detailed accounts of the biotic interactions on which biological control is based. PAUL L. ERRINGTON, *Of Predation and Life* (1967); and S. MARK HENRY, *Symbiosis,* 2 vol. (1966–67), survey the symbiotic associations of microorganisms, plants, marine organisms, invertebrates, birds, ruminants, and other organisms. See also L. MARGULIS, "Symbiosis and Evolution," *Scientific American,* 225:48–57 (1971). (P.S.M./Ed.)

Biological populations: MAURICE E. SOLOMON, *Population Dynamics,* 2nd ed. (1976), is a short introduction to the principles of density-dependent regulation, primarily in insects. G.F. GAUSE (GAUZE), *The Struggle for Existence* (1934, reprinted 1971), is a classic discussion of the first laboratory experiments with protozoa in the study of predator–prey interactions and competitive exclusion. DAVID LACK, *Population Studies of Birds* (1966), summarizes and relates many studies by a number of authors. T.T. MACAN, *Freshwater Ecology,* 2nd ed. (1974), discusses species behaviour and interrelationships. R.M.F.S. SADLEIR, *The Ecology of Reproduction in Wild and Domestic Mammals* (1969), includes information on litter size and age of breeding. T.R.E. SOUTHWOOD (ed.), *Insect Abundance* (1968), includes a study of the winter moth. ADAM WATSON (ed.), *Animal Populations in Relation to Their Food Resources* (1970), collects reviews of the influence of food on numbers in many types of animals. (D.L.L./Ed.)

Biological communities: JARED DIAMOND and TED J. CASE, *Community Ecology* (1986), provides a good introduction. J. BRAUN-BLANQUET, *Plant Sociology: The Study of Plant Communities* (1932, reissued 1987; originally published in German, 1928), is the standard European textbook on plant communities. S.R. EYRE, *Vegetation and Soils: A World Picture,* 2nd ed. (1968, reissued 1977), offers a concise treatment of world vegetation. S. CHARLES KENDEIGH, *Animal Ecology* (1961), discusses animal communities, niches, and biomes. A.W.F. SCHIMPER, *Plant-Geography upon a Physiological Basis* (1903, reprinted 1964; originally published in German, 1898), is one of the classic treatises on plant communities of the world. ROBERT H. WHITTAKER, *Classification of Natural Communities* (1962, reprinted 1977), *Ordination of Plant Communities,* 2nd ed. (1982), and *Classification of Plant Communities,* 2nd ed. (1982), review theories and approaches to classifying communities.

(R.H.W./Ed.)

Aquatic ecosystems: Introductions to various aspects of oceanic ecosystems include WILLIAM J. CROMIE, *The Living World of the Sea* (1968), a popular book on marine subjects of wide general interest; KLAUS GÜNTHER and KURT DECKERT, *Creatures of the Deep Sea* (1956; originally published in German, 1950), a classic work on deepwater animals; GEORGE E. MACGINITIE and NETTIE MACGINITIE, *Natural History of Marine Animals,* 2nd ed. (1968), an excellent, generalized treatment of the ecology of widely representative marine animals, with emphasis on those from the Pacific coast of North America; ROBERT C. MILLER, *The Sea,* 2nd ed. (1975), a fascinating book on marine life, habitats, and phenomena; HILARY B. MOORE, *Marine Ecology* (1958), an authoritative book on marine ecology, suitable for use as a college text; and J.A. COLIN NICOL, *The Biology of Marine Animals,* 2nd ed. (1967), a scholarly, exhaustive treatment of physiological topics and mechanisms pertaining to marine organisms.

Analyses of ecosystems in lakes and rivers can be found in DAVID G. FREY (ed.), *Limnology in North America* (1963), a geographically organized account of the lakes—and, to some extent, of the rivers, springs, estuaries, and wetlands—of North America, with a useful annotated bibliography; G. EVELYN HUTCHINSON, *A Treatise on Limnology,* 3 vol. (1957–75), an encyclopaedic work on lakes, probably the most comprehensive ever attempted for any type of aquatic ecosystem; H.B.N. HYNES, *The Ecology of Running Waters* (1970), a comprehensive book on the biology of rivers; GEORGE K. REID and RICHARD D. WOOD, *Ecology of Inland Waters and Estuaries,* 2nd ed. (1976), a unique attempt to cover the ecology of lakes, rivers, and estuaries in relatively simple fashion; BRIAN MOSS, *Ecology of Fresh Waters,* 2nd ed. (1988); and FRANZ RUTTNER, *Fundamentals of Limnology,* 3rd ed. (1963, reprinted 1974; originally published in German, 1962), a short introduction to lakes, with emphasis on their biology.

Studies of special ecosystems include JAMES GREEN, *The Biology of Estuarine Animals* (1968), which contains information about estuarine animals and how they fit into the ecosystem; JOHN TEAL and MILDRED TEAL, *Life and Death of the Salt Marsh* (1969), a lyrical but authoritative account for nonspecialists of the salt marsh ecosystem; N.A. HOLME and A.D. MCINTYRE (eds.), *Methods for the Study of Marine Benthos,* 2nd ed. (1984), a biological handbook with a section on coastal life and on primary production of coastal communities; J.R. LEWIS, *The Ecology of Rocky Shores* (1964, reissued 1976), a detailed, well-illustrated treatment of zonation on rocky shores in the British Isles; and T.A. STEPHENSON and ANNE STEPHENSON, *Life Between Tidemarks on Rocky Shores* (1972), a summary of the Stephensons' investigations of seashore life and zonation in both hemispheres, together with a review of other investigations in places not visited by the Stephensons.

(C.N./D.A.L./A.J.So./Ed.)

Terrestrial ecosystems: Information on terrestrial ecosystems can be found in the voluminous literature on biology and the Earth sciences.

Methods and results of research are provided in the series *Ecological Studies,* the first volumes of which are DAVID E. REICHLE (ed.), *Analysis of Temperate Forest Ecosystems* (1970); and H. ELLENBERG (ed.), *Integrated Experimental Ecology* (1971). Forest ecosystems are emphasized in the two volumes cited above and in V. SUKACHEV and N. DYLIS, *Fundamentals of Forest Biogeocoenology* (1968; originally published in Russian, 1964). Other major ecosystem types are treated in the following works: M.J. DUNBAR, *Ecological Development in Polar Regions: A Study in Evolution* (1968), on the adaptations of polar organisms, especially marine, to low temperatures and on the evolution of polar ecosystems; NICHOLAS POLUNIN, *Botany of the Canadian Eastern Arctic,* part 3, *Vegetation and Ecology* (1948); *Circumpolar Arctic Flora* (1959); HEINRICH WALTER, *Vegetation of the Earth and Ecological System of the Geobiosphere,* 3rd rev. and enlarged ed. (1985); originally published in German, 5th rev. ed., 1982), on the regional distribution of vegetation types over the world; J.W. BEWS, *The World's Grasses: Their Differentiation, Distribution, Economics, and Ecology* (1929), a comprehensive review of the nature of grasses and grasslands of Africa and their relation to those of other parts of the Earth; G.W. BROWN, JR. (ed.), *Desert Biology,* 2 vol. (1968–74), a symposium volume in which numerous specialists discuss their own fields of desert biology in varying depth, with attention concentrated on the deserts of North America; J.L. CLOUDSLEY-THOMPSON and M.J. CHADWICK, *Life in Deserts* (1964), a general account intended for use in colleges; and PAUL W. RICHARDS, *The Life of the Jungle* (1970), a lavishly illustrated popular account of rain forest ecology, and *The Tropical Rain Forest: An Ecological Study* (1952, reissued 1979), a standard work on the tropical rain forest.

(J.S.O./L.C.B./N.Po./F.W.We./R.T.C./D.W.G./P.W.R./Ed.)